Reconhecimento de entidades mencionadas em português

Documentação e actas do HAREM,
a primeira avaliação conjunta na área

Diana Santos e Nuno Cardoso
editores

Linguateca, 2007

Reconhecimento de entidades mencionadas em português

Documentação e actas do HAREM,
a primeira avaliação conjunta na área

Diana Santos e Nuno Cardoso
editores

Linguateca, 2007

1ª Edição, Novembro de 2007.
1st Edition, November 2007.

Publicação Digital. *Digital Print.*

ISBN 978-989-20-0731-1

O capítulo 12, "Functional Aspects of Portuguese NER", foi anteriormente publicado em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 2006. Proceedings.* p. 80-89, na série LNAI, Vol. 3960 da editora Springer Verlag, ISBN-10: 3-540-34045-9.
The chapter 12, "Functional Aspects of Portuguese NER", was republished from Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 2006. Proceedings. pp. 80-89, Springer Verlag, LNAI series, Vol. 3960, ISBN-10: 3-540-34045-9.

O capítulo 16, "Directivas para a identificação e classificação semântica na colecção dourada do HAREM", foi previamente publicado como Relatório Técnico DI/FCUL TR–06–18, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.
The chapter 16, "Directivas para a identificação e classificação semântica na colecção dourada do HAREM*", was previously published as Technical Report DI/FCUL TR–06–18, Department of Informatics, Faculty of Sciences, University of Lisbon.*

O texto do capítulo 17, "Directivas para a identificação e classificação morfológica na colecção dourada do HAREM", foi previamente publicado como Relatório Técnico DI/FCUL TR–06–19, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.
The chapter 17, "Directivas para a identificação e classificação morfológica na colecção dourada do HAREM*", was previously published as Technical Report DI/FCUL TR–06–19, Department of Informatics, Faculty of Sciences, University of Lisbon.*

O capítulo 18, "Avaliação no HAREM: Métodos e medidas", foi previamente publicado como Relatório Técnico DI/FCUL TR–06–17, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.
The chapter 18, "Avaliação no HAREM*: Métodos e medidas", was previously published as Technical Report DI/FCUL TR–06–17, Department of Informatics, Faculty of Sciences, University of Lisbon.*

Prefácio

Não quisemos que a divulgação do trabalho feito no HAREM sofresse um atraso tão significativo como o que ocorreu por ocasião das Morfolimpíadas (cujo livro saíu à luz quatro anos depois). Por isso, decidimos publicar a presente obra de forma electrónica e gratuita, de forma a maximizar o seu alcance e minimizar o tempo de saída.

Isso não obstou, naturalmente, a que tivéssemos seguido um processo editorial rigoroso, com revisão cruzada entre os autores, além de amplos comentários e sugestões pelos dois editores, numa tentativa de tornar os capítulos mais homogéneos entre si, e ainda a leitura crítica da primeira versão completa do livro por vários especialistas em processamento computacional do português, que resultou em várias sugestões valiosas e observações pertinentes.

Para que conste, aqui fica a nossa profunda gratidão a essa comissão informal de redacção, que foi constituída (por ordem alfabética) por António Teixeira, Daniel Gomes, Graça Nunes, Jorge Baptista, Luís Costa e Paulo Gomes. Agradecemos também a leitura aturada do primeiro capitulo pelo Eugénio Oliveira com valiosos comentários, e queremos fazer uma menção especial à Cristina Mota pelo cuidado e pormenor com que reviu todos os outros capítulos do livro, fazendo sugestões valiosíssimas. Embora como trabalho de bastidores, foi também muito importante a contribuição do Luís Miguel Cabral para o processamento das referências bibliográficas.

A organização de uma avaliação conjunta de raiz é algo que exige um grande empenhamento e muito trabalho, por isso nos parece importante que aquilo que se aprendeu e que foi feito possa ser reaproveitado por outros – os leitores do presente livro. Ao contrário de fechar aqui o trabalho nesta área e partir para outra, pretendemos também com este livro potenciar e possibilitar a preparação de futuras avaliações conjuntas em REM,

e em particular o Segundo HAREM que, à data de escrita deste prefácio, acaba de ser iniciado. Assim, tivemos o cuidado de republicar as directivas no presente volume e criar uma documentação mais cuidada dos próprios programas de avaliação, para facilitar a sua utilização e mesmo reprogramação.

Como nunca é demais ser repetido, na organização do HAREM não estivemos sós: contámos com a preciosa colaboração (por ordem alfabética) de Anabela Barreiro, Luís Costa, Paulo Rocha, Nuno Seco, Rui Vilela e Susana Afonso. E gostávamos de agradecer também a todos os participantes no Primeiro HAREM e também aos participantes no Encontro do HAREM no Porto pela participação e valiosas sugestões, participação e ideias essas que tudo fizemos para se encontrarem fielmente reflectidas pelo presente volume.

Como todo o trabalho feito no âmbito da Linguateca, o que nos moveu foi o desejo de uma melhoria significativa das condições do processamento computacional da língua portuguesa e, na esteira do modelo IRA (informação, recursos e avaliação), além da avaliação conjunta propriamente dita criámos recursos importantes para o REM em português (a colecção dourada, e os sistemas de avaliação). Com este livro, estamos a pôr em prática a terceira vertente, de informação.

Resta-nos agradecer a todos quantos tornaram este projecto (HAREM, e a própria Linguateca) possível, e acusar com gratidão o financiamento recebido, através dos projectos POSI/PLP/43931/2001 (2001-2006) e POSC 339/1.3/C/NAC (2006-2008).

Oslo e Lisboa, 5 de Novembro de 2007

Os editores

Diana Santos e Nuno Cardoso

Preface

This is a book about the First HAREM, an evaluation contest in named entity recognition in Portuguese, organized in the scope of the Linguateca project to foster R&D in the computational processing of Portuguese.

Although inspired by MUC, the path followed in HAREM was based on a different semantic model, aiming at identifying and classifying all proper names in text with the help of a set of 10 categories and 41 subcategories (called types), and allowing vague categories in the sense of merging two or more interpretations (as the geopolitical class in ACE, which conflates place and organization, but not only in that case).

HAREM had 10 participants in its first edition, which in fact included two evaluation events, the first event and Mini-HAREM (only for those who had participated before), which allowed us to perform some statistical validation studies and increase the evaluation resources. Because we had participants from non-Portuguese speaking countries (Denmark, Spain and Mexico), we have four chapters in English in this book, and therefore a preface in English is due as well.

This book reflects the participation and the discussion in the final HAREM workshop that took place in July 2006 after Linguateca's first summer school in Porto. It is organized in three parts, after an encompassing introduction:

1. Fundamentals of HAREM: history, preliminary studies, comparison with MUC and ACE, discussion of the semantic choices, statistical validation, a proposal for future venues, and a chapter summing up what was achieved and which future prospects we envisage.

2. Participation in HAREM: most participants wrote a chapter describing their systems,

approaches and results in HAREM evaluations, often also suggesting improvements or changes for the future.

3. HAREM documentation: the material produced by the organization, such as the guidelines for the annotation of the golden collection, the evaluation metrics, the evaluation software architecture, and the distribution of the golden collection as a regular corpus as well.

Following the usual procedure in Linguateca, abiding by the IRE model (information - resources - evaluation), we organized the evaluation contest, we made the resources therein available to the community, and we now gather and produce information about the whole endeavour, in the form of the present book.

We thank all participants in HAREM, our fellow organizers (Susana Afonso, Anabela Barreiro, Paulo Rocha, Nuno Seco and Rui Vilela), Luís Miguel Cabral who processed the book's references, and all those who participated as book reviewers (Luís Costa, Daniel Gomes, Paulo Gomes, Cristina Mota, Graça Nunes and António Teixeira) and whose help led to a considerable increase in quality.

All work in HAREM was done in the scope of the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract references POSI/PLP/43931/2001 and POSC/339/1.3/C/NAC.

Oslo and Lisbon, 5th November, 2007

The editors,

Diana Santos and Nuno Cardoso

Autores

Antonio Toral Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Andrés Montoyo Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Bruno Martins Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal, *agora* Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Christian Nunes Aranha Cortex Intelligence, Brasil.

Cristina Mota Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal, *agora* Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal / L2F, INESC-ID, Portugal / New York University, EUA.

Diana Santos Linguateca, SINTEF ICT, Noruega.

Eckhard Bick VISL, Institute of Language and Communication, University of Southern Denmark, Dinamarca.

José João Dias de Almeida Departamento de Informática, Universidade do Minho, Portugal.

Luís Sarmento Linguateca, CLUP, Faculdade de Letras da Universidade do Porto, Portugal, *agora* Faculdade de Engenharia da Universidade do Porto, Portugal.

Marcirio Chaves Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Marília Antunes Universidade de Lisboa, Faculdade de Ciências, Portugal.

Mário J. Silva Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Max Silberztein LASELDI, Université de Franche-Comté, França.

Nuno Cardoso FCCN, Linguateca, Portugal, *agora* Universidade de Lisboa, Faculdade de Ciências, LaSIGE, Portugal.

Nuno Seco Linguateca, Grupo KIS, Centro de Informática e Sistemas da Universidade de Coimbra, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Portugal.

Óscar Ferrández Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Paulo Rocha Linguateca, Grupo KIS, Centro de Informática e Sistemas da Universidade de Coimbra, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Portugal.

Rafael Muñoz Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Rui Vilela Departamento de Informática, Universidade do Minho, Portugal.

Thamar Solorio Human Language Research Institute, Universidade do Texas, Dallas, EUA.

Zornitsa Kozareva Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Espanha.

Capítulo 1

Breve introdução ao HAREM

Diana Santos e Nuno Cardoso

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 1, p. 1–16, 2007

Este capítulo apresenta o HAREM, tentando constituir algo interessante para leitores sem conhecimento prévio da área, passando por pessoas interessadas e conhecedoras do paradigma de avaliação conjunta, até aos próprios participantes no HAREM. Apresentamos a motivação para a realização do HAREM e consequente publicação deste volume, ao nível da necessidade de avaliação na área do processamento computacional da língua portuguesa em geral, e as razões que motivaram a escolha da área específica do reconhecimento das entidades mencionadas.

Prosseguimos com uma breve descrição sobre o evento que inspirou o HAREM, o MUC, assim como toda a história da organização do HAREM.

Depois de esclarecermos a terminologia e fixarmos as designações **HAREM**, **Primeiro HAREM** e **Mini-HAREM**, descrevemos o Primeiro HAREM em detalhe.

Essa descrição abarca, com o respectivo calendário:

- O trabalho preparatório;
- A criação dos recursos de avaliação;
- A organização da primeira avaliação;
- A organização do Mini-HAREM.

Produzimos depois um pequeno guia sobre onde encontrar mais documentação sobre o HAREM, fazendo uma espécie de inventário das publicações associadas, e terminamos o capítulo com uma pequena apresentação do presente livro, que marca a última contribuição do Primeiro HAREM.

1.1 O modelo da avaliação conjunta

Há poucos anos atrás, o processamento do português estava numa fase pré-científica, em que os (poucos) trabalhos publicados relatavam no máximo a sua própria auto-avaliação. Isso impedia, na prática, a reprodução dos resultados, inibindo o progresso na área e impedindo a formação de uma verdadeira comunidade científica que pudesse comparar abordagens e métodos aplicados a uma tarefa comum.

Essa situação foi identificada como um dos principais entraves ao progresso do processamento computacional da nossa língua em Santos (1999), e tem vindo a ser progressivamente modificada através da actuação da Linguateca nesse campo (Santos, 2007a).

A Linguateca possui três eixos de actuação: a informação, os recursos e a avaliação.[1] Nesta última vertente, promovemos desde o início o modelo da avaliação conjunta, tendo

[1] Para uma panorâmica da Linguateca através dos tempos veja-se entre outros Santos (2000, 2002); Santos et al. (2004); Santos e Costa (2005); Santos (2006c), assim como a lista de publicações constantemente actualizada no sítio da Linguateca.

organizado as Morfolimpíadas em 2002-2003 (Santos et al., 2003; Costa et al., 2007) e participando anualmente na organização do CLEF para o português desde 2004 (Rocha e Santos, 2007). Em 2005 iniciámos a organização do HAREM, a que se refere o presente volume e capítulo.

Ao possibilitar a comparação de diferentes abordagens de uma forma justa e imparcial, estas avaliações conjuntas fomentam o desenvolvimento de melhores sistemas e contribuem para a melhoria do desempenho destes. Além disso, permitem definir em conjunto uma área e avaliar e comparar tecnologias diferentes, além de fixarem e tornarem público um conjunto de recursos para avaliar e treinar sistemas no futuro. Para uma defesa alongada deste paradigma, veja-se Santos (2007b).

1.2 Entidades mencionadas

"Entidades mencionadas" (EM) foi a nossa tradução (ou melhor, adaptação) do conceito usado em inglês, *named entities*, e que literalmente poderá ser traduzido para "entidades com nome próprio".

A tarefa que nos propusemos avaliar era a de reconhecer essas entidades, atribuindo-lhes uma classificação (dentre um leque de categorias previamente definido e aprovado por todos) que representaria o significado daquela ocorrência específica da entidade no texto em questão.

Nós vemos o reconhecimento de entidades mencionadas (REM) como um primeiro passo na análise semântica de um texto. Separámos esse reconhecimento em duas subtarefas separadas: a **identificação** (de que uma dada sequência de palavras constitui uma EM) e a **classificação** (a que categoria semântica essa EM pertence, naquele contexto).

A razão para abordarmos esta tarefa foi a nossa convicção de que o REM é parte integrante da maioria dos sistemas inteligentes que processam e interpretam a língua, tais como sistemas de extracção de informação, de resposta automática a perguntas, de tradução automática, ou de sumarização de textos. Visto que a qualidade do REM nestes sistemas influencia decisivamente o seu resultado final, estamos convencidos de que a organização de avaliações específicas sobre REM pode beneficiar fortemente o progresso nestas tarefas.

A tarefa de REM necessita de uma clarificação das bases semânticas e pragmáticas do processamento de linguagem natural que não são necessariamente consensuais ou explícitas, pelo que a delimitação precisa do conceito de entidade mencionada e da sua operacionalização prática veio fazer correr muita tinta. O capítulo 4 deste livro é dedicado precisamente a este assunto, que não será portanto abordado aqui.

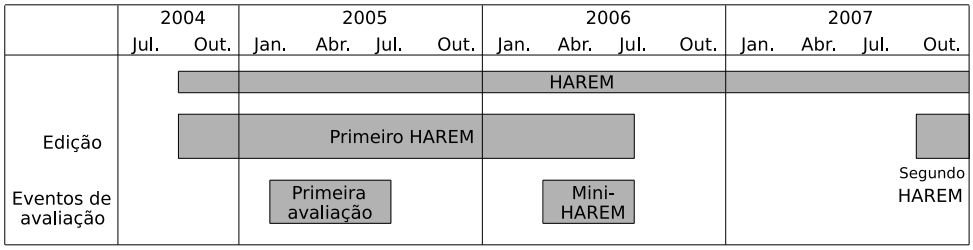


Figura 1.1: Diagrama temporal das edições e eventos de avaliação do HAREM.

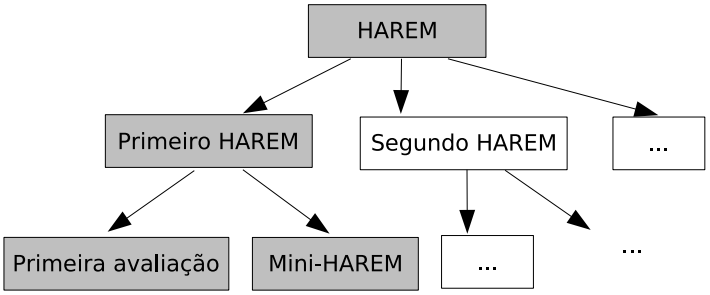


Figura 1.2: Terminologia usada no HAREM. Os eventos cobertos no presente livro estão marcados a cinzento.

1.3 A terminologia que emergiu do HAREM

Antes de prosseguirmos com uma análise histórica do desenvolvimento do HAREM, é essencial fixar a terminologia que vai ser usada neste livro e que foi surgindo muito pouco sistematicamente com as variadas fases da história do HAREM.

Assim sendo, a Figura 1.1 fornece um diagrama temporal das etapas do HAREM, enquanto que a Figura 1.2 indica graficamente as inclusões dos variados termos no contexto do HAREM.

1.4 Um pouco de história

Não fomos certamente os primeiros a achar que a detecção dos significados (ou categorias ontológicas) de nomes próprios seria uma sub-tarefa passível de avaliação separada. Cabe aqui contudo indicar como surgiu a inspiração, e até admitir que, no processo seguido, nem todas as outras fontes teoricamente possíveis de ser inspiradoras (porque já passadas) foram consultadas.

1.4.1 A inspiração

Foi o MUC (Message Understanding Conference), uma avaliação conjunta na área de extracção de informação (EI) existente desde 1987 (Hirschman, 1998), que propôs pela primeira vez, na sua sexta edição, que a tarefa de REM fosse medida de uma forma independente, após ter sido considerada durante vários anos como uma parte da tarefa mais geral de extrair informação de um texto (Grishman e Sundheim, 1996).

Embora os resultados da tarefa de REM, tal como definida pelo MUC, se tivessem situado a níveis muito altos de desempenho (mais de metade dos participantes obtiveram medidas F superiores a 90%), o que foi considerado um resultado comparável ao dos seres humanos, nem todos os investigadores aceitaram que isso indicava que a tarefa de REM já estava resolvida (veja-se por exemplo Palmer e Day (1997); Mikheev et al. (1999)). Por um lado, havia a questão da língua: "resolvido" para o inglês não significa resolvido para todas as línguas. Por outro lado, era preciso avaliar que métodos ou recursos eram necessários para essa tarefa.

Assim, após o MUC, vários outros eventos de avaliação focando o REM se seguiram, como o MET (Merchant et al., 1996), a tarefa partilhada do CoNLL (Sang, 2002; Sang e Meulder, 2003) ou o ACE (Doddington et al., 2004).

Enquanto o MET adoptou directamente a tarefa do MUC aplicando-a a japonês, espanhol e chinês, a tarefa partilhada do CoNLL procurou fomentar a investigação em sistemas de REM independentes da língua, usando textos em flamengo, espanhol, inglês e alemão mas reduzindo significativamente a grelha de classificação, que passou a conter apenas quatro categorias semânticas: `LOC` (local), `ORG` (organização), `PER` (pessoa) e `MISC` (diversos), simplificando portanto ainda mais a tarefa.

O ACE, pelo contrário, propôs a pista de *EDT - Entity Detection and Tracking*, em que o objectivo é fazer o reconhecimento de entidades, quer sejam quer não mencionadas através de um nome próprio, o que alarga consideravelmente a dificuldade da tarefa. O REM passa pois no ACE a compreender todo o reconhecimento semântico de entidades, sejam elas descritas por nomes comuns, próprios, pronomes, ou sintagmas nominais de tamanho considerável. Além disso, há um alargamento significativo das categorias usadas, como são exemplos as categorias armas, veículos ou instalações (em inglês, *facilities*), assim como a definição de uma "supercategoria" para locais+organizações, chamada "entidade geopolítica".

Deve ser referido que a inspiração directa e mais importante para o HAREM foi o MUC, e o nosso interesse de delimitarmos o problema em português e para o português, fez-nos duvidar ou não levar suficientemente a sério as iniciativas multilingues. Quanto ao ACE, foi tarde demais que soubemos das actividades deste, o que teve como consequência não nos termos inspirado nele para a organização do HAREM.

Por outro lado, convém lembrar que, em 2003 e 2004, altura em que surgiram várias

iniciativas de problematização e alargamento do REM, tais como o encontro de Guthrie et al. (2004), a Linguateca já estava em pleno no meio da organização do HAREM (ou do ensaio pré-HAREM), que será descrito em seguida.

1.4.2 Avaliação de REM em português antes do HAREM

O HAREM começou a ser planeado em Junho de 2003, por ocasião do Encontro AvalON.[2] Além de constituir o encontro final das Morfolimpíadas (Santos et al., 2003; Costa et al., 2007), nesse encontro foram discutidas e preparadas várias outras iniciativas, tendo sido lançadas as bases para um plano organizado de avaliações conjuntas em português, coadjuvado por uma comunidade científica interessada em participar em futuros iniciativas de avaliação semelhantes. Assim, foram convidadas várias pessoas a apresentar propostas concretas, uma das quais, da responsabilidade da Cristina Mota, era o culminar de um ensaio que visava medir ou auscultar o problema do REM em português.

Com efeito, esta investigadora tinha organizado nos meses antecedentes um ensaio, mais tarde documentado em Mota et al. (2007) e agora mais profusamente no capítulo 2 do presente livro, cujo objectivo era medir precisamente a dificuldade da tarefa de REM, abordando várias questões que ainda não tinham sido consideradas (ou, pelo menos, documentadas) em eventos anteriores.

O ensaio mostrou que:

- Muitos investigadores marcaram manualmente os textos usando uma hierarquia de classes semânticas bem mais vasta do que as hierarquias estipuladas por exemplo pelo MUC, o que mostra que a sua concepção de REM era diferente da reflectida pelos eventos de avaliação em REM da altura.

- A discordância entre anotadores era significativa, não só na interpretação do que é uma EM, mas também na identificação e na classificação das EM. Uma possível ilação a retirar foi a necessidade de incorporar o conceito de vagueza, quer na identificação quer na classificação, de forma a poder entrar em conta com as divergências, num ambiente de avaliação onde se mede e pontua o desempenho dos sistemas.

A apresentação das conclusões desse ensaio desencadeou uma discussão muito produtiva e participada sobre várias questões no encontro AvalON, tendo vários grupos sugerido que se começasse pelo REM geográfico. Contudo, pareceu-nos demasiado redutor cingir a futura tarefa de REM apenas à categoria dos locais em português, até porque um dos aspectos interessantes da avaliação seria medir a "confundibilidade" de nomes de locais com outras entidades.

[2] O Encontro AvalON, `http://www.linguateca.pt/avalon2003/`, foi um encontro sobre avaliação conjunta organizado pela Linguateca, que decorreu como um encontro satélite da 6ª edição do PROPOR em Faro (Mamede et al., 2003).

Este estudo serviu de inspiração para a organização do HAREM, que acabou por não incluir como organizadora a própria iniciadora do processo por razões relacionadas com a dedicação exclusiva desta nesse período à sua tese de doutoramento, e pelo facto de, além disso, pretender participar no HAREM, como veio a acontecer (veja-se o capítulo 15).

Embora tenhamos divergido em muitas questões da proposta original da Cristina Mota, é indubitavelmente a este ensaio que o HAREM mais deve a sua génese.

1.4.3 A preparação do Primeiro HAREM

O Primeiro HAREM teve o seu início oficial em Setembro de 2004, com um anúncio e chamada à participação através de mensagens nas listas e por mensagens directas aos já conhecidos possíveis interessados, saídos do ensaio inicial e da lista sobre avaliação mantida pela Linguateca.

Os autores do presente capítulo expuseram nessa altura a intenção da Linguateca de desenvolver uma metodologia nova para avaliar o REM, usando uma colecção de textos de diferentes géneros textuais e de várias variantes (a colecção do HAREM – CH), como base para criar uma colecção dourada (CD), ou seja, uma colecção devidamente anotada por seres humanos e que constituiria a bitola de comparação utilizada no HAREM.

As categorias semânticas seriam criadas por todos os participantes a partir da análise cuidada dos textos, e as directivas seriam continuamente aperfeiçoadas à medida que se progredia na tarefa de anotação da colecção dourada.

Nessa altura estabeleceu-se um grupo inicial de interessados, que se declararam participantes ou apenas observadores (por exemplo, interessados no problema mas que não tinham intenções ou condições de desenvolver um sistema REM para participar). Tivemos dez observadores, quatro dos quais participaram no exercício de anotação manual inicial (Débora Oliveira, Elisabete Ranchhod, John Cullen e Jorge Baptista), pelo qual manifestamos aqui a nossa gratidão.

Após coligir uma colecção de textos para a CD, o primeiro passo foi a divisão da CD em vários pedaços. A 26 de Outubro de 2004 foi entregue aos participantes (ou observadores) um pedaço diferente para o anotarem manualmente no prazo de duas semanas, seguindo uma proposta inicial de regras de etiquetagem e um conjunto inicial de categorias semânticas, meramente indicativas. Os participantes nessa anotação cooperativa foram mesmo instados a alargar ou mesmo "desobedecer" às directivas, e partilhar os seus argumentos com o resto da comunidade.

Com esta actividade, tentámos atingir vários objectivos:

- Em primeiro lugar, os participantes e observadores familiarizaram-se de imediato com as dificuldades da tarefa, nomeadamente a vagueza[3] da identificação e da classificação semântica, e a escolha das categorias e tipos semânticos a usar na hierarquia

[3] Sobre a questão da ubiquidade da vagueza em linguagem natural, ver Santos (1997).

final, que abranja adequadamente as EM reconhecidas. Desta forma, as discussões conjuntas em torno da metodologia do HAREM deixaram o reino do abstracto e foram muito mais produtivas e orientadas para os reais requisitos da tarefa em questão.

- A participação activa dos participantes e observadores nas etapas da organização da primeira avaliação do Primeiro HAREM tentou garantir que este correspondesse às necessidades da comunidade, e que os seus objectivos fossem ouvidos e levados em conta na metodologia em desenvolvimento. Ou seja, tentámos chegar a uma metodologia que traduzisse o que a comunidade entendia por REM em português, e que estaria implementada nos seus sistemas, evitando o erro de estipular uma tarefa desfasada da realidade que se pretende avaliar. Se tal foi ou não cabalmente conseguido, poderá ser julgado pelos capítulos de discussão no presente volume.

Durante o processo de anotação dos pedaços, várias dúvidas e casos "difíceis" (ou, simplesmente, casos que causaram discordâncias) foram debatidos, servindo de base para elaborar a primeira revisão às directivas, cuja discussão, pelos participantes, observadores e público em geral, teve como prazo final o dia 5 de Novembro de 2004. Os pedaços anotados foram entregues até ao dia 19 de Novembro de 2004.

Estes pedaços voltaram a ser reunidos numa verdadeira CD anotada, que foi exaustivamente revista por quatro anotadores da Linguateca: os autores do presente capítulo, Anabela Barreiro e Susana Afonso. Contudo, é preciso confessar que, no processo de revisão, as directivas não deixaram de ser aperfeiçoadas, quando assim achámos oportuno. A 16 de Dezembro de 2004, foi distribuído aos participantes um pedaço da CH etiquetado conforme as directivas em vigor, para poderem adaptar os seus sistemas e familiarizarem-se com o formato a empregar no HAREM. Até 10 de Janeiro de 2005, a organização dedicou-se aos aspectos associados com a medição dos sistemas, nomeadamente as directivas de avaliação e a definição da arquitectura de avaliação. Contudo, a CD continuou a ser revista aturadamente, com alterações pontuais às directivas oportunamente divulgadas. Entre 10 de Janeiro e 14 de Fevereiro de 2005 não foram realizadas mais alterações, para que se pudesse dar tempo aos participantes para adaptar os seus sistemas às directivas oficiais do HAREM.

1.4.4 O primeiro evento do Primeiro HAREM

O primeiro evento de avaliação teve início no dia 14 de Fevereiro de 2005. Os dez participantes (descritos na Tabela 1.1), oriundos de seis países diferentes (Brasil, Dinamarca, Espanha, França, México e Portugal), receberam a CH sem anotações, que tinham de devolver, marcada automaticamente passadas 48 horas. Foram-nos enviadas 18 saídas dentro do prazo e 3 saídas fora do prazo (não-oficiais, portanto).

Sistema	Participante	Instituição
CaGE	Mário J. Silva, Bruno Martins e Marcirio Chaves	Grupo XLDB, Universidade de Lisboa
Cortex	Violeta Quental	PUC-Rio/CLIC
ELLE	Isabel Marcelino	Pólo da Linguateca no LabEL
Malinche	Thamar Solorio	INAOE
NERUA	Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Rafael Muñoz e Andrés Montoyo	Universidade de Alicante
PALAVRAS-NER	Eckhard Bick	University of Southern Denmark
RENA	Edgar Alves e José João Dias de Almeida	Universidade do Minho
RSN-NILC	Graça Nunes, Ricardo Hasegawa e Ronaldo Martins	NILC
SIEMÊS	Ana Sofia Pinto, Luís Sarmento e Luís Miguel Cabral	Pólo do Porto da Linguateca
Stencil/NooJ	Cristina Mota e Max Silberztein	IST e LASELDI, Université de Franche-Comté

Tabela 1.1: Participantes na primeira avaliação do Primeiro HAREM

Passados mais dois dias, a colecção dourada (CD) (ou seja, o subconjunto anotado da colecção HAREM, CH) foi divulgada aos participantes, para eles próprios, se assim o desejassem, analisar as soluções e eventualmente alertar para possíveis erros.

Era tempo para desenvolver a plataforma de avaliação (capítulo 19 e Seco et al. (2006)), na qual, além dos autores do presente capítulo, participaram Nuno Seco e Rui Vilela.

O HAREM inspirou-se nas métricas de avaliação do MUC para a avaliação comparativa das saídas dos sistemas (Douthat, 1998). Contudo, foram introduzidos diversos melhoramentos para lidar com várias questões não contempladas no MUC, tais como a vagueza, a separação entre a avaliação da identificação e a da classificação semântica (categorias e tipos), o conceito de correcção parcial, e a avaliação separada por cenários distintos. Além disso, foram também aproveitados alguns conceitos da experiência anterior das Morfolimpíadas, tal como a distinção entre medidas absolutas e relativas (Santos et al., 2003; Costa et al., 2007). As métricas de avaliação, bem como as medidas, regras e as pontuações usadas no cálculo do desempenho dos sistemas, foram publicadas a 29 de Setembro de 2005. A última redacção desse texto (mas sem mudanças em relação à substância) encontra-se no capítulo 18 deste livro.

A 22 de Abril de 2005, foi apresentada aos participantes uma primeira arquitectura da plataforma de avaliação, permitindo a avaliação por cenários, e implementando na totalidade as directivas de avaliação entretanto colocadas públicas. Também nesta fase, os participantes podiam acompanhar o trabalho desenvolvido e opinar sobre as regras de avaliação e a pertinência das medidas, já com a ajuda dos exemplos concretos disponibili-

zados com a documentação dos programas.

A 20 de Maio de 2005 foram enviados aos participantes os primeiros resultados do HAREM, respeitantes à tarefa de identificação. Os resultados globais, devidamente anonimizados, foram tornados públicos a 9 de Junho de 2005. Uma semana depois, eram divulgados os resultados relativos à classificação morfológica.

É preciso mais uma vez salientar que as directivas de avaliação foram continuamente revistas (e tornadas mais pormenorizadas), pois, à medida que se desenvolviam os programas de avaliação, algumas situações particulares iam sendo detectados e resolvidos.

A grande demora na publicação dos resultados ficou no entanto também a dever-se ao facto de quase todas as saídas submetidas ao HAREM não respeitarem as regras de etiquetagem, o que levou à necessidade de normalizar manualmente as saídas enviadas, e interagir com os participantes no sentido de resolver estes problemas.

Assim sendo, só a 6 de Setembro de 2005 (sensivelmente sete meses após os participantes terem enviado o resultado dos seus sistemas) é que foi possível divulgar os resultados finais da tarefa de classificação semântica, juntamente com uma revisão ligeira dos valores para a tarefa de identificação, que não apresentou alterações significativas na ordenação dos participantes. Seguiram-se os resultados da tarefa da classificação morfológica, publicados em 29 de Setembro de 2005. Finalmente, o processo foi dado por concluído com o envio dos resultados individuais, para todas as tarefas, aos participantes, a 28 de Outubro de 2005.

1.4.5 O Mini-HAREM: medição do progresso e validação estatística

Considerando que os resultados do HAREM já não representavam fielmente o estado dos sistemas concorrentes, e que o atraso na publicação destes tinha resultado em alguma desmotivação da comunidade, resolvemos repetir, ainda dentro do Primeiro HAREM, a comparação entre os sistemas que estivessem dispostos a enviar novas saídas. Uma vez que a arquitectura de avaliação se encontrava concluída e os programas prontos, livremente disponíveis e amplamente testados com os mesmos sistemas que iriam participar, não se previam atrasos substanciais na publicação dos resultados da nova avaliação conjunta.

A este novo evento de avaliação chamou-se o Mini-HAREM, e a participação no dito foi restrita apenas aos participantes do primeiro evento. O Mini-HAREM empregou a mesma metodologia do HAREM – com excepção de algumas pequenas alterações nas categorias. Muito brevemente,

- o tipo `PRODUTO` da categoria `OBRA` foi suprimido;
- o tipo `MEMBROCLASSE` foi adicionado à categoria `COISA`;
- os URL e os endereços de correio electrónico deixaram de ser considerados EM.

Os participantes foram evidentemente informados com antecedência destas ligeiras mudanças, mas não de qual colecção de textos os seus sistemas iriam classificar. De facto, foi distribuída aos participantes a mesma CH; a diferença residia no uso de uma nova CD. A constituição desta segunda CD usada no Mini-HAREM, a que chamamos CD 2006, é semelhante à da primeira CD, chamada CD 2005, e os seus documentos são disjuntos.

O Mini-HAREM teve os seguintes objectivos (mais detalhados em Cardoso (2006a)):

- A obtenção de mais dados sobre cada sistema participante: ao rever/anotar manualmente mais uma parcela da CH, conseguimos o dobro do material no qual podemos basear a avaliação, ao concatenar as duas CD.
- A obtenção de material para a validação estatística dos resultados dos sistemas participantes (ver capítulo 5): com dois eventos usando a mesma colecção, pode-se medir os sistemas sobre duas colecções douradas e sobre o conjunto destas (ao todo, três recursos de avaliação).
- A medição da evolução dos sistemas ao longo do tempo (desde a altura do primeiro evento até ao Mini-HAREM medeou um ano).
- Uma melhor caracterização do estado da arte em REM para o português.

Para evitar que problemas inesperados na formatação dos resultados dos sistemas atrasassem novamente esta comparação, para o Mini-HAREM foi também desenvolvido um verificador de sintaxe das saídas (ver secção 19.2.1),que permitia que os participantes verificassem se a marcação produzida pelos seus sistemas estava conforme as regras do HAREM e os requisitos dos programas de avaliação do mesmo, antes de enviarem as saídas oficialmente para o HAREM.

Com os programas de avaliação e de geração de relatórios já desenvolvidos, o Mini-HAREM decorreu com maior rapidez. A chamada à participação foi realizada no início de 2006, e o Mini-HAREM foi marcado para o dia 3 de Abril de 2006. Infelizmente, nem todos os participantes no Primeiro HAREM se mostraram interessados, e alguns sistemas tinham mudado de mãos ou sido completamente reestruturados.

O Mini-HAREM contou assim apenas com cinco participantes (descritos na Tabela 1.2), metade dos participantes originais, mas que enviaram 20 saídas, todas oficiais. Os participantes tiveram igualmente um prazo de 48 horas para devolver a colecção do HAREM devidamente etiquetada, um prazo que terminou no dia 5 de Abril de 2006, ao meio-dia, hora de Lisboa.

Não obstante ter sido facultado o validador e termos informado os participantes dos problemas no caso do evento anterior, foi necessário mesmo assim rever manualmente as saídas e corrigir a sua sintaxe para que pudessem ser processadas.

Assim, dois meses depois, a 9 de Junho de 2006, foram divulgados os resultados globais do Mini-HAREM, e os relatórios individuais enviados aos participantes. A comparação dos

Sistema	Participante	Instituição
CaGE	Mário J. Silva, Bruno Martins e Marcirio Chaves	Grupo XLDB, Universidade de Lisboa
Cortex	Violeta Quental e Christian Nunes	PUC-Rio
SIEMÊS2	Luís Sarmento	FEUP/Pólo do Porto da Linguateca
SMELL	Elisabete Ranchhod e Samuel Eleutério	LabEL
Stencil-NooJ	Cristina Mota e Max Silberztein	L2F/INESC e LASELDI, Université de Franche-Comté

Tabela 1.2: Participantes na segunda avaliação do Primeiro HAREM, o Mini-HAREM

dois resultados foi apresentada no Encontro do HAREM no Porto, a 15 de Julho de 2006 (Cardoso, 2006b), além de ser pormenorizadamente discutida em Cardoso (2006a).

1.5 Uma breve descrição da participação no Primeiro HAREM

A participação no Primeiro HAREM foi muito variada, englobando desde sistemas desenvolvidos de raiz para participar no HAREM, como o SIEMÊS (ver capítulo 14) e o ELLE (Marcelino, 2005), até sistemas que participaram "de raspão" para verificar ou estudar questões relativamente marginais, tais como o reconhecimento de entidades geográficas apenas, como o CaGE (capítulo 8), ou a simples identificação de entidades mencionadas através de métodos de aprendizagem automática, como o MALINCHE (capítulo 10).

No meio do espectro tivemos sistemas já existentes, que faziam portanto já alguma forma de REM completo, mas sem necessariamente conceberem o problema do REM como implementado no HAREM (aliás, isso nunca aconteceu), tais como o PALAVRAS-NER (capítulo 12), o Stencil-NooJ (capítulo 15), o NERUA (capítulo 11) ou o Cortex (capítulo 9). Podemos contudo ainda subdividir os sistemas entre aqueles que tentaram de certa forma adaptar o seu funcionamento para participar no HAREM e aqueles que se ficaram por experimentar — sem adaptação -– até onde o seu sistema original conseguia ir, dada a tarefa de avaliação proposta.

Ao contrário das Morfolimpíadas, em que todos os sistemas pertenciam à categoria de sistemas já existentes e bem desenvolvidos, antes da avaliação conjunta, o HAREM parece-nos ter conseguido estimular interesse específico e novo no problema, não só devido ao facto de terem de facto surgido sistemas novos, como pelo interesse unânime em participar em novas edições, expresso por todos os participantes no Encontro do HAREM, e que esperamos poder confirmar-se na prática num futuro breve.

Mais uma vez por oposição às Morfolimpíadas, também temos de reconhecer que não conseguimos que o HAREM cobrisse outras zonas limítrofes. Ou seja, enquanto que

um radicalizador e um corrector ortográfico também participaram nas Morfolimpíadas, desta forma aumentando o âmbito desta avaliação conjunta, a nossa tentativa de alargar o HAREM ao simples reconhecimento de nomes próprios em texto falhou, visto que o NILC (o único sistema que tinha concorrido sob esta perspectiva) preferiu retirar-se por achar que esta última tarefa era demasiado distinta para fazer sentido ser englobada numa avaliação de REM.

1.6 Mais informação sobre o HAREM: um pequeno guia

Ao longo dos mais de três anos de trabalho da Linguateca na área de REM, foi sendo criada documentação variada, não só a nível das páginas na rede no sítio da Linguateca, como também sob a forma de diversos artigos e apresentações e uma tese de mestrado, todos eles sobre o HAREM.

Neste livro parece-nos mais indicado mencionar onde se encontra a informação em relação aos variados temas, em vez de a repetir, embora tenhamos tentado incluir neste volume as especificações fundamentais do HAREM, ao republicar as directivas de anotação e a descrição das medidas, respectivamente nos capítulos 16, 17 e 18.

1.6.1 Ensaio pré-HAREM

O estudo organizado pela Cristina Mota e que inspirou o HAREM foi inicialmente documentado em Mota et al. (2007), por ocasião do livro dedicado ao paradigma de avaliação conjunta (Santos, 2007a). O capítulo 2 constitui uma documentação mais pormenorizada, em que podemos seguir a experiência de anotação de textos do CETEMPúblico e do CETENFolha, que contou com a colaboração de nove investigadores e que foi fundamental para detectar muitos dos problemas que vieram a ser tratados no HAREM.

1.6.2 Metodologia

Quase todos os artigos ou apresentações relativos ao HAREM dão bastante ênfase às inovações metodológicas, quer na definição da própria tarefa, quer na forma de a avaliar. Veja-se pois Santos et al. (2006), Santos (2006a), Santos (2006b) e Seco et al. (2006) para formas diferentes de apresentar o HAREM nessa perspectiva. No capítulo 3 podemos encontrar uma comparação detalhada entre a metodologia do HAREM, e a metodologia adoptada pelo MUC, enquanto o capítulo 4 discute a questão específica do modelo semântico contrastando-o com o do MUC e o do ACE.

De qualquer forma, um prato forte de quase todos os capítulos da parte de discussão do presente volume são as questões metodológicas.

1.6.3 A colecção dourada

Uma parte importante da metodologia refere-se ao conjunto das soluções presentes na CD. Em Santos e Cardoso (2006) detalha-se a criação e as características da CD, bem como a motivação subjacente à decisão em adoptar um leque mais diversificado de categorias e de tipos, e como a vagueza se encontra codificada nas etiquetas usadas pelo HAREM.

Para conhecer a fundo as categorias e as opções utilizadas na criação das colecções douradas, é imprescindível consultar as directivas (capítulos 16 e 17 deste volume). Visto que os sistemas de REM participantes podiam escolher se participavam na classificação semântica, na classificação morfológica, ou em ambas, sendo apenas obrigatória a tarefa de identificação, dividimos as directivas em duas. Como tal, durante a avaliação, a tarefa de identificação encontrava-se descrita em ambos os documentos.

Finalmente, o capítulo 4 de Cardoso (2006a) destila as CD usadas, nomeadamente na sua composição por géneros textuais, categorias semânticas e variantes. Muito desse material foi republicado no capítulo 20 deste volume.

1.6.4 Quantificação: Métricas, medidas, pontuações e regras de cálculo

Embora também apresentadas junto com a metodologia do HAREM (e portanto delineadas nos artigos e capítulos mencionados acima), a apresentação pormenorizada das medidas e métricas do HAREM é feita no capítulo 18, compreendendo as pontuações por cada alinhamento, as regras para lidar com alternativas de identificação, as várias medidas contempladas para cada tarefa, e as métricas usadas para a atribuição de um valor de desempenho às saídas dos sistemas.

1.6.5 A arquitectura e os programas da plataforma de avaliação

A arquitectura da plataforma de avaliação do HAREM foi apresentada em Seco et al. (2006), e detalhada na secção 4.3.3 de Cardoso (2006a). No capítulo 19 apresenta-se a documentação detalhada e definitiva de todos os programas que fazem parte da arquitectura proposta, cujo código fonte se encontra também disponível desde a realização do Mini-HAREM.

1.6.6 Validação estatística

A tarefa de validação estatística aos resultados do HAREM foi o assunto principal da tese (Cardoso, 2006a), onde se descreve o método estatístico utilizado, a metodologia de validação, a sua adaptação aos requisitos do HAREM, e onde se demonstra que o tamanho das colecções usadas nos eventos HAREM é suficiente para comparar adequadamente os sistemas. O capítulo 5 do presente volume resume o trabalho de validação estatística efectuado.

1.6.7 Resultados do HAREM

No capítulo 5 (página 69) e na secção 5.3 de Cardoso (2006a), faz-se uma primeira análise dos resultados globais do HAREM, fornecendo um primeiro panorama de REM em português. Uma selecção dos próprios resultados encontra-se como apêndice deste volume.

1.6.8 Discussão e primeiro balanço

O encontro presencial do HAREM constituiu um primeiro balanço da iniciativa, quer do ponto de vista da organização, quer do ponto de vista dos participantes. As contribuições (ver sítio do Encontro do HAREM) e a discussão ocorrida formaram o ponto de partida para o presente volume, que passamos a descrever brevemente.

1.7 O presente livro

Após variadas reformulações, decidimos dividir o livro em três partes:

1. a parte relacionada com o REM em português;
2. a parte de descrição conjuntural dos sistemas participantes no Primeiro HAREM;
3. a parte de documentação desta primeira avaliação conjunta.

A primeira parte é a que pode ser mais interessante de um ponto de vista teórico, porque descreve questões quer de organização quer de conteúdo de uma avaliação conjunta que são pertinentes para o futuro da área. Não é, contudo, possível nem desejável ficar a um nível de abstracção tão elevado que impeça o leitor de compreender de que tipo de sistemas e/ou problemas estamos a falar.

Para isso é fundamental consultar e compreender a documentação dos próprios sistemas e a explicação dos princípios de funcionamento subjacentes, que constitui a segunda parte do livro, e que poderá servir não só para ilustrar a grande variedade de abordagens e preocupações do leque de participantes, mas também para inspirar a criação de novos sistemas ou a reutilização de técnicas de outros sistemas.

A terceira e última parte é, em grande parte, uma mera republicação das directivas utilizadas, mas a que se juntaram dois capítulos originais: o primeiro sobre a arquitectura dos programas de avaliação, e o segundo sobre a disponibilização das colecções douradas através do projecto AC/DC (Santos e Sarmento, 2003).

Finalmente, pensamos ser necessário que fique fixado e empacotado em forma de livro a destilação do que foi o Primeiro HAREM: as directivas seguidas na anotação da CD e as medidas e métodos de cálculo empregues. Não porque achamos que devam permanecer imutáveis e usadas sempre daqui para a frente, mas porque é preciso que possam ser facilmente referidas (e eventualmente revogadas, ou melhoradas) em futuras edições do HAREM.

Agradecimentos

Embora tenhamos acabado por escrever este capítulo apenas no nosso nome, não queremos deixar de reconhecer que a organização do Primeiro HAREM foi partilhada, em maior ou menor grau, com o Nuno Seco, o Rui Vilela, a Anabela Barreiro, a Susana Afonso e o Paulo Rocha.

E que, claro, sem os participantes e/ou observadores do HAREM não teria havido HAREM.

Quanto ao texto propriamente dito, estamos muito gratos a todos os investigadores que se deram ao árduo trabalho de rever com toda a atenção a nossa primeira versão, e cujas sugestões e recomendações nos levaram a mudanças por vezes substanciais. Foram eles, por ordem alfabética, António Teixeira, Cristina Mota, Daniel Gomes, Eugénio Oliveira, Graça Nunes, Jorge Baptista, Luís Costa e Paulo Gomes. Esperamos que possam reconhecer as melhorias que eles próprios sugeriram.

Este texto, assim como o trabalho que descreve, insere-se no âmbito do trabalho da Linguateca, financiada através dos projectos POSI/PLP/43931/2001 e POSC 339/1.3/C/NAC, e co-financiada pelo POSI.

Parte I

Capítulo 2

Estudo preliminar para a avaliação de REM em português

Cristina Mota

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 2, p. 19–34, 2007.

O presente capítulo visa relatar, de forma mais completa do que em Mota et al. (2007), uma actividade de prospecção realizada em 2003 que serviu de inspiração à organização do HAREM. Essa actividade consistiu na anotação manual ou semi-automática de uma pequena série de extractos do CETEMPúblico (Rocha e Santos, 2000), um corpus que integra artigos extraídos de 1500 edições diárias do jornal *Público*, e do CETENFolha, um corpus correspondente de português do Brasil criado com base no jornal Folha de São Paulo, de 1994. O seu principal objectivo foi preparar e motivar a participação numa futura avaliação conjunta dedicada a sistemas de REM, numa tentativa de compreender quais as categorias das entidades que os sistemas deveriam anotar, bem como estabelecer as directivas que deviam ser seguidas. Salienta-se desde já que, embora os participantes pudessem usar um sistema de base que os auxiliasse na anotação, o objectivo não era comparar o desempenho de sistemas mas sim o que os participantes consideravam como correcto. Apresentamos uma descrição da tarefa levada a cabo e uma análise dos resultados.

No âmbito do seu modelo de trabalho, IRA (Informação-Recursos-Avaliação), a Linguateca iniciou em 2002 actividades que visavam promover a avaliação conjunta de sistemas de processamento de linguagem natural. Estas actividades pioneiras para o processamento de textos escritos em português, bem como os seus primeiros resultados, encontram-se documentados em Santos (2002), Santos et al. (2004) e Santos (2007b). Uma das áreas de actuação escolhida foi a do REM, que começou por ficar a cargo do pólo da Linguateca no LabEL. Essa escolha deveu-se ao facto da presente autora, que na altura era colaboradora no pólo, ter já experiência no desenvolvimento de uma ferramenta de reconhecimento de entidades mencionadas para português.

O HAREM veio então no seguimento deste estudo preliminar, no qual em parte se inpirou. No entanto, houve modificações importantes que se encontram discutidas em vários outros capítulos deste livro, e por isso faz sentido documentar este estudo inicial de forma independente. A primeira tentativa de cristalizar esses passos iniciais foi realizada em Mota et al. (2007), mas dadas as restrições de tamanho (uma secção num capítulo de livro), apresentamos aqui uma descrição mais detalhada.

O arranque do processo deu-se no dia 29 de Janeiro de 2003 com o envio para a lista `avalia@linguateca.pt`, uma lista de divulgação para os investigadores interessados em avaliação conjunta, de uma mensagem com uma primeira proposta de avaliação. Essa proposta solicitava aos interessados na avaliação que anotassem manualmente, ou de forma automática combinada com revisão manual, um conjunto de extractos do CETEMPúblico e do CETENFolha. Esses exctractos anotados deveriam ser enviados até ao dia 21 de Fevereiro de 2003, tendo este prazo inicial sido adiado por coincidir com o prazo de submissão de artigos de várias conferências internacionais. Assim, a nova data estabelecida foi dia 10 de Março de 2003. Os exctractos enviados, bem como uma análise preliminar da classificação feita pelos participantes, foram disponibilizados no sítio da Linguateca logo em

29 de Janeiro de 2003	Envio da proposta inicial
10 de Março de 2003	Data limite para envio dos textos anotados
22 de Maio de 2003	Divulgação dos resultados
28 de Junho de 2003	Sessão de trabalho no AvalON 2003
Setembro de 2004	Início do HAREM

Tabela 2.1: Calendário da actividade preparatória.

seguida. A discussão dos resultados e a preparação de uma futura avaliação conjunta teve lugar no AvalON 2003, a 27 de Julho, na Universidade do Algarve. A Tabela 2.1 apresenta um calendário com as etapas desta actividade preparatória.

Neste capítulo, começamos por descrever a tarefa proposta, apresentamos a análise de resultados e, em jeito de conclusão, alguns comentários finais.

2.1 Descrição da Proposta

A proposta enviada sugeria duas linhas de acção a serem seguidas: a criação cooperativa de directivas; e a criação de recursos de avaliação.

Para a primeira linha de acção, numa primeira fase, pretendia-se estabelecer e caracterizar as entidades que os sistemas teriam de identificar, bem como de que forma as entidades deveriam ser anotadas no texto. Foram exemplificadas algumas entidades, adaptando a classificação do MUC (Grishman e Sundheim, 1995; Chinchor e Marsh, 1998) para português:

- Nomes próprios de
  - Pessoas (ex: Fernando Pessoa, Maria do Carmo, Sampaio)
  - Organizações (ex: IST, Instituto Superior Técnico, Portugal Telecom)
  - Lugares (ex: Sintra, Serra da Estrela, Minho)
- Expressões temporais
  - Datas (ex: 24 de Janeiro de 2000, segundo semestre de 1992, anos 60)
  - Horas (ex: meio-dia, 13:40, 4 horas da manhã)
- Expressões numéricas
  - Monetárias : (ex: 20 milhões de euros, 900 mil contos)
  - Percentuais : (ex: 10,5%, sete por cento)

Além disso, estabeleceu-se que as entidades deveriam ser marcadas com etiquetas SGML, tendo sido fornecidos exemplos de anotação em contexto, adoptando o esquema de marcação original do MUC, tal como se ilustra na Tabela 2.2.

PESSOA	(...) aquilo que <ENAMEX TYPE="PERSON">Fernando Pessoa</ENAMEX> tão expressivamente denominou (...)
ORGANIZAÇÃO	(...) a <ENAMEX TYPE="ORGANIZATION">Portugal Telecom</ENAMEX> voltou a ultrapassar (...)
LUGAR	(...) vai do <ENAMEX TYPE="LOCATION">Minho</ENAMEX> à região do (...)
DATA	Foi durante o <TIMEX TYPE="DATE">segundo semestre de 1992</ENAMEX> que a inflacção (...)
HORA	(...) se estipula as <TIMEX TYPE="TIME">4 horas da manhã</ENAMEX> como limite de (...)
MONETÁRIA	(...) com <NUMEX TYPE="MONEY">900 mil contos</ENAMEX> a fundo perdido (...)
PERCENTAGEM	(...) aos <NUMEX TYPE="PERCENT">sete por cento</ENAMEX> do capital (...)

Tabela 2.2: Exemplos de utilização de cada uma das etiquetas do MUC em extractos da Parte 20 do CETEMPúblico.

Esta linha de acção resultaria num conjunto de critérios e de recomendações (*directivas*) que deveria igualmente conter exemplos que ilustrassem o que devia e não devia ser marcado. A proposta chamava a atenção para algumas das muitas questões que se poderiam colocar e cuja resposta deveria ser tornada clara nas recomendações:

- Quais os tipos de nomes próprios que os sistemas deveriam ser capazes de identificar (e classificar)? Deveria um nome de um estabelecimento comercial (livraria, cinema, discoteca, etc.) ser identificado como uma organização?
- Os sistemas deveriam reconhecer entidades que incluíssem léxico não português, como por exemplo *Empire State Building*, *New York Times*, *BBC* ou *Manchester United*?
- O que fazer no caso de uma entidade estar encaixada noutra? Por exemplo, deveria *Lisboa* fazer parte do nome da organização, como no caso a), e não ser marcada como nome de lugar, ou deveria ser marcada como tal uma vez que não faz parte do nome da instituição, como no caso b) ?

  a) *(...) Crise na faculdade influencia eleições de amanhã para a reitoria da Universidade Técnica de Lisboa (...)*

  b) *(...) A Polícia Judiciária de Lisboa anunciou ontem a conclusão de três inquéritos respeitantes (...)*

A segunda linha de acção consistia na criação de recursos para a avaliação, que seriam anotados manualmente de acordo com os critérios e a classificação estabelecidos nas recomendações. Esses recursos de avaliação constituiriam uma colecção dourada que se-

ria usada como referência na comparação com os resultados produzidos pelos sistemas a partir do mesmo texto sem anotação.

Dado que estas duas linhas de acção poderiam ser desencadeadas em paralelo, foi então sugerido que se começasse por fazer a anotação de dois pequenos conjuntos de textos. A sua dimensão era pequena, apenas os dez primeiros extractos do CETEMPúblico (versão 1.7) e os primeiros vinte[1] do CETENFolha (versão 1.0), porque o objectivo era sobretudo motivar os investigadores para a tarefa. Apesar de tanto o CETEMPúblico como o CETENFolha serem públicos, os extractos para anotar foram disponibilizados no sítio da Linguateca. Deste modo, todos estariam certamente a usar a mesma versão do conjunto de textos. Alternativamente, também foi sugerido que os participantes, em vez de usarem extractos do CETEMPúblico e do CETENFolha, enviassem os textos que preferissem. Talvez por se ter chamado a atenção para o facto de que esta solução tornaria a comparação de resultados mais difícil, ninguém optou por escolher novos textos.

Findo o prazo de duas a três semanas para anotação, ter-se-ia material suficiente para observar a percepção que cada participante tinha sobre o REM, donde poderiam ser tirados resultados comparativos.

A mensagem enviada sugeria ainda que se adoptasse a classificação do MUC adaptada para português e continha o extracto 26 do CETEMPúblico com todos os nomes próprios anotados, quer estivessem ou não contemplados pela classificação do MUC (ver Figura 2.1).

Depois de ter sido enviada a mensagem inicial, precisou-se um pouco melhor a tarefa, aquando da disponibilização da informação no sítio da Linguateca. O objectivo seria que todas as sequências consideradas pelos participantes como sendo nomes próprios deveriam ser delimitadas com a etiqueta SGML `NOMEPROP`, em que o atributo TIPO deveria ter um dos seguintes valores: `PESSOA`, `ORGANIZACAÇÃO`, `LUGAR` ou `OUTRO`. Em alternativa, em vez de `OUTRO`, poderiam ser usadas etiquetas mais específicas, da escolha do participante.

2.2 Descrição dos textos

Como mencionado acima, foram anotados os primeiros dez extractos da versão 1.7 do CETEMPúblico e os vinte primeiros extractos da versão 1.0 do CETENFolha. As Figuras 2.2 e 2.3 mostram respectivamente a distribuição por semestre e por tópico nos dois conjuntos de extractos.

A variedade de semestres no CETEMPúblico deve-se ao facto de o corpus corresponder a 16 semestres compreendidos entre 1991 e 1998, enquanto o CETENFolha só contém edições do ano de 1994. Naturalmente que o conjunto destes extractos é demasiado pequeno para poder tirar quaisquer conclusões que sejam aplicáveis aos corpora completos.

[1] Foi inicialmente sugerido usar também os primeiros 10 extractos do CETENFolha; no entanto, se assim fosse, o número de nomes próprios dos dois subconjuntos seria muito díspar por isso o número de extractos deste corpus foi duplicado.

```
<ext n=26 sec=soc sem=91b>
<p>
<s>O caso ocorreu numa noite de 1978, na ilha de <NOMEPROP TIPO="LUGAR">
 Carvalo</NOMEPROP>, ao largo da <NOMEPROP TIPO="LUGAR">Córsega
 </NOMEPROP>.</s>
<s>O príncipe jantava com amigos num restaurante deste paraíso para
milionários, quando um grupo barulhento de jovens da alta sociedade
italiana acostou na enseada de
<NOMEPROP TIPO="LUGAR">Palma</NOMEPROP>, ao lado do seu iate, o
<NOMEPROP  TIPO="BARCO">L'Aniram</NOMEPROP>.</s>
<s>Os advogados da defesa sublinharam no processo que este facto perturbou
altamente o "senhor de <NOMEPROP TIPO="LUGAR">Sabóia</NOMEPROP>".</s>
<s>Naquele ano, as <NOMEPROP TIPO="ORGANIZAÇÃO">Brigadas Vermelhas
</NOMEPROP> (<NOMEPROP TIPO="ORGANIZAÇÃO">BR</NOMEPROP>) estavam no
auge da actividade terrorista, o líder cristão-democrata <NOMEPROP
TIPO="PESSOA">Aldo Moro</NOMEPROP> acabara de ser raptado, e o príncipe
-- proibido de entrar em <NOMEPROP TIPO="LUGAR">Itália</NOMEPROP>
desde o exílio do pai em 1946 -- teria mesmo recebido ameaças das
<NOMEPROP TIPO="ORGANIZAÇÃO">BR</NOMEPROP>.</s>
</p>
<t>Uma vida por um barco</t>
<p>
<s>O certo é que, pouco depois, <NOMEPROP TIPO="PESSOA">Vítor-Emanuel
 </NOMEPROP> apercebeu-se que um barco pneumático fora deslocado do seu
iate e atracado ao <NOMEPROP TIPO="BARCO">Cocke</NOMEPROP>, o navio dos
jovens italianos.</s>
<s>"Irritado com este acto de apropriação", foi buscar uma espingarda
<NOMEPROP TIPO="ARMA">US 30</NOMEPROP> semiautomática, utilizada em
safaris, e 31 cartuchos, e dirigiu-se para o <NOMEPROP TIPO="BARCO">Cocke
</NOMEPROP>.</s>
<s>Um dos jovens, <NOMEPROP TIPO="PESSOA">Nicola Pende</NOMEPROP>,
 acorda com um grito:</s>
<s>"Roubaste o meu barco, vais pagar."</s>
<s>Pouco depois, o príncipe aponta-lhe a arma ao ventre.</s>
<s>Na confusão que se segue, parte um primeiro tiro, depois um segundo, e
os dois homens caem ao mar.</s>
</p>
</ext>
```

Figura 2.1: Extracto 26 do CETEMPúblico, anotado pela autora.

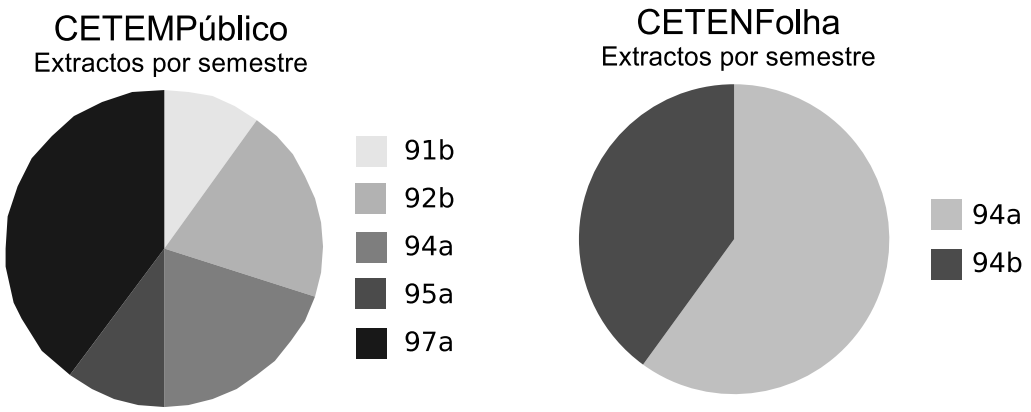


Figura 2.2: Distribuição dos extractos por semestre.

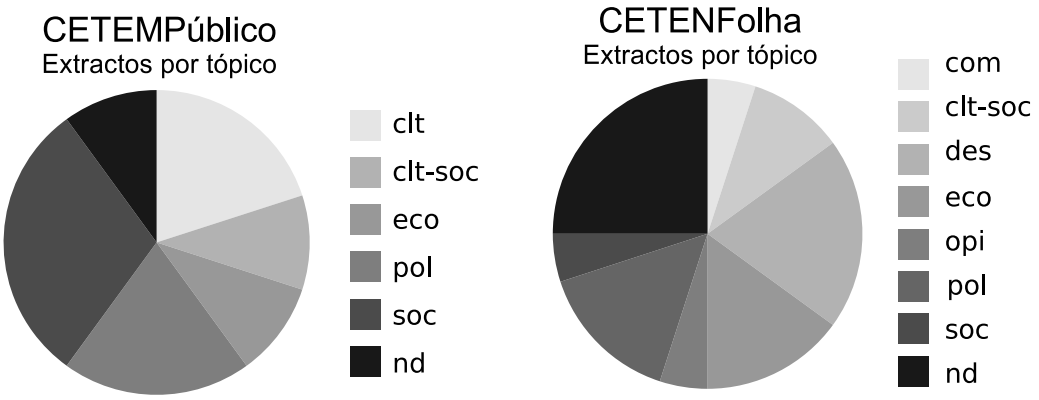


Figura 2.3: Distribuição dos extractos por tópico.

Para além dos extractos não terem sido escolhidos de modo a serem representativos do corpus completo, basta dizer que o semestre com mais extractos no corpus completo é o primeiro semestre de 1992 (92a), que nem sequer se encontra representado no conjunto dos dez extractos seleccionados.

Quanto aos tópicos, o CETENFolha apresenta mais variedade do que o CETEMPúblico. Tal como foi referido anteriormente, inicialmente tinham sido escolhidos também apenas dez extractos do CETENFolha. No entanto, como se pode constatar na Tabela 2.3, em média o subconjunto do CETENFolha apresenta um número significativamente inferior de palavras quer por parágrafo quer por frase, apesar de ter mais do dobro do número de frases e de parágrafos do subconjunto do CETEMPúblico (ver Figura 2.4).

Na Figura 2.4 mostra-se a frequência de várias unidades textuais. Entende-se por *átomo* qualquer sequência de caracteres delimitados pelo espaço; *palavra* são sequências de letras

Número médio	Palavras		Palavras com maiúsculas	
	CETEMPúblico	CETENFolha	CETEMPúblico	CETENFolha
Por parágrafo	82,60	28,37	7,80	3,39
Por frase	25,29	14,36	2,39	1,72

Tabela 2.3: Número médio de palavras e de palavras em maiúsculas por frase e por parágrafo nos dois subconjuntos seleccionados.

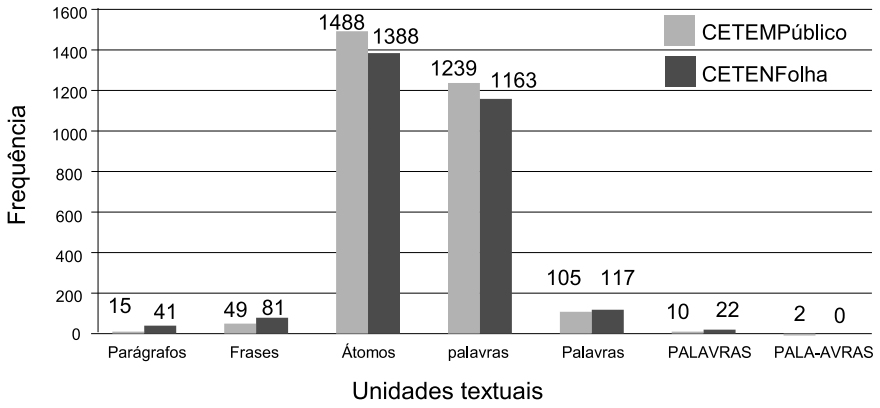


Figura 2.4: Número de ocorrências de várias unidades textuais.

e de caracteres, como o hífen e a barra; *Palavra* é qualquer palavra que comece por uma letra maiúscula; *PALAVRA* é qualquer sequência de letras em maiúsculas e *PALA-AVRAS* qualquer sequência de letras maíusculas e também hífens e barras.

Para se ficar também com uma ideia da variedade das sequências contíguas de palavras em maiúsculas (ou seja, sem considerar que um nome próprio pode conter determinadas palavras que podem não estar em maiúscula, como certas preposições), contabilizou-se o comprimento dessas sequências e o correspondente número de ocorrências (ver Figura 2.5). No CETEMPúblico existem sequências que variam entre comprimento 1 e 6 (não existindo sequências de comprimento 5), enquanto as sequências no CETENFolha variam entre 1 e 3.

2.3 Resultados

Participaram no exercício de anotação manual (ou automática com revisão) 9 participantes/anotadores. Na Tabela 2.4 encontra-se o nome dos participantes e das instituições a que pertenciam na altura.

Os resultados que a seguir se apresentam têm em conta as seguintes noções:

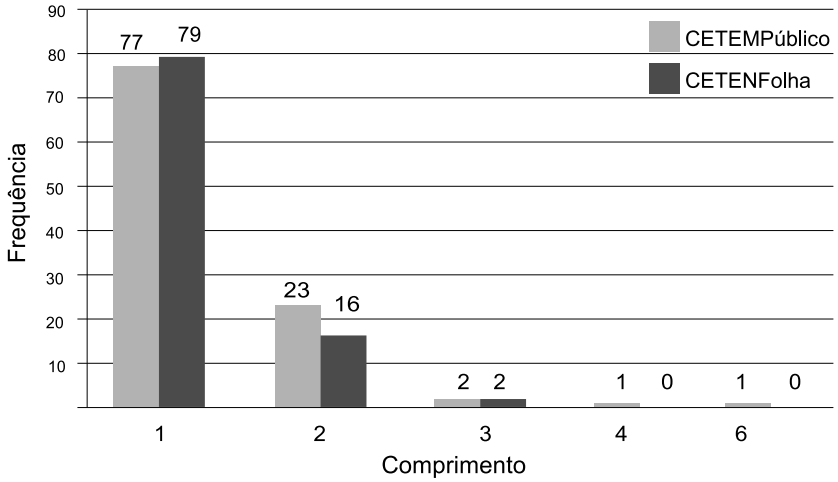


Figura 2.5: Número de ocorrências de sequências de palavras em maiúsculas de comprimentos *n*.

Identificador	Participante	Instituição
AS	Alberto Simões	Linguateca, Pólo do Minho
Prib	Cláudia Pinto	Priberam
CM	Cristina Mota	Linguateca, Pólo do LabEL
DS	Diana Santos	Linguateca, Pólo do SINTEF
EB	Eckhard Bick	Southern Denmark University
LO	Lucelia de Oliveira	NILC
Lab	Paula Carvalho	LabEL
RM	Raquel Marchi	NILC
VM	Vanessa Maquiafavel	NILC

Tabela 2.4: Participantes na tarefa de anotação.

- *entidade* corresponde a qualquer sequência delimitada com etiquetas SGML pelos anotadores;
- *nome próprio* corresponde a uma entidade marcada com a etiqueta `NOMEPROP`;
- *entidade (ou nome próprio) em comum* corresponde a uma sequência identificada por pelo menos um anotador, ou seja, uma sequência identificada consensualmente por um ou mais anotadores. Se para uma mesma sequência um anotador tiver identificado, por exemplo, *secretário de Estado* e outro tiver identificado apenas *Estado*, nenhuma das entidades contribuirá para o total de entidades em comum.

Foram calculadas três medidas de concordância na classificação:

- CE1: concordância relativa ao total de entidades em comum (ou seja, identificadas por pelo menos um anotador);
- CNP1: concordância relativa ao total de nomes próprios em comum (ou seja, identificados por pelo menos um anotador);
- CNPT: concordância relativa ao número total de nomes próprios identificados igualmente por todos os anotadores.

Foram tidos em conta os seguintes aspectos:

1. No caso de CE1 e CNP1, se um anotador não identificou uma entidade que outros reconheceram, essa entidade conta para o total de entidades em comum, mas não para o número de entidades em que há acordo;
2. Não se entrou em linha de conta com a subcategorização feita por alguns anotadores, ou seja, a concordância é relativa apenas à classificação feita usando o atributo `TIPO` ;
3. Dado que um dos anotadores propôs um conjunto bem variado de etiquetas que não contempla algumas das classes inicialmente sugeridas, estabeleceu-se a equivalência entre `ANTROPÓNIMO` e `PESSOA` e entre `TOPÓNIMO` e `LUGAR` (o estabelecimento desta última equivalência obrigou adicionalmente a substitutir a classificação das entidades marcadas originalmente por esse anotador como `LUGAR` por `LUGAR1`);
4. Ignorou-se igualmente que possa haver classes que são equivalentes por classificarem com nomes diferentes o mesmo conjunto de entidades (ou de nomes próprios), ou classes que possam estar completamente contidas noutras;
5. Não foram contabilizadas as entidades identificadas dentro das etiquetas SGML que já se encontravam nos extractos, uma vez que essas etiquetas correspondem a meta-informação estrutural do próprio corpus e como tal não deveriam ter sido analisadas[2].

2.3.1 Identificação de entidades

Como se pode ver na Figura 2.6, no CETEMPúblico foram identificadas de 81 a 106 entidades, enquanto no CETENFolha (Figura 2.7) o número de entidades identificadas variou entre 98 e 134. Destaca-se ainda que três dos nove anotadores identificaram exclusivamente nomes próprios, deixando sem marcação as expressões temporais e numéricas.

Combinando as entidades identificadas por pelo menos um anotador obtêm-se um conjunto de 140 entidades diferentes para o CETEMPúblico e de 163 para o CETENFolha. Desses conjuntos, respectivamente 63 e 70 entidades foram consensualmente identificadas

[2] Esta é uma das situações que mostra a falta de clareza nas instruções dadas aos anotadores.

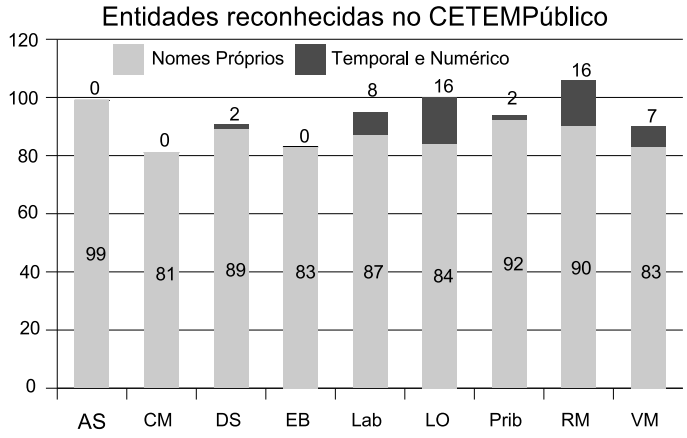


Figura 2.6: Total de entidades identificadas no CETEMPúblico por anotador.

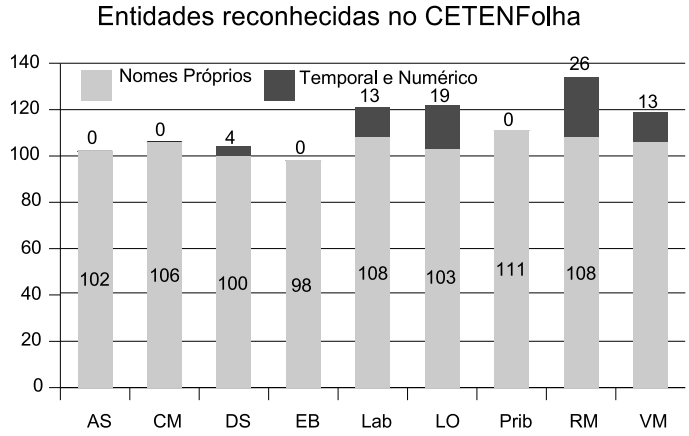


Figura 2.7: Total de entidades identificadas no CETENFolha por anotador.

por todos os anotadores, o que corresponde a 45% de concordância na identificação das entidades no CETEMPúblico e a 42,95% de concordância na identificação das entidades no CETENFolha. Se tivermos em conta apenas os nomes próprios então existe acordo na identificação em respectivamente 54,78% (63 em 115) e 56% (70 em 125) dos nomes distintos.

A lista das entidades comuns – ou seja, que foram identificadas por pelo menos um anotador e que não envolvem encaixe nem sobreposição com outras – e respectiva classificação encontram-se no apêndice B. Estas entidades correspondem a 67,86% (95 em 140) das entidades distintas do CETEMPúblico e a 74,85% (122 em 163) das entidades distintas

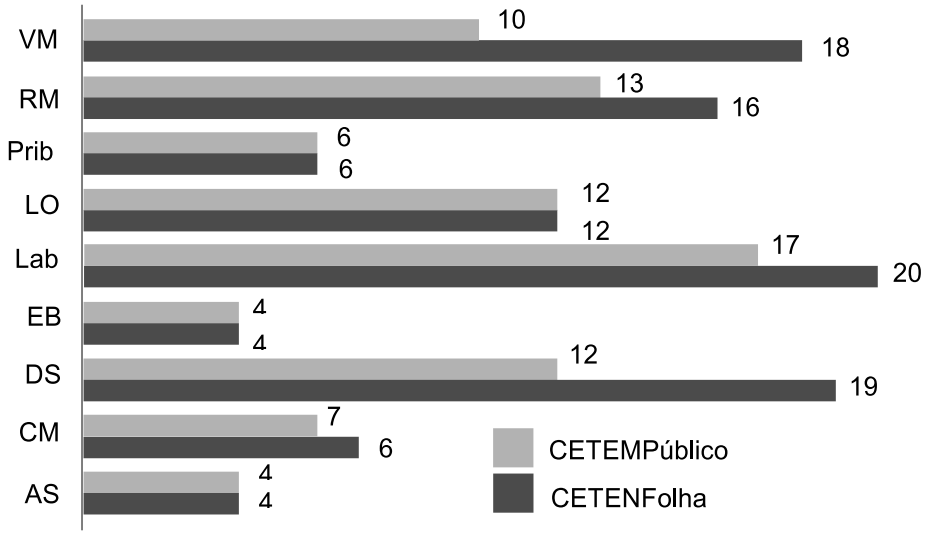


Figura 2.8: Total de categorias diferentes usadas por anotador.

do CETENFolha. O apêndice B também mostra as entidades para as quais não houve consenso na identificação, que também inclui as entidades que foram estruturadas (ou seja, têm outras entidades encaixadas. Apenas um anotador considerou este tipo de entidades.)

2.3.2 Classificação de entidades

Apesar do número de entidades ser bastante pequeno (cerca de uma centena), e de o número de categorias por anotador variar entre 4 e 20 (ver Figura 2.8, de facto, o número de diferentes categorias combinando as categorias de todos os anotadores é substancialmente elevado: 63 categorias no CETEMPúblico e 81 categorias diferentes usadas no CETENFolha. Esta variedade de categorias está bem patente em Mota et al. (2007, Figura 14.1) e que aqui se reproduz na Figura 2.9.

Naturalmente que, dada a variedade de etiquetas, a concordância quanto à classificação foi baixa (ver Tabelas 2.2 a 2.4). Note-se que os valores destas três tabelas não entram em consideração com as entidades que envolvem encaixe ou sobreposição com outras.

Se entrarmos também em consideração com os nomes próprios identificados por todos os anotadores que possam envolver encaixe ou estar sobrepostos com outros então obtemos 47,62% de concordância no CETEMPúblico (30 em 63) e 45,86% de concordância na classificação no CETENFolha (31 em 70).

Figura 2.9: União das etiquetas usadas no CETEMPúblico e no CETENFolha. Salienta-se a negrito as etiquetas originalmente propostas, gravitando à sua volta as etiquetas sugeridas pelos participantes.

	Entidades em comum (E1)	Entidades com mesma classificação	CE1
CETEMPúblico	95	30	31,58%
CETENFolha	122	30	24,59%

Tabela 2.5: Concordância na classificação das entidades comuns (CE1).

	Nomes próprios em comum (NP1)	Nomes próprios com a mesma classificação	CNP1
CETEMPúblico	79	30	37,97%
CETENFolha	98	30	30,61%

Tabela 2.6: Concordância na classificação dos nomes próprios comuns (CNP1).

	Nomes próprios identificados por todos os anotadores (NPT)	Nomes próprios com a mesma classificação	CNPT
CETEMPúblico	59	30	50,85%
CETENFolha	66	30	45,45%

Tabela 2.7: Concordância na classificação dos nomes próprios identificados por todos os anotadores (CNPT).

	AS	CM	DS	EB	Lab	LO	Prib	RM	VM	Média	Desvio Padrão
AS	100	89,47	93,42	90,79	92,11	86,84	93,42	93,42	86,84	90,79	2,63
CM	97,14	100	100	97,14	98,57	92,86	100	100	94,29	**97,50**	2,55
DS	95,95	94,59	100	95,95	97,3	93,24	100	100	93,24	96,28	**2,51**
EB	97,18	95,77	100	100	98,59	94,37	100	100	92,96	97,36	2,58
Lab	89,74	88,46	92,31	89,74	100	93,59	92,31	100	92,31	92,31	3,33
LO	81,48	80,25	85,19	82,72	90,12	100	85,19	95,06	85,19	85,65	*4,53*
Prib	95,95	94,59	100	95,95	97,3	93,24	100	100	93,24	96,28	**2,51**
RM	76,74	76,74	80,23	76,74	83,72	80,23	80,23	100	86,05	*80,09*	3,21
VM	89,19	89,19	93,24	89,19	97,3	93,24	93,24	100	100	93,07	3,73
Média	90,42	*88,63*	93,05	89,78	94,38	90,95	93,05	**98,56**	90,52		
Desvio Padrão	*7,25*	6,47	6,82	6,63	4,97	4,61	6,82	**2,53**	3,54		

Tabela 2.8: Acordo entre pares de anotadores na identificação das entidades no CETEMPúblico.

2.3.3 Quadros comparativos entre pares de anotadores

De modo a perceber até que ponto é que as entidades identificadas por um dado anotador são consensuais, calculámos em relação às entidades que cada um dos anotadores reconheceu a percentagem de entidades identificadas também por cada um dos outros anotadores. As Tabelas 2.8 e 2.9 apresentam esses valores para o CETEMPúblico e para o CETENFolha, respectivamente.

Por exemplo, a célula (CM,DS) na Tabela 2.8 indica que todas as entidades identificadas por CM foram igualmente identificadas por DS; a célula (DS,CM) na mesma tabela indica que das entidades identificadas por DS, 94,5% foram igualmente identificadas por CM. Isto significa que DS identificou todas as que CM identificou e mais algumas. A média e o desvio padrão de uma coluna dão uma ideia de quanto é que o anotador representado na coluna concorda com os anotadores representados nas linhas; a média e o desvio padrão de uma linha indicam quanto é que anotadores representados nas colunas concordaram com a anotação do anotador representado nessa linha. Ou seja, se o desvio padrão for alto para uma linha, isso significa que esse anotador é polémico, pois há uns anotadores que concordam mas outros que discordam muito dele; se o desvio padrão for alto na coluna, isso significa que o anotador discorda mais de uns anotadores do que de outros.

2.4 Comentários finais

Tal como já referido anteriormente, todos os resultados aqui apresentados, incluindo os textos marcados por cada um dos anotadores bem como as entidades integradas em concordâncias, ficaram públicos no sítio da Linguateca antes do encontro presencial ter decorrido.

Aquando dessa sessão, além de como chegar a um consenso quanto à escolha das categorias, foram ainda levantadas mais algumas questões, que ficaram também em aberto

	AS	CM	DS	EB	Lab	LO	Prib	RM	VM	Média	Desvio Padrão
AS	100	97,59	93,98	92,77	97,59	93,98	98,8	96,39	93,98	**95,64**	**2,08**
CM	86,17	100	91,49	91,49	100	95,74	100	97,87	95,74	94,81	4,5
DS	88,64	97,73	100	92,05	97,73	94,32	98,86	96,59	94,32	95,03	3,21
EB	88,5	89,9	93,1	100	98,9	96,6	100	98,9	95,4	95,16	4,09
Lab	79,41	92,16	84,31	84,31	100	95,1	92,16	98,04	95,1	90,07	6,15
LO	74,29	85,71	79,05	80	92,38	100	86,67	94,29	89,52	*85,24*	6,51
Prib	85,42	97,92	90,63	90,63	97,92	94,79	100	96,88	95,83	93,75	4,17
RM	69,64	80,36	74,11	74,11	86,61	83,93	82,14	100	88,39	79,91	6,23
VM	74,29	85,71	79,05	80	92,38	86,67	94,29	89,52	100	*85,24*	*6,51*
Média	69,73	79,65	74,08	85,67	83,08	80,57	81,62	83,7	81,61		
Desvio Padrão	27,09	30,73	28,76	6,61	31,66	30,73	31,4	31,74	30,96		

Tabela 2.9: Concordância entre pares de anotadores na identificação das entidades no CETENFolha.

para a futura realização da avaliação conjunta, nomeadamente:

1. Que sequências considerar como entidades mencionadas? Nomes próprios? Ou também expressões temporais e numéricas?

2. Deveria ser considerada a constituição interna das entidades permitindo a delimitação de entidades encaixadas noutras? Por exemplo, **<EM>** `Escola de Medicina de Harvard` **</EM>** versus **<EM>** `Escola de Medicina de` **<EM>** `Harvard` **</EM> </EM>**.

3. O que fazer com cargos, títulos e funções? Integrá-los na delimitação da entidade como em **<EM>** `Presidente Jorge Sampaio` **<EM>** ou ignorar, pretendendo-se `Presidente` **<EM>** `Jorge Sampaio` **<EM>**? Mas e se o cargo, por exemplo, não começar por maiúscula como em `major Carlos Barbosa`?

4. Atribuir-se-á a etiqueta em função do contexto? Compare-se por exemplo *(...) feira especializada que teve lugar em Basileia(...)* com *(...)chegará o dia em que a Rússia ajudará(...)*.

5. O que fazer quando não é possível decidir? Anotar ou ignorar?

Além disso, delineou-se um primeiro esboço dos passos a tomar na primeira avaliação conjunta de sistemas de REM, no sentido de continuar o trabalho iniciado com a experiência que relatámos:

1. Estabelecer o conjunto de etiquetas e regras de anotação a adoptar;

2. Realizar um nova anotação manual com os mesmos textos usando o novo conjunto de etiquetas, tendo se sugerido a utilização de uma ferramenta auxiliar de anotação, como por exemplo o Alembic Workbench (Day et al., 1997) que facilitaria não só o processo de anotação manual como também o de revisão e comparação das anotações;

3. Seleccionar e preparar os textos. Uma sugestão consistia em utilizar os mesmos textos que fossem utilizados na avaliação de recuperação de informação e sumarização automática, com o objectivo de ter um recurso reutilizável e mais rico;

4. Fazer uma pré-inscrição;

5. Propor um calendário para a avaliação.

Após quatro anos decorridos, penso que as conclusões mais salientes do presente ensaio foram que ele demonstrou indubitavelmente haver interesse da parte da comunidade, mas grande necessidade de consenso, o que talvez tenha motivado os organizadores a tomar uma atitude mais impositiva na condução da própria avaliação conjunta.

Capítulo 3

MUC vs HAREM: a contrastive perspective

Nuno Seco

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 3, p. 35–41, 2007.

This chapter presents a brief overview of two pioneering evaluation contests in the field of Named Entity Recognition (NER) and delves into the conceptual underpinnings of each. The intention is not one of divulging the referred events, as that has been done in Grishman e Sundheim (1996) and Santos et al. (2006), but rather one of contrastive scrutiny. The reader should be attentive of the fact that I am comparing two events that took place in completely different time frames. Notwithstanding, this comparison is relevant because both correspond to the genesis of the joint evaluation paradigm in the field of NER of two different languages, English and Portuguese, respectively.

The field of Natural Language Processing (NLP) has faced many obstacles since its birth. While some have been somewhat overcome, others still remain. One such obstacle is the **identification** and **classification** of named entities. It is in the classification facet of named entities that HAREM differs quite significantly from the Message Understanding Conferences (MUC) (Sundheim, 1995; Grishman e Sundheim, 1996; Hirschman, 1998). Nonetheless, there are evolutions of MUC contests (namely the Automatic Content Extraction (ACE) (Doddington et al., 2004) that address some of the shortcomings pointed out in this chapter. Arguably, many may refute the relevance of this paper because of the time gap between the two events; even so, the discussion is still appropriate as both correspond to the origins of the evaluation event in each language.

The reader should also take into account the fact that by comparing two evaluation events pertaining to two different languages certainly raises issues of authority of such comparisons as is pointed out in Cardoso (2006a, Section 5.3.3). Nonetheless, my concern is not one of comparing the results of the events but one of comparing the underlying assumptions and motivations of these events.

The rest of this chapter is organized in the following manner: Section 3.1 provides a brief overview of MUC, focusing on the aspects dealing with NER. Section 3.2 presents HAREM, contrasting it with MUC along with its guiding principles that motivated the construction of a new evaluation methodology. Section 3.3 presents the fine grained evaluation metrics employed along with their possible combinations. Finally, Section 3.4 concludes the paper summarizing the main differences identified.

3.1 An Overview of MUC

Prior to MUC, several Information Extraction (IE) systems were developed, but most of them were developed having specific domains in mind. Consequently, it was impossible to compare systems and strategies in a just way. As such, the need for a common evaluation environment that would enable fair comparison of systems was acknowledged. In order to quench the need, an informal poll of NLP groups was carried out to determine which groups had running text processing systems, and whether these groups would be interested in coming together to assess the state of NLP systems.

The first MUC event took place in 1986 (Grishman e Sundheim, 1996) and had the main goal of gathering researchers interested in the topic of IE. For the first time, a common corpus of real messages was used, from a common theme (the naval domain). System performance was compared using this common corpus, and the output of each system was discussed.

In 1989 a second MUC event took place and introduced the notion of template filling along with several evaluation metrics. In this edition, the participants had to fill templates that had several types of attributes with their corresponding values extracted from the given text. Introducing such templates and manually pre-calculating the correct values allowed, for the first time, the use of evaluation metrics such as precision, recall or F-measure to measure and compare the system's performances.

From 1991 up to 1993, MUC organized three more evaluation events. The main characteristics of these events was the change in target domains, the size of the corpus, the complexity of the templates and, finally, the inclusion of more languages such as Japanese, Spanish and Chinese.

MUC-6 took place in 1995 and had 3 main goals in mind:

1. Promote the development of reusable components that could be easily used in other NLP related tasks besides IE.

2. Promote, as much as possible, an effortless portability of systems to other domains for which they were not initially conceived.

3. Look into issues concerned with deeper understanding of the texts, such as anaphoric references and relations between attributes of different templates.

Thus, it was in the context of MUC-6 guidelines that NER was identified as being an autonomous component prone task and received diligent attention. MUC-7 took place in 1998 and did not diverge when compared to its preceding event, being that the basic difference was in the number of texts used in the contest.

3.2 Named Entity Recognition

Named entities, from a MUC viewpoint, were defined as: (Sundheim, 1995)

> *"... markables [named entities] includes names of organizations, persons, and locations, and direct mentions of dates, times, currency values and percentages. Non-markables include names of products and other miscellaneous names ('Macintosh', 'Wall Street Journal', 'Dow Jones Industrial Average') ..."*

This definition alone represents a major difference between HAREM and MUC, a discussion postponed to Section 3.3.

NER is considered to be domain independent and task independent, according to MUC's guidelines. The results obtained in MUC's NER task seem to suggest that NER is an easy task, with more than half of the systems obtaining results above 90% in terms of precision and recall (the best system obtained an F-measure of 0.9642).

Before accepting that the NER task is a solved case, one should address the issue of what exactly is being evaluated: The MUC-6 NER task used a golden collection of 30 articles taken from the Wall Street Journal (WSJ) from January of 1993 to June of 1994. MUC-7 used 100 articles from same collection. The named entities of this golden collection were manually identified and classified according to three different categories and subtypes (Sundheim, 1995):

1. `ENAMEX` – Entity names with subtypes organization, people and location.
2. `TIMEX` – Temporal expressions with subtypes date and time.
3. `NUMEX` – Numeric expressions with subtypes money and percent.

Summing up, the classification facet of NER in MUC evaluations was done according to the above mentioned categories. The next section discusses the HAREM evaluation and delineate the underlying conceptual differences in the evaluation.

3.3 HAREM

In HAREM, the classification system of MUC-6 was challenged, questioning its appropriateness to real applications, and if it really represents the NER issue. Note that the categories chosen for MUC were accomplished in a top down manner. On the contrary, HAREM took a bottom-up approach by manually analyzing text, identifying relevant entities and then attributing them a classification in context. As a consequence, a much finer grained classification hierarchy with 10 categories and 41 types was established (Santos e Cardoso, 2006):

1. **PESSOA**:INDIVIDUAL, CARGO, GRUPOIND, GRUPOMEMBRO, MEMBRO, GRUPOCARGO
2. **ORGANIZACAO**:ADMINISTRACAO, EMPRESA, INSTITUICAO, SUB
3. **TEMPO**:DATA, HORA, PERIODO, CICLICO
4. **LOCAL**:CORREIO, ADMINISTRATIVO, GEOGRAFICO, VIRTUAL, ALARGADO
5. **OBRA**:PRODUTO, REPRODUZIDA, PUBLICACAO, ARTE
6. **ACONTECIMENTO**:EFERMIDE, ORGANIZADO, EVENTO
7. **ABSTRACCAO**:DISCIPLINA, ESTADO, ESCOLA, MARCA, PLANO, IDEIA, NOME, OBRA

8. **COISA**:CLASSE, SUBSTANCIA, OBJECTO, MEMBROCLASSE
9. **VALOR**:CLASSIFICACAO, QUANTIDADE, MOEDA
10. **VARIADO**:OUTRO

Note: COISA:MEMBROCLASSE appeared only on 2006 event. In 2005, OBRA:PRODUTO was discarded.

These finer grained categories lead to a finer grained NER classification task, therefore making the HAREM NER task much more intricate when compared to MUC's task and of other events. Another important aspect that HAREM took into account was **context**, that is, the surroundings in which a named entity appears determines its meaning and, therefore, its category (or categories). For example, in MUC the term **Brasil** would be considered an ENAMEX regardless of the context it appeared in. On the other hand, HAREM dealt with the issue of sense extensions such as metonymy. Consequently, the term **Brasil** could be classified differently according to the surrounding context. Consider the following examples taken from Santos (2006a):

*O **Brasil** venceu a copa...* (PESSOA:GRUPOMEMBRO)
*O **Brasil** assinou o tratado...* (ORGANIZACAO:ADMINISTRACAO)
*O **Brasil** tem muitos rios...* (LOCAL:ADMINISTRATIVO)

In each example, the same term is classified according to the context it appears, an aspect not dealt by MUC. Nonetheless, ACE, for instance, takes this aspect into consideration (Doddington et al., 2004).

Another aspect, and probably the most distinctive aspect is that HAREM, takes vagueness into account during identification and classification. That is, the possibility of a named entity simultaneously being identified or interpreted according to different referents both of which are correct. The issue of vagueness is more carefully discussed in Chapter 4. Consider the following example:

*...era um teólogo seguidor de **Emmanuel Swendenborg**.*
(PESSOA:INDIVIDUAL or ABSTRACCAO:OBRA ?)

In this example, both interpretations are equally acceptable (the writings of the person or the actual person), and most probably they occur simultaneous in our conceptual system and discourse structure (Pustejovsky, 1994). For an in-depth discussion on vagueness in the realm of HAREM we refer the reader to Santos e Cardoso (2006). Nonetheless, MUC also allowed alternative identifications through the use of the ALT tag attribute, but regarding semantic classification was more conservative. For example, the MUC guidelines state that *the White House* should be marked up as ORGANIZATION or have no markup at all in the answer key. This is a highly conservative approach when compared to HAREM that allowed different categories to occur simultaneously.

3.4 Evaluation

In HAREM, a golden collection of 129 (and later another set of 128 different texts for the Mini-HAREM[1] event) texts manually tagged was used as the reference for evaluation purposes. The collection comprised several different text genres written according to several different language varieties, mainly from Portugal, and Brazil, but also from Angola, Mozambique, East Timor, Cape Verde, India and Macao. As well as identifying and semantically classifying named entities, HAREM took into consideration the gender and number of the entities, introducing two new facets of evaluation with subtypes. HAREM proposed 3 subtasks: Identification (correct delimitation of the entity), Semantic Classification and Morphological Classification (gender and number).

Each of these dimensions was evaluated using different configuration scenarios. These have been clearly explained in Chapter 18 and as such it will suffice to say that there are 12 different possible evaluation scenarios for the participant. The motivation for such flexibility is that many participants are only concerned with certain aspects of classification (e.g. only interested in the PESSOA category).

Another issue worth stressing is that the HAREM evaluation software deals with partial alignments. In other words, it can cope with inexact matches of named entities between source and target texts. This aspect was never considered in other evaluation events. A finer discussion of the evaluation aspects of HAREM may be found in Seco et al. (2006).

The metrics used in HAREM subsume the ones proposed and employed in MUC, HAREM introduced many new evaluation metrics (Cardoso, 2006a). Nonetheless, regarding the metrics that were employed in both, the results obtained were drastically different. The best system in the first HAREM event attained an F-measure of 0.63 (considering an evaluation configuration equivalent to that of MUC). At first sight this seems to indicate that the state of the art of NER for Portuguese is substantially inferior to that of English. But from another standpoint one may argue that it is not the quality of NER systems that is inferior to that of English, but that the evaluation standards are much more meticulous in HAREM, resulting in a more demanding task and yielding lower performance values. It is the author's belief that the last perspective correctly mirrors the reality of HAREM.

3.5 Final Remarks

In conclusion, HAREM has brought significant contributions to the field of NER, specifically regarding the Portuguese language, where previous work did not exist. A finer grained classification system has been proposed that was obtained using bottom-up analysis approach of actual corpora. Named entities were classified in context according the classification system proposed; the number of different interpretations in HAREM was con-

[1] The interested reader should see Cardoso (2006a) for details.

siderably larger than in MUC (see Chapter 4). Vagueness, a ubiquitous characteristic of language, was taken into account in the HAREM evaluation. Morphological classification (gender and number) was also considered for the first time in the field of NER. The golden collection employed and used in the evaluation process was substantially wider-ranging when compared to MUC. MUC used the Wall Street Journal, which can be considered a domain specific journal, while HAREM used documents from general newspapers in Portugal and Brazil, Web texts, literary fiction, transcribed oral interviews and technical text. Finally, the evaluation framework showed to be very powerful, fulfilling the assorted needs of the several participants in a very flexible manner.

Acknowledgements

I would like to thank Bruno Antunes, Diana Santos, Nuno Cardoso and Cristina Mota for their valuable comments and suggestions.

Capítulo 4

O modelo semântico usado no Primeiro HAREM

Diana Santos

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 4, p. 43–57, 2007.

Este capítulo fundamenta o modelo semântico desenvolvido para o Primeiro HAREM. Em primeiro lugar, são levantadas algumas questões de fundo sobre a semântica da linguagem natural, e a sua aplicação no caso específico do REM. Segue-se uma apologia relativamente longa, tentando rebater diversos contra-argumentos levantados em algumas ocasiões (como por exemplo o Encontro do HAREM), e justificando as bases teóricas do modelo adoptado.

Como mencionado no capítulo 1, a razão por que começámos a tratar de REM na Linguateca foi porque nos pareceu a tarefa mais básica possível a nível de semântica. Contudo, isto não significa que o REM seja propriamente uma tarefa fácil, ou que a maior parte das questões associadas ao PLN não acabe por surgir, quando se pretende delimitar rigorosamente o âmbito e o propósito desta tarefa.

4.1 O que é semântica?

Sendo um capítulo sobre a definição de uma tarefa semântica, é preciso começar por lembrar que não há realmente um grande consenso entre o que é a esfera da semântica, o que leva a que seja necessário que, até a esse nível!, estabeleçamos uma definição para que o capítulo possa fazer sentido.

Muito simplificadamente, a semântica ocupa-se da relação entre a forma (a língua) e o "mundo exterior à língua". Deixemos neste momento de parte a questão complexa de o que é este mundo e se ele existe realmente na Natureza ou apenas nas nossas mentes (ver Santos (1940)). Por outras palavras, a semântica tenta relacionar objectos linguísticos com objectos não linguísticos. Visto que o mundo em si não está acessível para as nossas análises, existe sempre um modelo ou conceptualização que medeia entre ele e a língua, ou seja, os investigadores em semântica constroem modelos que pretendem representar a realidade e tentam mapear a língua nesses modelos.

Devido à grande complexidade da tarefa, um mapeamento directo é raramente sugerido (mesmo quando se está a falar da relação entre a língua e um modelo conceptual parecido, como por exemplo a lógica de primeira ordem). As teorias semânticas recorrem a estruturas intermédias (como a DRT de Kamp e Reyle (1993)), a tipos especiais de raciocínio (como lógica não monotónica, ver Ginsberg (1987)) ou a representações especificamente desenhadas para emparelhar propriedades conhecidas da linguagem natural, tais como mundos possíveis (Hughes e Cresswell, 1968) para interpretar modalidade, ou guiões (Schank e Rieger, 1974) para fazer sentido de algum tipo de descrições esperadas.

Seja qual for a teoria que nos agrade mais, estou convencida de que ninguém discordará do seguinte: delimitar o conceito de entidade mencionada, como conceito semântico, tem a ver com a relação entre a língua e o mundo exterior à língua, mundo esse que é mediado/representado por um conjunto de símbolos que representam esse mundo. A tarefa

de REM, como qualquer tarefa semântica, passa por um conjunto de categorias, sobre as quais se tenta chegar a um entendimento partilhado.

Existem duas grandes escolas de análise semântica: a **denotacionalista**, onde os símbolos são um substituto de objectos exteriores, e a **funcionalista**, em que os símbolos representam a relação entre os objectos, ainda dentro da própria língua. Assim, uma parte importante do significado de um texto (ou sintagma, ou palavra) é a função que desempenha relativamente aos outros elementos do texto. Pode ver-se esta análise como mais um nível entre a língua (forma) e o mundo; em paralelo com a denotação, deve também ter-se em conta a função. (E a função é geralmente obtida de um conjunto de poucos valores, tais como os casos de Fillmore (1968)). Esta é uma forma de tentar explicar sistematicamente porque é que uma mesma expressão em contextos diferentes tem ou pode ter significados diferentes, que é uma das propriedades mais básicas e mais importantes da linguagem natural. Por outro lado, existe ainda outra escola a que chamarei **pragmática**, que defende que é o contexto que define o sentido, e que não há denotação fixa. Ou seja, as funções de cada elemento no texto dão-lhe um significado, juntamente com o contexto real de produção da frase.

Em qualquer caso, a análise semântica pressupõe sempre uma classificação em categorias, e essa classificação não é nada consensual na forma como é estruturada: são conjuntos baseados em semelhanças, ou em diferenças (Ellis, 1993)? Todos os membros de uma categoria são iguais, ou há membros mais fortes do que outros? Quais os limites e as relações entre as categorias? São mutuamente exclusivas ou, pelo contrário, hierarquicamente ou funcionalmente definidas?

Para não tornar este capítulo demasiado geral, vou apenas discutir estas questões na subtarefa de dar sentido aos nomes próprios, o REM. Antes disso, vou fazer uma digressão necessária pela questão da vagueza.

4.1.1 A importância da vagueza para a semântica

Um dos meus cavalos de batalha é a questão da vagueza na língua. Ao contrário de uma concepção bastante divulgada, que considera a vagueza como uma fraqueza da linguagem natural que deve ser reparada, reduzida ou pelo menos tratada (como doença), eu considero que a vagueza é uma das qualidades mais importantes e positivas da linguagem natural, que deve ser admirada e tratada (com respeito) de forma a não se perder o seu conteúdo.

Ao contrário de outras abordagens que apenas reconhecem o fenómeno da vagueza em ocorrências concretas da língua, eu considero que a vagueza existe tanto ao nível da competência como ao nível do desempenho, ou seja, quer globalmente como propriedade dos itens lexicais e das estruturas da língua (fora do contexto) – a competência – , quer ao nível da língua concreta, das frases em contexto — o desempenho.

Felizmente, existem vários linguistas e filósofos que partilham esta opinião, donde não é necessário começar por argumentar longamente sobre a necessidade de lidar com este tema. Basta-me remeter para maiores autoridades (Burns, 1991; Pustejovsky, 1995; Lakoff, 1987; Buitelaar, 1998; Cruse, 2004) que lidam com a vagueza, se bem que sob perspectivas diferentes, ou para outros textos meus (Santos, 1997, 2006d) que já tratem a vagueza em pormenor.

De forma a restringir o âmbito do presente capítulo, discutirei apenas a questão da vagueza associada a abordagens computacionais relacionadas com a formalização dos nomes próprios, portanto directamente relacionadas com a questão do HAREM.

4.2 O que é o REM?

Qualquer definição de REM depende fortemente do modelo semântico adoptado, e em particular, do seu lado extra-linguístico. No MUC definiram-se três conceitos principais que representam generalizações que se supunha existirem no mundo real: pessoas (PERSON), organizações (ORGANIZATION) e locais (LOCATION), e a tarefa de REM propunha reconhecer nomes próprios (uma restrição de forma) que apontassem ou correspondessem a essas categorias (fixadas de princípio) em textos jornalísticos escritos em inglês. Quando os nomes próprios encontrados no texto se referiam a outro tipo de entidades que não locais, pessoas ou organizações, não deviam ser reconhecidos, e assumiu-se que uma pessoa, uma organização e um local nunca poderiam coincidir (o que não é propriamente surpreendente).

No HAREM, nós estávamos interessados em **todos** os nomes próprios (definidos de forma bastante liberal), ao contrário de apenas um subconjunto de nomes próprios que tivessem uma dada denotação, para ter uma ideia do que a tarefa de REM significava para o português. Por isso começámos por tentar categorizar todos essas ocorrências em vários tipos de texto.

4.2.1 Metonímia

Porque não estávamos só à procura de casos simples, depressa nos demos conta do que muitos outros investigadores já tinham notado e formalizado antes de nós: que há muitos casos em que um nome originalmente usado para denotar um certo objecto é usado como substituto para outros objectos (ou entidades) que pertencem a um tipo completamente diferente. Por exemplo, em *Fontes próximas do Palácio de Belém desmentiram que...*, a entidade *Palácio de Belém* não se refere a um edifício, mas sim ao Presidente da República português, eventualmente secundado também pelo seu gabinete.

Ao contrário das opções que muitos seguiram, de formalizar e sistematizar essas substituições, nós adoptámos uma solução mais radical, ao marcar a entidade final de acordo

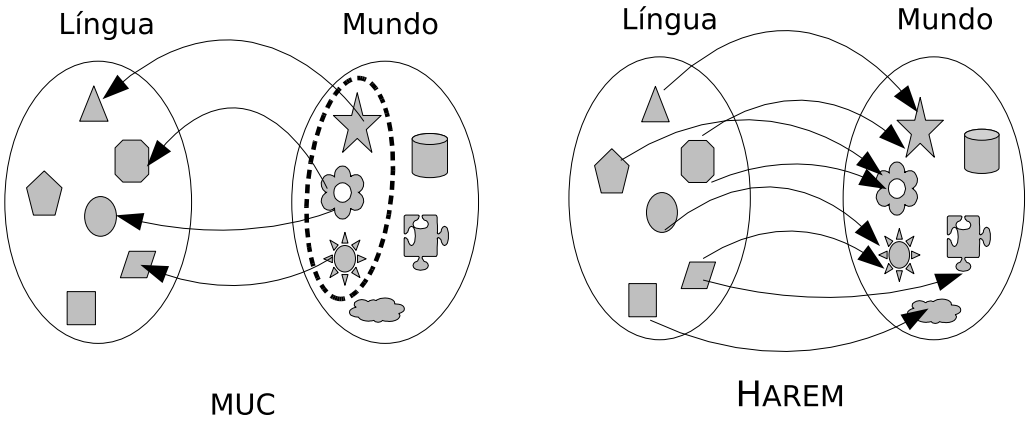


Figura 4.1: Dois pontos de partida diferentes para abordar a questão da semântica (do REM).

com o objecto denotado. (Que, no caso anterior, seria uma pessoa ou grupo de pessoas).

Este fenómeno é vulgarmente chamado **metonímia**, e pode ser definido como o caso em que uma expressão é usada para referir outro referente relacionado (veja-se Lakoff e Johnson (1980)). Exemplos conhecidos na literatura são o uso de *Vietname* para a guerra do Vietname, ou a *tosta mista* para o cliente que a encomendou, respectivamente nos seguintes exemplos:

> *Vietname nunca mais.*
>
> *A tosta mista queixou-se.* (dito por um criado ao cozinheiro do mesmo restaurante, e referindo-se, naturalmente, ao cliente que encomendou a tosta mista).

Qualquer pessoa que se debruce sobre a interpretação de nomes próprios em texto defronta-se com estes casos, muito comuns em textos jornalísticos. Markert e Nissim (2002) listam um número apreciável de padrões metonímicos associados a lugar ("place-for-event", "place-for-people", "place-for-product", etc.), assim como critérios detalhados para classificar "nomes de locais" nos vários padrões. Além disso, propõem uma organização hierárquica das metonímias, a existência de uma categoria "mista"[1] e o (implícito) reconhecimento de vagueza. Também Leveling e os seus colegas (Leveling e Veiel, 2006; Leveling e Hartrumpf, 2006) estudam a metonímia em termos de recolha de informação geográfica (RIG) em textos jornalísticos em alemão e concluem que, se retirarem os casos em que os "locais" são usados metonimicamente, obtêm resultados melhores no GeoCLEF[2]. Ou seja,

[1] Para tratar de casos como o seguinte exemplo (inventado), em que *Moçambique* aparece como local para a primeira oração, e como governo para a segunda: *quando chegou a Moçambique, que até essa altura se tinha mostrado contra as sanções, recebeu a desagradável notícia de que...*

[2] O GeoCLEF é uma avaliação conjunta de recolha de informação geográfica, integrada no CLEF (Rocha e Santos, 2007), e que desde 2005 inclui também o português (Gey et al., 2007; Mandl et al., 2007).

se o sistema só considerar casos em que os locais são mesmo locais, obtém maior precisão nos resultados relativos a tópicos geográficos.

Note-se que também os organizadores do MUC estavam conscientes deste fenómeno, embora as directivas do MUC ditassem que a marcação devia ser feita de acordo com a categoria original. Por outras palavras, o facto de um nome próprio em inglês ser usado numa construção metonímica "place-for-product" não impedia que fosse classificado como LOCATION. (Para sermos totalmente precisos, Chinchor e Marsh (1998, secção A.1.6) discute de facto os casos de metonímia, dividindo-os entre "proper" e "common", e trata-os diferentemente, mas sem explicar a razão.)

No HAREM optámos precisamente pela abordagem oposta: em casos de "place-for-product", o nome próprio seria marcado como PRODUCT, e não como LOCATION.

Mais do que isso, e dada a nossa aderência ao modelo da vagueza como propriedade fundamental da língua, o modelo semântico que abraçámos não recorre a metonímia, mas sim a vagueza de nomes próprios, que podem (e costumam) ter mais de uma interpretação associada.

Veja-se o caso mais comummente discutido, o nome de um país. Na minha opinião, o conjunto de interpretações "pessoas/povo, governo administrativo, local, e cultura/história" em relação a um nome de país fora de contexto é indissociável.

Ou seja, os usos mais prototípicos de *país* incluem ou podem incluir todas estas vertentes. É certo que, em alguns casos, um país/nação pode não ter um local, ou não ter um governo reconhecido, ou não ter mesmo ainda uma cultura/história.[3] Por outro lado, o facto de conter estas quatro (e mais) vertentes no seu significado não quer dizer que o seu nome não possa ser usado apenas numa vertente (em particular apenas como lugar, mas não especialmente como lugar), como ilustram os seguintes exemplos:

> *Portugal orgulha-se dos descobrimentos.* (história/cultura)[4]
> *Portugal tem um clima ameno* (local, geografia física)
> *Portugal tem uma taxa de natalidade baixa* (povo)
> *Portugal decretou o feriado do Carnaval.* (governo, administração, geografia política)

A diferença fundamental entre a abordagem de Markert e Nissim e de Leveling, por um lado, e a abordagem do HAREM, por outro, é que a primeira considera que primariamente países ou cidades são locais, e só por um processo mais complicado (metonímia) deixam a sua interpretação "básica" e passam a exprimir outras coisas não básicas, enquanto que a segunda abordagem, seguida no HAREM (assim como por outras correntes de semântica

[3] Só não consigo imaginar um país deserto, ou seja, que nunca tenha tido pessoas.

[4] Agradeço ao Nuno Cardoso o exemplo mais garantidamente histórico/cultural de: *A influência de Portugal foi grande no Japão.* Mantive, contudo, o exemplo original por causa da argumentação que se segue, à qual convém o uso de *Portugal* como sujeito.

mencionadas acima, muito particularmente a de Pustejowsky[5]) não privilegia a interpretação local em relação às demais interpretações. Como argumento para não privilegiar a vertente lugar, note-se que todos os casos mencionados acima podem ser anaforicamente relacionados usando a palavra país, mas não usando a palavra *local* ou *lugar* (só o segundo):

Portugal é um país com tradições (ou *é um país que se orgulha dos seus Descobrimentos*)
Portugal é um país de clima ameno.
Portugal é um país com taxa de natalidade abaixo de...
Portugal foi o único país da EU que decretou feriado na terça feira passada.

A perfeita aceitabilidade de *Portugal, local de sonho para muitos turistas, orgulha-se dos seus Descobrimentos* foi apresentada por Cristina Mota (c.p.) como uma prova de que a palavra *Portugal*, mesmo noutras acepções/vertentes, pode ser identificado como *local*. Eu discordo. Para mim, o autor da frase apenas está a ligar duas vertentes num argumento que se espera coerente, e não a referir-se à segunda vertente como `LOCAL`.

Voltando ao REM, o HAREM requer uma distinção entre `ABSTRACCAO`, `LOCAL`, `PESSOA` (povo), `PESSOA` (governo), ou mais do que uma vertente simultaneamente, ao contrário do MUC, que classificaria todos os casos acima como `LOCATION`.

No modelo semântico subjacente ao HAREM, portanto, a palavra *Portugal* não significa imediatamente um lugar. O contexto no qual o nome *Portugal* se insere é vital para seleccionar a vertente da palavra. Além disso, e aqui está a importância da vagueza para o modelo, pode muitas vezes significar mais do que uma única vertente. Se apenas classificássemos *Portugal* como `País`, que é uma alternativa por vezes sugerida (que será debatida mais abaixo), ficava muito por compreender. E se classificássemos `País` como (apenas) `Lugar` (como se fez no MUC), estávamos a deitar fora mais de metade do significado de *Portugal*.

4.2.2 REM como aplicação prática

Ninguém discorda que, para determinar a vertente semântica em que é empregue qualquer expressão, é preciso compreender o texto em questão, e que haverá diferentes casos em que um utilizador estará interessado em diferentes vertentes de um mesmo conceito (por exemplo, política portuguesa contemporânea vs. aspectos da natureza em Portugal).

Poucos compreendem, contudo, que isso significa que ao nome *Portugal* não pode então ser associada sempre a mesma classificação se se quer distinguir entre as várias vertentes.

[5] Pustejovsky (1995) sugere um mecanismo complicado de formalização semântica, estruturado em quatro eixos/estruturas (argumental, de acontecimentos, de modos de explicação (*qualia*), e de herança lexical), separando além disso o que ele chama tipos unificados (*unified types*) e tipos complexos (*complex types*). Pustejovsky (1995, p. 141–157) analisa, por exemplo, *book* e *newspaper* como tipos complexos `informação.matéria-impressa`, sendo além disso *newspaper* um tipo complexo (`informação.matéria-impressa`)`.organização`. Conforme o contexto, um dado texto pode referir-se a modos particulares de explicação (os quatro que ele considera são forma, propósito, constituição e criação), ou a mais do que um desses modos.

Ou seja, se o REM pretende ajudar os sistemas e as pessoas a distinguir entre diferentes significados, é preciso que estejam separados e não aglomerados.

Questões a que o modelo semântico do HAREM (com a correspondente criação da colecção dourada) permite responder é, por exemplo, quantas vezes é que a palavra *França* foi utilizada na acepção LOCAL – ao contrário da pergunta, a que o MUC responde, de quantas vezes é que *França* (ou *France*) foi usada como um país (assumindo que os países são considerados LOCATION no MUC).

Em ambos os casos, não se está a entrar em conta com *França* quando classifica pessoas, ou organizações (fábrica de roupa, ou de sapatos), claro. Por isso é que ambas as tarefas são correctamente compreendidas como análise semântica dos textos, visto que requerem mais do que uma classificação de acordo com um dicionário de nomes próprios ou almanaque. Note-se contudo que se for a pastelaria França a referida na frase *Encontramo-nos hoje às duas na França*, neste caso *França* seria classificada como um LOCAL no HAREM, e como ORGANIZATION no MUC.

4.2.3 REM como classificação semântica tradicional

É muitas vezes apresentado como alternativa ao REM, ou como outro modelo de REM, a classificação directa dos nomes próprios nas classes mais descritivas que os compõem, tal como país, artista, político, monumento, jornal, para evitar escolher em que vertente cada uma destas classes deverá ser colocada. Ou seja, livro é um objecto ou uma obra de arte? Jornal é uma organização, um porta-voz, ou um papel? País é um lugar, um povo, ou um conceito? Não interessa, dizem os defensores deste modelo, o que interessa é ter classificado um nome próprio como Livro, ou Jornal, ou País.

Isto na minha opinião é simplesmente escamotear o problema. Primeiro, porque acaba por não se atribuir uma classificação segundo uma grelha pré-determinada mutuamente exclusiva (como é o caso da divisão do MUC entre LOCATION/PERSON/ORGANIZATION ou da categorização do HAREM com 10 grupos). O REM deixa assim de ser um problema especificável *a priori*, porque em princípio há um número infinito de classes a que cada expressão pode ser atribuída. E ainda há outra objecção importante, relacionada outra vez com a vagueza essencial da língua, que é mais facilmente compreendida por um exemplo. Atentemos nas seguintes frases:

> *Património de Sintra ameaçado por construção selvagem*
> *Freixo de Espada à Cinta atrai turismo com festival de música*
> *Douro com problemas de poluição*

Todos os "lugares" com nome podem ser empregues para denotar um conjunto de pessoas, uma cultura, etc., mas exactamente que tipo de lugar (ou entidade) referida não é geralmente tornado explícito na comunicação, porque não é necessário. Nas frases acima,

Sintra refere-se a concelho, a vila, ou a serra? *Freixo de Espada à Cinta* descreve a cidade ou a região? E *Douro* é o rio, a região, ou a população ribeirinha?

Ou seja, não é claro que classificações semânticas se devem atribuir a estas entidades mencionadas (`Concelho`, `Vila`, `Serra`, `Cidade`, `Região`, etc.?), bem como continua a não ser óbvia qual a acepção (ou vertente) em que elas são usadas nos contextos dados (`Rio` não parece poder nunca englobar a população ribeirinha desse mesmo rio, embora para `País` isso pareça ser aceitável).

Isto demonstra que a opção de classificar as EM segundo os seus tipos semânticos imediatos (`País`, `Rio`, `Cidade`, etc.) causaria mais problemas do que os que resolveria.

Na minha opinião, a maior objecção a este modelo é que, em muitos casos, senão na sua esmagadora maioria, o falante não quer decidir se se está a referir à cidade, à serra ou a todo o concelho... quanto mais a pessoa que recebe a informação e não sabe o que passa (ou passou) na mente do falante. *Sintra*, na maior parte das vezes, é vaga entre as três interpretações "cidade", "população" e "serra".

4.3 O ACE como uma alternativa ao MUC: outras escolhas

Para que fique mencionado, a inspiração do HAREM foi o MUC. Não nos debruçámos na altura suficientemente sobre o ACE (Doddington et al., 2004), convencidos de que representava um estádio mais elevado, demasiado complexo para principiantes na tarefa do REM.

Agora, estou convencida de que foi um erro grave não termos estudado aturadamente o processo seguido no ACE, pois parece que, em paralelo, chegámos independentemente a muitas conclusões semelhantes, embora também enveredado por caminhos diferentes.

Comecemos por salientar que a questão da metonímia (ou várias vertentes de, principalmente, nomes de lugares) foi resolvida no ACE através da introdução da categoria "locais geopolíticos" (para países ou cidades que são comummente mencionadas como actores políticos). Esta é uma forma um pouco original de lidar com a questão da vagueza na língua, mas apenas neste caso particular (criando a categoria `LOCAL+ORG`, que pode além disso ser especializada através da escolha de uma das possibilidades).

Segundo a interpretação de Maynard et al. (2003a), repetida em Cunningham (2005), o ACE teve a intenção de melhorar o processo seguido pelo MUC de uma forma semelhante ao HAREM: nas palavras de Maynard, em vez de análise "linguística", tentaram uma análise "semântica": *where the MUC task dealt with the **linguistic** analysis, ACE deals with its **semantic** analysis*. Nas palavras do ACE (Doddington et al., 2004, p. 838), este está interessado no reconhecimento de "entidades, não apenas nomes" (*entities, not just names*). Pese embora a imprecisão desta terminologia (que opõe linguístico a semântico), o que eles querem dizer é que o MUC partiu da forma, e o ACE do conteúdo (denotação). Algo surpreendentemente, até mencionam a versão inglesa da nossa terminologia: *In ACE these*

names are viewed as ***mentions*** *of the underlying entities.* Não podíamos ter confirmação mais evidente para a nossa escolha de nome em português, nem demonstração mais óbvia de que o HAREM e o ACE identificaram o mesmo problema no MUC. Contudo, abordaram-no de uma forma diametralmente oposta.

O problema do MUC, que refinamos aqui, é que partia de uma definição arbitrária com base nos dois campos ligados pela semântica (a língua/forma, e o conteúdo/denotação), delimitada por um subconjunto deste último: a tarefa do MUC tinha como alvo nomes próprios (forma) com significado de organização, local, etc. (denotação), como se aprecia nas palavras de Chinchor e Marsh (1998): «*the expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times [...] and quantities [...] The task requires that the system recognize what a string represents, not just its superficial appearance*».

O ACE escolheu o **lado do conteúdo** e pediu para -– independentemente da forma — os sistemas marcarem tudo o que fosse organização, local, pessoa, etc., sem restrições de forma (podiam ser realizados linguisticamente como substantivo, pronome, nome próprio, sintagma nominal, etc.).

O HAREM, ao contrário, **escolheu o lado da forma**: partiu de tudo o que é nome próprio em português (ver capítulo 16) e pediu para os sistemas identificarem e classificarem – sem restrições de sentido numa primeira fase, mas, depois de um estudo empírico inicial – com base na classificação proposta pela organização. (Note-se, no entanto, que aceitamos uma categoria `OUTRO`, ou seja, não garantimos que todas e quaisquer ocorrências de nomes próprios no texto podem ser enquadrados no produto cartesiano das categorias do HAREM.)

A parecença entre as duas extensões ao MUC (ambas reconhecem o MUC como inspiração) é também visível no aumento da variedade em tipo de textos: em vez de alargar em género como fizemos no HAREM, contudo, o ACE alargou em qualidade de texto ou meio de obtenção desse texto. Além de notícias impressas, usou textos obtidos a partir de reconhecimento óptico, e de reconhecimento automático de fala. Também alargou o assunto (em vez de um único domínio, passou a ter notícias sobre vários domínios ou assuntos). Interessante que, no caso do HAREM, usámos a extensão em termos de variante e sobretudo de estilo/género textual, alargando em termos de meio ou de qualidade apenas quando tal derivava de um género textual diferente: em particular, para cobrirmos a Web, tivemos de incluir textos de pouca qualidade, e para incluir entrevistas, tivemos de recorrer à transcrição da linguagem oral.

Outra semelhança entre o ACE e o HAREM foi o aumento significativo da complexidade na anotação humana, que, de acordo com (Maynard et al., 2003a), atingiu apenas 82,2% de consenso no ACE.

Outra diferença em relação ao MUC partilhada pelo HAREM e pelo ACE é a utilização neste último duma métrica baseada em custo (Maynard et al., 2002), que, embora mais geral do que a do HAREM, tem pontos de semelhança com a medida da classificação se-

mântica combinada do HAREM, permitindo a quantificação de uma dificuldade *a priori*.

Contudo, há diferenças entre o ACE e o HAREM, que nos impedem de rotular este de "o ACE português", mesmo de forma aproximada.

Em primeiro lugar, o ACE mistura a tarefa de reconhecimento de EM com a de reconhecimento de co-referências, o que significa que a forma de avaliar a identificação e/ou classificação é diferente. Desse ponto de vista, o HAREM emparelha com o MUC, ao separar (e no caso do HAREM, ignorar) a tarefa de co-referência da da identificação.

Mas a distinção mais importante é filosófica mesmo, e está relacionada com o tema principal do presente texto: o ACE exige uma única resposta correcta (através da possível criação de categorias vagas, tal como as entidades geopolíticas ou as instalações), enquanto no HAREM estamos interessados, não numa cristalização oficial dessas categorias, mas na detecção empírica de todas as perspectivas possíveis oferecidas pela língua. Ou seja, em vez de resolver o problema da vagueza do lado da organização com categorias fixas codificando essa vagueza (ou tipos complexos, na terminologia de Pustejowsky) aceitámos *a priori* qualquer conjunto de categorias como sendo representável pelo HAREM e que os anotadores decidiram atribuir como tal no contexto.

Para sermos completamente justos, convém realçar, mais uma vez agradecendo à Cristina Mota por nos ter tornado cientes desse facto, que o ACE permite, opcionalmente, a marcação da vertente (local, pessoa, organização) para as entidades geopolíticas[6]. Embora isso seja uma forma de resolver (para um conjunto limitado) a questão das múltiplas vertentes, parece-nos que a diferença é maior que a semelhança: por um lado, no HAREM não é só a categoria `<LOCAL | ORGANIZACAO>` que pode ser vaga, mas todas; por outro, quando uma expressão é só `LOCAL`, deve ser marcada como tal no HAREM, e não duplamente como "`<LOCAL | ORGANIZACAO>` vertente `LOCAL`", como no ACE.

4.4 A abordagem do HAREM como processamento da linguagem natural em geral

Um modelo conceptual ingénuo de um reconhecedor de EM é concebê-lo como um sistema com listas de nomes próprios previamente classificados (um almanaque) que atribui essa classificação quando os nomes se encontram no texto. E, de forma igualmente ingénua, se pode conceber que é esse o papel de um dicionário na análise sintáctica computacional.

De facto, dado o peso e relevância do contexto, não é preciso que as mesmas categorias se encontrem em ambos os lados da análise (ou seja, tanto no dicionário como no resultado da análise (sintáctica) de texto, ou tanto no almanaque como no resultado da análise semântica do texto), embora tenha de haver uma maneira de se fazer a ponte.

[6] Entidades semelhantes, tais como o marcador `<civ>`, são chamadas *híbridas* por Bick no capítulo 12. No HAREM são simplesmente codificadas através do operador |, ou seja `<LOCAL | ORGANIZACAO>`.

Concretizando: num almanaque, faz sentido que esteja armazenada a informação de que *França é um país*, mas na gramática de REM daria jeito que estivesse "um país pode ser um local, um povo, etc...". Tal qual como num dicionário pode estar que "*perfeito* é um adjectivo", mas na gramática terá de estar (ou seria desejável que estivesse) "um adjectivo pode ser usado como um substantivo, como um pós-nominal, como um pré-nominal, como uma parte de um composto, como uma exclamação, etc.", de forma a compreender ocorrências como, respectivamente, *o perfeito é inacessível, um perfeito disparate, um casal perfeito, amor-perfeito, Perfeito!*, etc. (vejam-se mais exemplos em Santos, 2006a).

Estamos pois a defender neste capítulo implicitamente um nível intermédio de processamento, ou melhor, uma forma de fazer PLN mais dirigida pelo contexto e menos pelo léxico. Em última análise, o desenvolvimento do sistema de REM fica ao critério do seu autor e dos seus objectivos, e muitos investigadores provavelmente quererão codificar "tudo" num dicionário ou num almanaque. No entanto, é importante salientar que estas duas abordagens são possíveis e que, a nível teórico pelo menos, uma não tem prioridade sobre a outra.

Note-se também que toda a discussão neste capítulo até agora tem sido sobre vagueza, ou seja a possibilidade de diversas interpretações simultaneamente. Outro assunto diferente é a ambiguidade, que talvez convenha também mencionar, para prevenir mal-entendidos em relação ao que até aqui se expôs.

Um caso claro de ambiguidade em REM é o seguinte: *Washington* representando o governo americano (ou o conjunto das pessoas correntemente fazendo parte do governo americano) e *Washington* representando o primeiro presidente dos Estados Unidos. Embora ambas sejam classificadas como `PESSOA` no HAREM: nome de uma cidade (capital) como menção a um grupo de pessoas pertencendo a entidade governativa[7] (`<PESSOA TIPO="GRUPOIND">`) e nome de uma pessoa que deu (por acaso) origem ao nome dessa mesma cidade (`<PESSOA TIPO="INDIVIDUAL">`), deve ser claro que qualquer texto em contexto ou se refere a um ou a outro.[8]

Igualmente no caso do adjectivo acima exemplificado, o facto de haver um substantivo ou adjectivo com o sentido de "categoria verbal" implica que *perfeito* é ambíguo, mas que, em qualquer contexto, ou significa uma ou outra das acepções.

É pois sempre preciso resolver a ambiguidade (isto é, escolher uma das várias opções mutuamente exclusivas), o que é uma tarefa completamente diferente de lidar como deve ser com a vagueza, que significa preservar vários sentidos relacionados.

[7] Veja-se o capítulo 16, para a distinção entre *Washington* como governo (`<ORGANIZACAO TIPO="ADMINISTRATIVO">`) e como grupo de pessoas pertencentes ao governo (`<PESSOA TIPO="GRUPOIND">`).

[8] Convém contudo notar que a medida de classificação semântica por categorias (capítulo 18) não permite entrar em conta com o facto de que as duas interpretações de *Washington* como `PESSOA` correspondem a dois sentidos diferentes. Um sistema que tivesse feito a escolha errada (entre as duas `PESSOA`s) não seria por isso penalizado no HAREM.

4.5 Alguma discussão em torno do modelo de REM do Primeiro HAREM

Um dos argumentos apresentado contra o modelo utilizado no HAREM, expressamente vocalizado durante a sessão plenária do Encontro do HAREM e também já presente em Bick (2006a,b), é a questão da relação entre o significado "intrínseco" (aquele que aparece num dicionário, sem contexto) de um nome próprio, e o papel que esse nome próprio desempenha em contexto. Segundo Eckhard Bick, ambos são necessários e devem ser marcados, mas a ligação com a realidade (se é rio, se é cidade, se é país) está no dicionário, e o resto provém da interpretação sintáctico-semântica, em termos mais gerais, na forma de papéis semânticos como os propostos originalmente por Fillmore (1968) (agente, paciente, direcção, instrumento, etc.). Nesta perspectiva, a conjugação dos dois tipos de informação permite inferir o que estamos a anotar no HAREM: `País` + `Agente` = `Governo` (ou `Povo?` ou `Equipa`); `País` + `Instrumento` = `Governo`; etc.[9]

É preciso, contudo, confirmar na prática se de facto se consegue: i) definir consensualmente os papéis semânticos necessários (algo que até agora não parece ter sido possível) e aplicá-los a texto real de forma satisfatória; e ii) definir uma álgebra que dê de facto as mesmas (ou mais satisfatórias) distinções do que as empregues no HAREM.

Admitindo que tal seja possível, ou seja, que usar uma classificação composta por um papel semântico genérico mais um conjunto de marcações específicas no léxico consegue produzir o mesmo que o HAREM procurou atingir, penso que tal funcionará mais como uma demonstração de que o nosso modelo de interpretação dos nomes próprios no HAREM é apropriado, do que como uma crítica ao nosso objectivo. Parece-me que esta posição — que é baseada num modelo de como fazer REM — acaba por redundar em mais um argumento a favor da anotação usada no HAREM para avaliar o REM.

4.6 Outros trabalhos

Uma distinção semelhante fora feita já por Johannessen et al. (2005) no âmbito de uma comparação entre vários sistemas para línguas nórdicas. Estes autores discutem a definição da tarefa de REM, identificando duas estratégias distintas, que baptizam como "forma com prioridade sobre a função" e "função com prioridade sobre a forma"[10]. A primeira estratégia pretende identificar formas com uma dada denotação, independentemente da sua função em contexto; a segunda pretende identificar um conjunto de funções com base no contexto. Talvez devido ao grande número de autores, a conclusão do artigo é de "indecisão quanto à estratégia preferível" (p. 97). Mais interessante é a afirmação de que os sistemas que dão prioridade à função são mais robustos em relação à diminuição drástica do tamanho dos almanaques. Não fica, contudo, claro como é que os autores podem

[9] Por `Governo` estamos aqui a abreviar a notação do HAREM correcta, que seria `<ORGANIZACAO TIPO="ADMINISTRACAO">`.
[10] Em inglês, *form over function* e *function over form*.

comparar os sistemas com uma mesma avaliação se de facto a tarefa a que se propõem é diferente nos dois casos.

4.7 Comentários finais

Este capítulo tentou apresentar o modelo semântico pressuposto pelo Primeiro HAREM, quer através da aplicação básica de conceitos semânticos genéricos ao REM, quer através de uma comparação detalhada com os modelos respectivos do MUC e do ACE.

Os dois pressupostos mais importantes dizem respeito à importância do contexto na interpretação, e à ubiquidade da vagueza na linguagem natural.

Contudo, o capítulo é profundamente teórico no sentido de não fornecer dados empíricos sobre a extensão das diferenças entre os modelos apresentados, e sugere imediatamente algumas tarefas que propomos também no capítulo 7.

Com efeito, seria muito interessante anotar a colecção dourada do HAREM com uma marcação estilo MUC (cujas directivas para o português ainda estão contudo por fazer a um nível de detalhe suficiente) e depois medir e analisar objectivamente os resultados: quantas vezes é que a diferença entre os modelos implicaria diferença a nível da classificação final?

Outra medição é a da dificuldade da tarefa proposta pelo HAREM, quer a nível de concordância entre anotadores, quer a nível de dispersão intracategorial e intercategorial dos nomes próprios em português. É preciso quantificar quantas EM são ambíguas e/ou vagas tanto em potência (no almanaque ideal) como na realidade (aproximada), em texto. Se fizermos uma nova anotação no estilo MUC, poderemos ter ainda outra medida da diferença entre a dificuldade das duas tarefas.

Um outro trabalho natural como continuação do HAREM é comparar os resultados obtidos por Markert e Nissim (2002) para o inglês analisando 2000 EM, com os resultados das colecções douradas do HAREM (mais de 9000 EM), investigados sob a perspectiva da metonímia.

Finalmente, talvez a questão mais interessante para uma teorização mais rigorosa do REM será investigar a redutibilidade de um problema ao outro. Será que do "tipo MUC" mais papel semântico se pode derivar o "tipo HAREM"? Será que do "tipo HAREM" mais o papel semântico de uma população poder-se-á inferir o "tipo MUC"?

Esperamos poder um dia vir a responder a estas perguntas, com a ajuda da comunidade reunida em torno do Segundo HAREM, visto que a criação de recursos dourados e a sua disponibilização não deve morrer com a comparação na primeira avaliação conjunta, mas sim produzir matéria prima para muitos estudos empíricos e mesmo futuras avaliações.

Agradecimentos

Este capítulo deve um número apreciável de agradecimentos: em primeiro lugar, a todos os presentes no Encontro do HAREM no Porto pelo interesse e entusiasmo dos debates, que tiveram uma influência decisiva na concepção do presente texto; em segundo lugar, aos meus colegas na organização do HAREM sem a qual não estaríamos aqui, muito em particular ao Nuno Cardoso com quem revi em conjunto toda a colecção dourada e, como tal, partilhei muitas horas dedicadas à compreensão das EM em texto em português.

Além de um agradecimento natural a todos os participantes no HAREM, é forçoso salientar, admirar e agradecer a postura do Eckhard Bick e da Cristina Mota, que participaram segundo as normas do HAREM apesar de, desde o início, terem discordado dessas normas no que se refere precisamente ao modelo semântico utilizado.

No que se refere ao presente texto, tenho de agradecer especialmente a revisão cuidada e as muitas sugestões de melhoria da Cristina Mota, do Nuno Cardoso e do Jorge Baptista em relação a versões anteriores, que levaram a uma reescrita quase completa do capítulo. Gostava também de mencionar o entusiasmo genuíno e sempre carente de mais fundamentação que foi exibido pelo Nuno Seco em relação ao modelo do HAREM, quando chegou, mais tardiamente, à organização do mesmo, como aliás é patente no capítulo anterior. Ele foi assim, embora inconscientemente, um dos inspiradores do presente texto.

Finalmente, este capítulo foi escrito integralmente no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC 339/1.3/C/NAC.

Capítulo 5

Validação estatística dos resultados do Primeiro HAREM

Nuno Cardoso, Mário J. Silva e Marília Antunes

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 5, p. 59–77, 2007.

Nos últimos tempos, tem havido um aumento do número de conferências dedicadas à avaliação de sistemas inteligentes, sobretudo no que respeita às suas capacidades de Processamento de Linguagem Natural (PLN). Cada conferência organiza periodicamente eventos de avaliação de tarefas específicas, que têm contribuído significativamente para melhorar a eficácia dos sistemas na resolução de vários problemas específicos de PLN.

As tarefas de avaliação realizadas no HAREM recriam ambientes de avaliação onde os diversos factores que podem influenciar a medição podem ser minimizados, controlados ou mesmo eliminados. Ao garantir que os desempenhos dos sistemas são calculados segundo um ambiente e critérios de avaliação comuns a todos, torna-se possível realizar uma comparação justa e imparcial entre sistemas.

Nas avaliações conjuntas, os resultados obtidos mostram quais as melhores estratégias e fornecem dados importantes para a compreensão do problema. Contudo, aos resultados obtidos vem sempre associada uma margem de erro, relacionada com a aproximação que a tarefa tem ao problema real. Neste aspecto, as colecções de textos usadas nas tarefas de avaliação têm suscitado algumas reticências:

- Certas colecções de textos, como a *web*, são muito difíceis de representar numa amostra estática. Esta dificuldade está relacionada com a diversidade de assuntos, formatos, autores e estilos de escrita, ou a volatilidade dos seus conteúdos (Gomes e Silva, 2006). Como saber se a colecção de textos usada é uma amostra representativa da colecção real de textos?

- Qual é o tamanho mínimo da colecção para poder ser considerada como válida uma amostra da colecção real que se pretende representar? Como se pode determinar esse tamanho mínimo?

- Os resultados dos eventos de avaliação podem ser extrapolados para a colecção real? Se o sistema *A* se revela superior ao sistema *B* numa dada instância de avaliação, será que o mesmo sucede fora do ambiente de avaliação?

Se fosse possível calcular o erro global inerente ao processo de avaliação, conseguir-se-ia quantificar o ruído das medições dos resultados com significado estatístico, obtendo-se valores de desempenho absolutos dos sistemas. Contudo, é muito difícil quantificar o efeito de todos os erros associados à aproximação que a tarefa faz ao problema.

No entanto, é possível calcular o erro associado a comparações relativas, determinando-se desta forma se as diferenças verificadas entre duas saídas são significativas ou se são fruto de erros de medição, e se o tamanho da colecção usada é suficiente para realizar essa comparação. Assim sendo, é possível extrapolar, com elevado grau de confiança, se as diferenças observadas entre sistemas resultam exclusivamente de terem sido usados dife-

rentes métodos de REM pelos sistemas, e se também se podem verificar fora do ambiente de avaliação.

Como tal, a realização de uma **validação estatística** completa aos resultados obtidos pelos sistemas REM participantes permite calcular o nível de confiança possível nas diferenças de desempenhos observadas nas avaliações conjuntas do HAREM. Adicionalmente, a validação analisa se o tamanho das colecções usadas nas avaliações permite extrair conclusões fundamentadas sobre as estratégias empregues pelos diversos sistemas.

Este capítulo apresenta o trabalho de selecção e de implementação de um teste estatístico adequado para a validação dos resultados. Na secção 5.1 referem-se as validações estatísticas usadas em eventos de avaliação REM passados, e faz-se uma resenha dos testes estatísticos adoptados: o *bootstrap* e o teste de aleatorização parcial. Na secção 5.2 detalha-se o teste de aleatorização parcial e a sua adaptação à metodologia do HAREM. A secção 5.3 descreve uma experiência realizada para analisar a influência do tamanho da colecção nos resultados, e na secção 5.4 apresentam-se os resultados da validação estatística da primeira edição do HAREM.

5.1 Validação estatística para REM

A validação estatística de avaliações conjuntas em PLN deve adoptar os testes estatísticos adequados às especificidades da tarefa. Podemos encontrar exemplos de estudos sobre a aplicabilidade de testes estatísticos a diversas áreas, como é o caso de recuperação de informação (Savoy, 1997; Sakai, 2006) ou da tradução automática (Koehn, 2004; Riezler e Maxwell III, 2005).

Antes de iniciar a validação estatística dos resultados, é preciso seleccionar o teste estatístico mais adequado para a tarefa de REM tal como é apresentada pelo HAREM. No caso de REM, desconhecemos qualquer estudo exaustivo sobre o teste estatístico mais adequado para a comparação de resultados.

O MUC adoptou, para a tarefa de REM, o mesmo teste de aleatorização parcial (*Approximate Randomization*) usado nas restantes tarefas propostas (Chinchor, 1995, 1998a). O objectivo era determinar se as diferenças observadas entre sistemas são realmente significativas, e a validação estatística foi realizada sobre as métricas de precisão, abrangência e medida F. Não há referências sobre se foram considerados e avaliados outros testes estatísticos para a validação.

Nas tarefas partilhadas de REM do CoNLL (Sang, 2002; Sang e Meulder, 2003), foi usado o *bootstrap* para calcular os intervalos de confiança dos resultados da avaliação, somente para a medida F. Também em relação a esta avaliação conjunta, não há informação sobre se foi realizado um estudo sobre o método estatístico mais adequado para validar os resultados da avaliação.

Ambos os métodos – aleatorização parcial e *bootstrap* – são baseados em *testes não-paramétricos*, ou seja, testes que não fazem suposições prévias sobre a distribuição real nem se baseiam nos parâmetros desta, utilizando ao invés os dados disponíveis para gerar uma distribuição empírica, que representa uma aproximação à distribuição real. Para mais informação sobre os os métodos estatísticos referidos neste capítulo, recomenda-se a consulta dos livros (Sheskin, 2000; Good, 2000; Efron, 1981; Moore et al., 2002).

Riezler e Maxwell III (2005) compararam os dois testes estatísticos para algumas métricas usadas na avaliação em PLN, e observaram que a aleatorização parcial apresenta uma margem de erro inferior ao *bootstrap*. Adicionalmente, verifica-se que o *bootstrap* é mais sensível à qualidade do conjunto de observações iniciais, o que pode originar reamostragens enviesadas e levar por vezes à rejeição indevida da hipótese nula (Noreen, 1989).

No caso presente da validação da metodologia do HAREM, questiona-se se a aplicabilidade do método *bootstrap* à tarefa, já que:

- a metodologia de geração de reamostragens do *bootstrap* não tem em consideração as fortes dependências que existem entre EM. Ao invés, o método de aleatorização parcial permite preservar as dependências entre as observações.
- não há garantias de que todas as EM marcadas pelos sistemas sejam usadas nas reamostragens, ao contrário do método de aleatorização parcial. Assim, não há certeza de que as reamostragens geradas sejam representativas da saída do sistema.

Assim sendo, o teste de aleatorização parcial revela-se o teste estatístico mais adequado para a tarefa de validação estatística do HAREM. O teste implementado para a validação estatística foi inspirado pelo trabalho de Chinchor (1992) para o MUC, e descrito em detalhe na secção seguinte.

5.2 Teste de aleatorização parcial

O teste de aleatorização parcial é, na sua essência, um teste de permutações. Estes testes baseiam-se no princípio de que, se a diferença observada entre duas amostras para a métrica M, d, é significativa, então a permuta aleatória de dados entre as amostras irá alterar consideravelmente os valores de d. No caso oposto de a diferença ser ocasional, a permuta de dados não terá um impacto significativo nos valores de d.

O teste de hipóteses pode ser formulado pela seguinte hipótese nula:

> H_0: *A diferença absoluta entre valores da métrica M para as saídas A e B na tarefa de avaliação T, é aproximadamente igual a zero.*

A hipótese nula postula que as duas amostras são semelhantes, afirmando que a diferença d não é significativa. Num cenário com duas amostras semelhantes, é provável que

um certo número n_m de reamostragens apresente valores de d^* iguais ou superiores a d. Por outro lado, se as duas amostras são distintas, isso reflecte-se num valor inicial de d elevado. As n_r reamostragens geradas apresentam uma tendência para obter valores de d menores do que o valor inicial de d, sendo menos frequente observar reamostragens onde $d^* \geq d$.

5.2.1 Metodologia

O teste de aleatorização parcial é levado a cabo através dos seguintes passos:

1. Calcular a diferença absoluta d entre valores da métrica M, para as saídas A e B.

$$d = \mid M_A - M_B \mid \tag{5.1}$$

2. Gerar n_r reamostragens. Para cada reamostragem:

   a) Percorrer o conjunto de todas as observações de A, $O_A = \{O_A^1, O_A^2, ..., O_A^n\}$, e de B, $O_B = \{O_B^1, O_B^2, ..., O_B^n\}$.

   b) Permutar cada par de observações $\{O_A^i, O_B^i\}$, com uma probabilidade θ igual a 0.5.

   c) Calcular a diferença d^* entre os valores da métrica M para as reamostragens A^* e B^*.

$$d^* = \mid M_{A^*} - M_{B^*} \mid \tag{5.2}$$

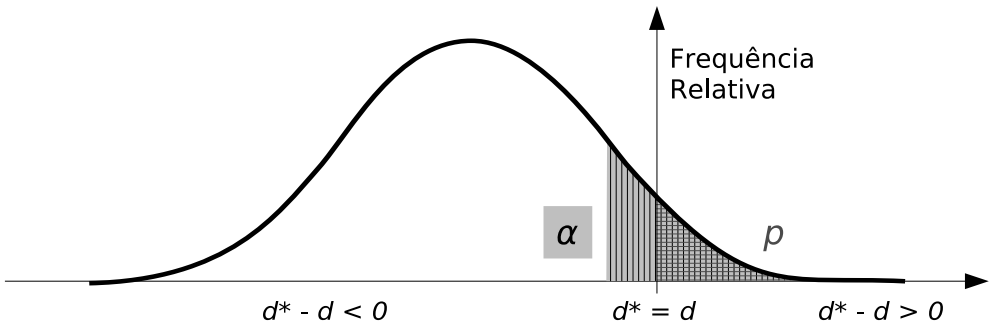
3. Contar o número de vezes (n_m) que o valor de d^* foi igual ou superior a d.

$$n_m = \sum_{i=1}^{n_r} w_i, \quad w_i = \begin{cases} 1 & \text{se } (d^* - d) \geq 0 \\ 0 & \text{se } (d^* - d) < 0 \end{cases} \tag{5.3}$$

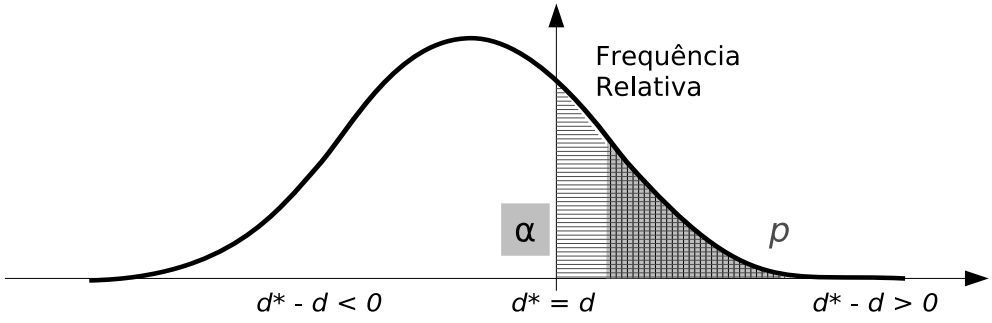
4. Calcular o valor de p:

$$p = \frac{(n_m + 1)}{(n_r + 1)} \tag{5.4}$$

O valor de p é a razão entre n_m, o número de reamostragens onde se observa que $d^* \geq d$, e n_r, o número de reamostragens total. Para valores de p inferiores a um determinado nível de significância α, rejeita-se a hipótese nula, ou seja, a diferença observada entre A e B é significativa. O α representa a probabilidade de se rejeitar a hipótese nula quando esta é verdadeira (e, portanto, não deve ser rejeitada), o denominado *erro de tipo I* (ver Figura 5.1). Por outras palavras, representa a probabilidade de se concluir que as saídas A e B são significativamente diferentes, quando na realidade não o são.



(a) Cenário favorável à rejeição de H_0



(b) Cenário desfavorável à rejeição de H_0

Figura 5.1: Aproximação da distribuição empírica de $d^* - d$ resultante das reamostragens.

Quando n_r cobre o universo de todas as permutações possíveis entre amostras, o teste é denominado aleatorização completa (*Exact Randomization*). No entanto, para amostras com muitas observações, torna-se impraticável gerar todas as permutações possíveis entre amostras, mesmo para a capacidade computacional actual. O teste de aleatorização parcial é uma aproximação ao teste de aleatorização completa, limitado a um determinado número n_r de reamostragens, e a sua distribuição revela-se uma boa aproximação à distribuição real para um número elevado de reamostragens, podendo ser desprezados os erros derivados da aproximação.

5.2.2 Aplicação ao HAREM

A simplicidade e versatilidade do teste de aleatorização parcial permite adaptá-lo facilmente à avaliação de várias tarefas de PLN, como a tradução automática ou a análise

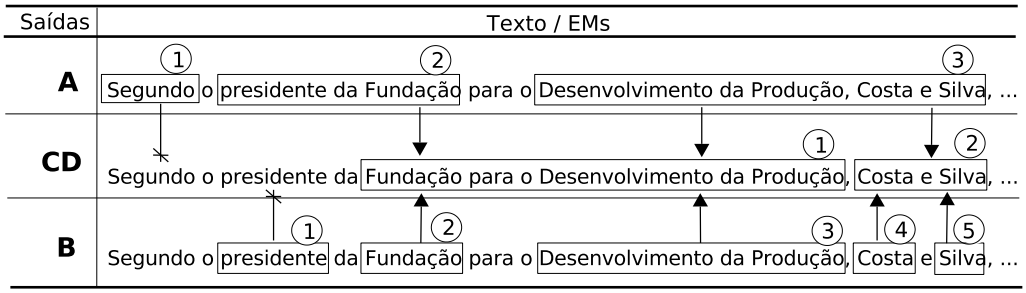


Figura 5.2: Excerto de texto marcado com EM nas saídas *A* e *B*, e respectivos alinhamentos com a CD representados por setas.

morfossintáctica (Morgan, 2006). Um dos pressupostos do teste de aleatorização parcial postula que as observações entre as saídas devem ser permutáveis entre si, o que não é directamente satisfeito pelas saídas dos sistemas de REM participantes no HAREM, uma vez que:

- É frequente encontrar observações espúrias ou em falta na saída *A* que não têm correspondência na saída *B* e vice-versa. Assim, não há um par de observações, mas sim apenas uma observação, para permutar.
- As alternativas das EM vagas na tarefa de identificação podem totalizar números diferentes de observações para as saídas *A* e *B*.
- As observações da saída *A* podem depender de várias observações da saída *B*, e vice-versa. Como tal, em certos casos, o emparelhamento de observações não se pode restringir a pares de EM.

O problema é ilustrado no exemplo da Figura 5.2, onde se pode observar que a CD identifica 2 EM, a saída *A* identifica 3 EM e produz 4 alinhamentos, e a saída *B* identifica 5 EM e produz 5 alinhamentos. A diferença entre o número de alinhamentos para as saídas *A* e *B* viola o pressuposto de permutabilidade dos testes de permutações. Outra situação relevante ilustrada nos alinhamentos respeitantes à EM *presidente da Fundação*, onde se pode verificar que a observação 2 da saída *A* depende das observações 1 e 2 da saída *B*. A permutação destas três observações não pode violar o pressuposto de independência entre observações permutadas.

Apontam-se duas estratégias para resolver os problemas encontrados:

1. Reduzir as observações ao seu elemento mínimo comum, ou seja, permutar os termos do texto.
2. Agrupar as observações ao seu elemento máximo comum, ou seja, permutar blocos de observações do texto.

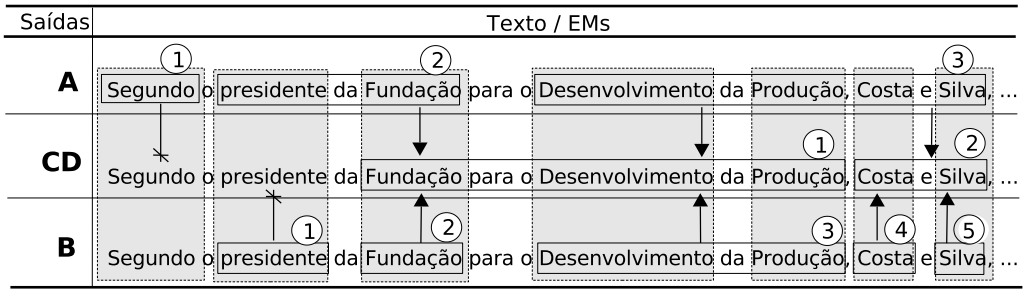


Figura 5.3: Permutações por termos para o exemplo da Figura 5.2.

Permutação por termos e por blocos

A Figura 5.3 ilustra o exemplo da Figura 5.2 com as possíveis permutações segundo a estratégia de permutação por termos. A permutação por termos procura reproduzir a estratégia de REM denominada BIO, no qual o sistema processa sequencialmente os termos do texto (Sang e Meulder, 2003). Segundo esta estratégia, usada nas colecções de texto da tarefa partilhada de REM do CoNLL, os termos são etiquetados com os seguintes marcadores:

B (***B**egin*), se o termo está no início de uma EM.

I (***I**nside*), se o termo pertence a uma EM, mas não a inicia.

O (***O**utside*), se o termo não pertence a nenhuma EM.

Contudo, a permutação por termos possui os seguintes problemas:

- A permutação por termos pode partir as EM em pedaços. Ao partir alinhamentos correctos com uma pontuação de valor igual a 1 em vários pedaços parcialmente correctos, cujo somatório das pontuações possui um valor máximo limitado a 0,5, a pontuação original é alterada. Assim, é muito provável que o valor absoluto da métrica final para as saídas *A* e *B* seja prejudicado pelas quebras de EM, o que pode ter consequências nefastas na decisão de rejeição da hipótese nula.
- Após a quebra das EM e a permuta dos termos, é necessário unir os termos para restaurar as respectivas EM originais. No entanto, no caso da classificação semântica, a reconstrução pode gerar EM com diferentes categorias semânticas (ver Figura 5.4).
- A quebra das EM implica recalcular as pontuações de cada saída. Para tal, é necessário reavaliar as EM em relação à CD para cada reamostragem.

A Figura 5.5 ilustra o exemplo da Figura 5.2 com as possíveis permutações segundo a estratégia de permutação por blocos de EM. A permutação por blocos de EM pode ser

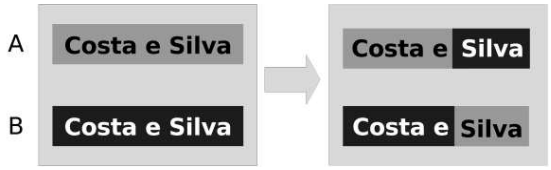


Figura 5.4: Permutações por termos com classificações semânticas diferentes. As saídas *A* e *B* marcaram a EM "Costa e Silva" com categorias diferentes, representadas na figura por tons diferentes.

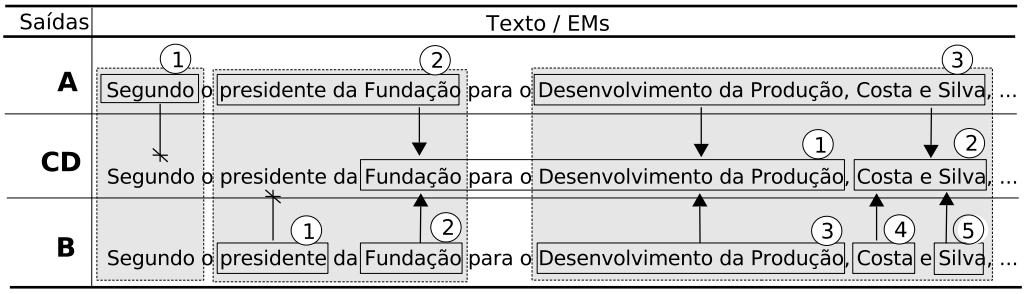


Figura 5.5: Permutações por blocos para o exemplo da Figura 5.2.

interpretada como uma permutação ao nível de determinadas unidades de texto, como unidades lexicais multipalavra, frases ou mesmo parágrafos. A estratégia mantém a independência entre observações, sendo mais adequada aos objectivos apontados para a validação estatística do HAREM.

A permutação por blocos apresenta as seguintes vantagens comparativamente à permutação por termos:

- a permutação por blocos não quebra as EM, evitando os inconvenientes da permutação por termos.
- A pontuação de cada alinhamento não sofre alterações com a permuta, não sendo necessário recalcular as pontuações para cada reamostragem.
- Para alinhamentos sobre EM vagas na sua identificação, a permutação por blocos não é afectada pelo número diferente de alinhamentos que pode existir entre saídas.

Com base nesta análise, adoptou-se a estratégia de permutação por blocos para os testes de aleatorização parcial.

5.3 Experiências com o tamanho da colecção

Como acontece em todos os métodos estatísticos, o número de observações (n_O) tem influência directa na margem de erro do teste. Buckley e Voorhees (2000); Voorhees e Buc-

kley (2002) estudaram a relação que há entre as diferenças observadas entre saídas do TREC (Harman, 1993), o número de observações efectuadas, e o erro associado à conclusão final. Concluíram que existe uma relação e que esta pode ser determinada empiricamente.

Posteriormente, Lin e Hauptmann (2005) conseguiram provar matematicamente o que Voorhees e Buckley tinham concluído empiricamente, mostrando que há uma relação exponencial entre o erro da avaliação e a diferença entre valores de métricas e o número de observações efectuadas. Logo, um aumento do número de observações resulta na diminuição do erro do teste estatístico.

No caso do HAREM, é importante determinar a relação entre o tamanho das colecções douradas usadas nas avaliações e a margem de erro nos resultados obtidos. Para tal, realizou-se uma experiência sobre duas saídas reais do Mini-HAREM. A experiência consistiu em aplicar o teste de permutações a subconjuntos de blocos das saídas A e B cada vez menores, e verificar os valores de p de cada teste.

5.3.1 Selecção dos blocos

Ao restringir o teste estatístico a um subconjunto aleatório de X blocos, está-se a diminuir o tamanho da colecção. Há dois métodos de selecção aleatória de blocos:

1. A selecção realiza-se no início do teste, e as n_r reamostragens são feitas a este subconjunto.

2. A selecção realiza-se antes de cada reamostragem.

Ao implementar o primeiro método de selecção na experiência, observou-se que, para subconjuntos pequenos de blocos, o risco de escolher subconjuntos de blocos pouco representativos da população de blocos aumenta. Assim sendo, os valores do teste estatístico oscilavam consideravelmente consoante o subconjunto de blocos inicial, o que não permitia retirar conclusões.

Consequentemente, optou-se por usar o segundo método de selecção de blocos na experiência aqui descrita. Este método revela-se bem mais robusto quando aplicado em situações em que as amostragens são pouco representativas, obtendo-se resultados mais conclusivos.

5.3.2 Resultados da experiência

As duas saídas usadas na experiênca descritas na Tabela 5.1.

Se se adoptar o critério (subjectivo) de Jones e Bates (1977), que refere que "*differences of 5% are noticeable, and differences of 10% are material*", pode-se estimar *a priori* que a saída *A* é melhor do que a saída *B* com base nos valores das métricas apresentadas na Tabela 5.1. Contudo, esta experiência irá determinar a veracidade desta afirmação com maior certeza.

	Saída A	Saída B	Diferença
Número de EM na saída	4.086	4.191	105
Número de EM na CD	3.663	3.661	2
Número de blocos	4.312	4.312	-
Precisão	79,77%	72,84%	6,93%
Abrangência	87,00%	69,58%	17,42%
Medida F	0,8323	0,7117	0,1206

Tabela 5.1: Resultados da tarefa de identificação para duas saídas do Mini-HAREM.

Observa-se que há uma diferença de 2 EM no número total de EM na CD entre as duas saídas. Esta diferença explica-se pela opção feita por diferentes alternativas em dois casos de EM vagas na sua identificação, por cada saída. O número de blocos (4 312) é aproximadamente 4% maior do que o número de EM marcadas nas saídas, uma discrepância que é causada pelo número de alinhamentos de cada saída com pontuação espúria e em falta que não tem contrapartida na saída oposta, gerando blocos semelhantes ao primeiro bloco do exemplo da Figura 5.5.

A Tabela 5.2 mostra que as médias nas reamostragens das saídas *A* e *B* se mantêm constantes para os subconjuntos de blocos usados. A Tabela 5.3 mostra que o desvio padrão das diferenças entre reamostragens aumenta à medida que o número de blocos diminui, enquanto que a média das diferenças entre reamostragens mantém-se aproximadamente constante.

A precisão é a primeira métrica a registar valores de p acima de α para um nível de confiança de 99% (α = 1%), uma vez que apresenta a diferença inicial mais baixa entre as três métricas. Esta experiência mostra que, quando se diminui o número de blocos no teste de permutações, o desvio padrão da distribuição empírica das métricas aumenta até se atingir um ponto em que o valor de p excede o valor de α (ver Figura 5.10(a)). Como a significância estatística dos resultados depende da métrica usada no teste estatístico e da diferença inicial de valores entre as saídas, não é possível determinar um tamanho mínimo absoluto para a CD.

5.4 Resultados

As Figuras 5.6, 5.7, 5.8 e 5.9 apresentam os resultados das avaliações conjuntas de 2005 e de 2006, para as tarefas de identificação e de classificação semântica (na medida combinada). Nestas figuras estão representados os resultados da validação estatística aos resultados, realizado sobre o conjunto das duas CD, com um nível de confiança de 99% (α = 1%), e com a geração de 9.999 reamostragens para cada teste.

Os resultados da validação estatística estão apresentados sob a forma de caixas cinzentas, que agrupam as saídas onde não é possível concluir que a diferença observada

Reamostragens de A

Nº Blocos	Média			Desvio padrão		
	Precisão	Abrang.	Medida F	Precisão	Abrang.	Medida F
4.312	0,7653	0,7830	0,7741	0,0035	0,0040	0,0032
2.000	0,7653	0,7717	0,7685	0,0080	0,0091	0,0072
1.000	0,7655	0,7650	0,7652	0,0125	0,0145	0,0115
500	0,7654	0,7610	0,7631	0,0187	0,0214	0,0173
250	0,7656	0,7596	0,7623	0,0271	0,0310	0,0252
200	0,7655	0,7593	0,7620	0,0305	0,0348	0,0284
100	0,7657	0,7587	0,7615	0,0437	0,0491	0,0406
75	0,7657	0,7591	0,7614	0,0497	0,0564	0,0464
50	0,7652	0,7579	0,7601	0,0616	0,0685	0,0572
25	0,7665	0,7612	0,7612	0,0860	0,0945	0,0799

Reamostragens de B

Nº Blocos	Média			Desvio padrão		
	Precisão	Abrang.	Medida F	Precisão	Abrang.	Medida F
4.312	0,7654	0,7831	0,7741	0,0035	0,0040	0,0032
2.000	0,7654	0,7719	0,7687	0,0080	0,0091	0,0072
1.000	0,7655	0,7648	0,7650	0,0127	0,0145	0,0116
500	0,7653	0,7609	0,7630	0,0187	0,0216	0,0174
250	0,7655	0,7595	0,7622	0,0272	0,0312	0,0253
200	0,7652	0,7595	0,7620	0,0322	0,0365	0,0295
100	0,7650	0,7582	0,7609	0,0430	0,0494	0,0405
75	0,7655	0,7586	0,7611	0,0506	0,0567	0,0468
50	0,7656	0,7598	0,7613	0,0616	0,0674	0,0566
25	0,7668	0,7618	0,7617	0,0860	0,0951	0,0804

Tabela 5.2: Médias e desvios-padrão para as métricas das saídas A e B, para subconjuntos de blocos de tamanho decrescente, e número de reamostragens n_r igual a 9.999.

Nº Blocos	Valor de p			Média			Desvio padrão		
	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F
4.312	0,0001	0,0001	0,0001	-0,00005	-0,00006	-0,00006	0,0071	0,0081	0,0065
2.000	0,0001	0,0001	0,0001	-0,00013	-0,00021	-0,00017	0,0105	0,0119	0,0095
1.000	0,0001	0,0001	0,0001	-0,00005	-0,00006	-0,00005	0,0147	0,0166	0,0134
500	0,0012	0,0001	0,0001	0,00015	0,00007	0,00011	0,0207	0,0232	0,0188
250	**0,0195**	0,0001	0,0001	0,00009	0,00008	0,00009	0,0293	0,0325	0,0265
200	**0,0320**	0,0001	0,0001	0,00021	-0,00019	0,00001	0,0322	0,0365	0,0295
100	**0,1391**	0,0013	0,0049	0,00070	0,00048	0,00058	0,0461	0,0514	0,0419
75	**0,1925**	0,0029	**0,0121**	0,00016	0,00048	0,00035	0,0532	0,0589	0,0481
50	**0,2909**	**0,0166**	**0,0430**	-0,00042	-0,00193	-0,00120	0,0657	0,0747	0,0608
25	**0,4585**	**0,0946**	**0,1582**	-0,00035	-0,00064	-0,00052	0,0931	0,1047	0,0858

Tabela 5.3: Valores de p, médias e desvios-padrão para as diferenças entre métricas das saídas A e B, para subconjuntos de blocos de tamanho decrescente, e número de reamostragens n_r igual a 9.999.

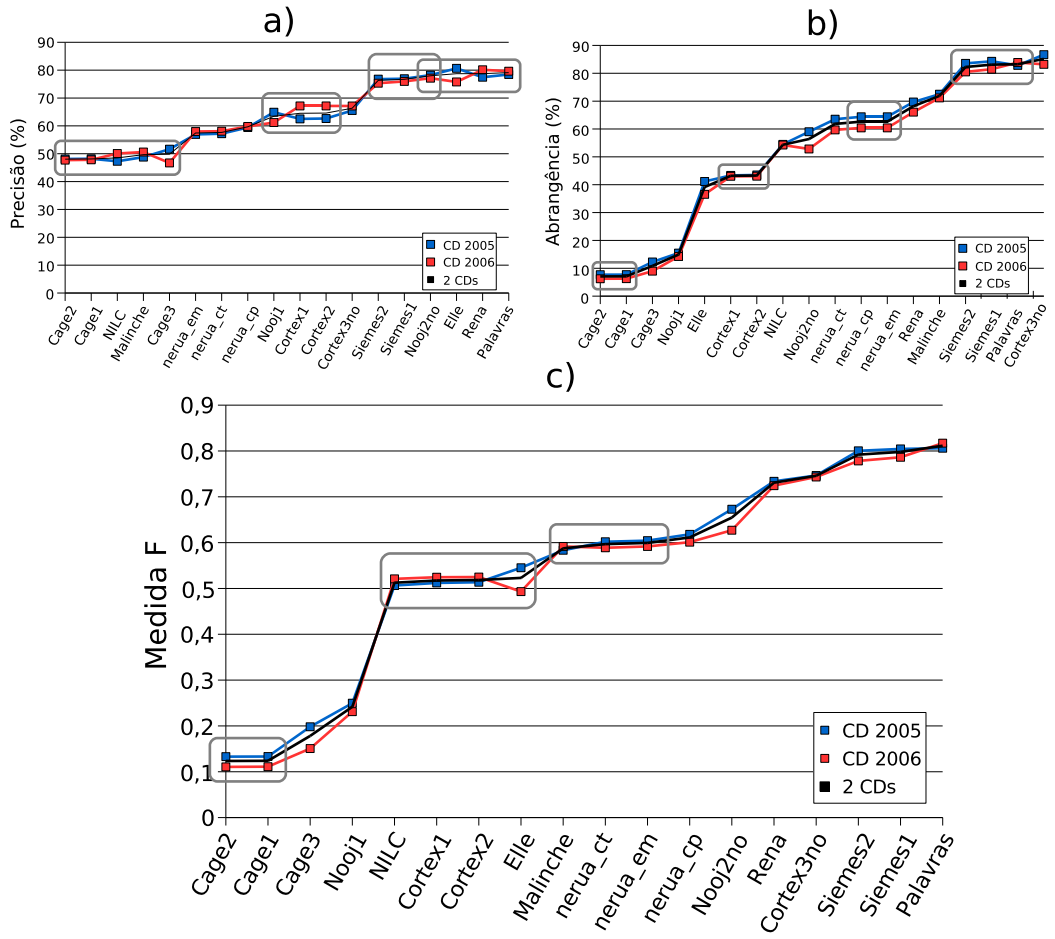


Figura 5.6: Desempenho dos sistemas para a tarefa de identificação no Primeiro HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.

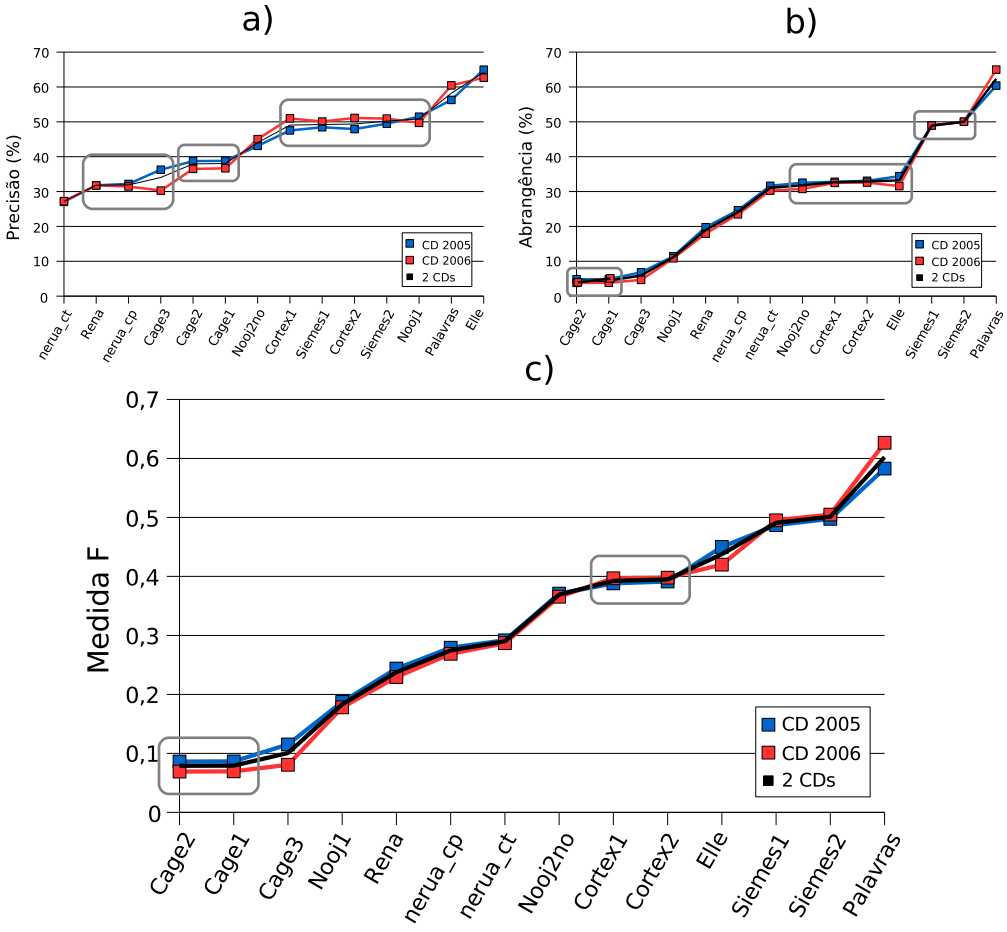


Figura 5.7: Desempenho dos sistemas para a tarefa de classificação semântica (na medida combinada) no Primeiro HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.

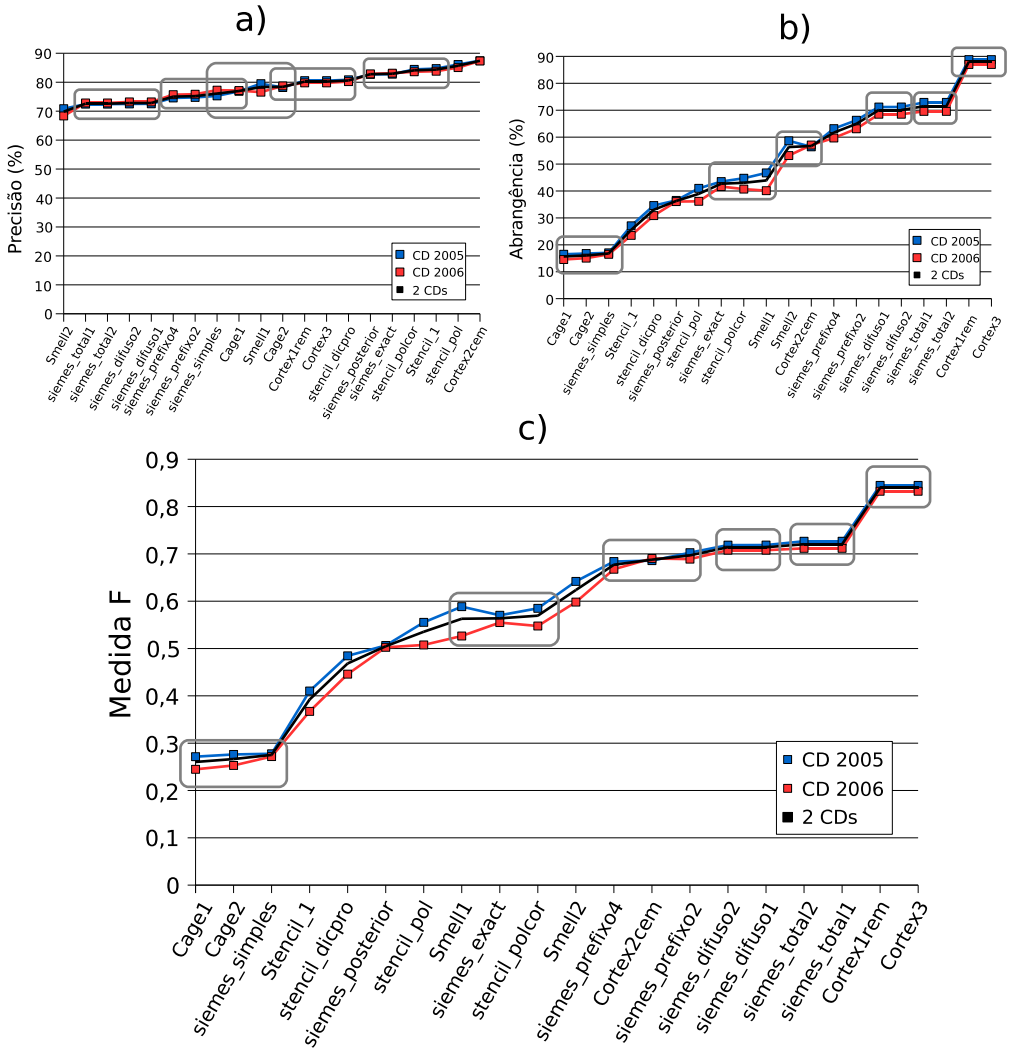


Figura 5.8: Desempenho dos sistemas para a tarefa de identificação no Mini-HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.

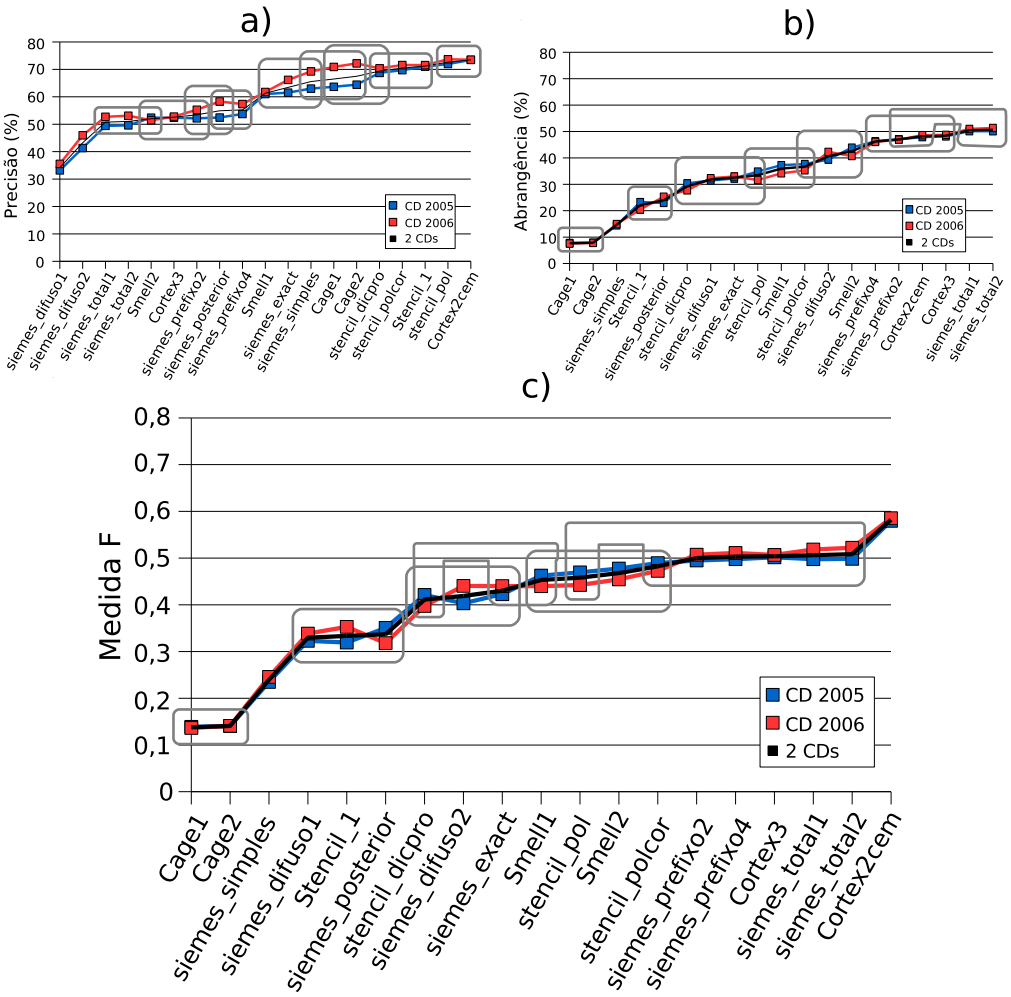
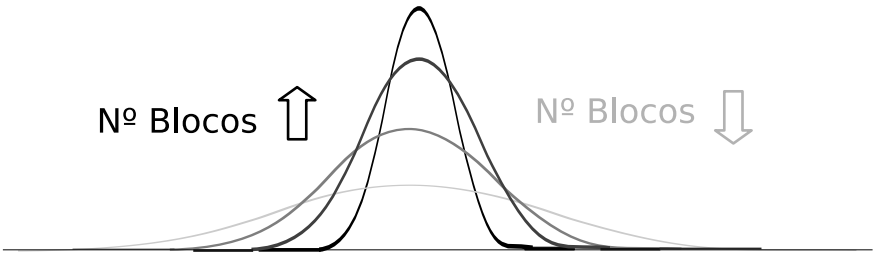
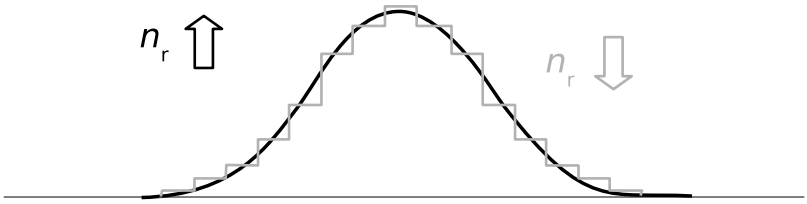


Figura 5.9: Desempenho dos sistemas para a tarefa de classificação semântica (na medida combinada) no Mini-HAREM, para a **a)** precisão, **b)** abrangência e **c)** medida F.



(a) Com a variação do número de blocos para o mesmo número de reamostragens n_r.



(b) Com a variação do número de reamostragens n_r para o mesmo número de blocos.

Figura 5.10: Comportamento da distribuição empírica das métricas.

é significativa (ou seja, o respectivo valor de p é igual ou superior a α. Os valores de p calculados estão apresentados no apêndice C).

O número n_r de reamostragens não afecta o valor de p, mas afecta o seu número de algarismos significativos. Para valores de n_r elevados, a distribuição gerada aproxima-se mais da distribuição real, o que permite ter maior confiança nos valores de p calculados.

Para $n_r = 9.999$, o valor de p é calculado até à décima de milésima (0,0001), o que implica que são precisas 100 ou mais reamostragens que verifiquem a condição $d^* \geq d$ para que $p \geq \alpha$ (para 99% de confiança). No caso de $n_r = 99$, o valor de p é calculado até à centésima (0,01), bastando somente 1 reamostragem que verifique a condição $d^* \geq d$ para que $p \geq \alpha$. Como tal, um número reduzido de reamostragens n_r torna o teste vulnerável à geração de reamostragens excepcionais, e condiciona a confiança que se pode ter no resultado do teste (ver Figura 5.10(b)).

A Tabela 5.4 apresenta os resultados do teste de aleatorização parcial para os subconjuntos de 2.000, 200 e 25 blocos, usando valores de 9.999, 999 e de 99 para o número de reamostragens. Os resultados mostram que o número de reamostragens não tem influência nos valores de média e desvio padrão das métricas.

Nº Blocos	n_r		Valor de p			Média			Desvio padrão		
			Prec.	Abr.	Med.F	Prec.	Abr.	Med.F	Prec.	Abr.	Med.F
2.000	9.999	Saída A				0,7653	0,7717	0,7685	0,0079	0,0092	0,0072
		Saída B	0,0001	0,0001	0,0001	0,7654	0,7719	0,7687	0,0080	0,0091	0,0072
		Diferença				-0,0001	-0,0002	-0,0002	0,0105	0,0119	0,0095
	999	Saída A				0,7654	0,7717	0,7685	0,0081	0,0094	0,0074
		Saída B	0,001	0,001	0,001	0,7655	0,7720	0,7687	0,0079	0,0091	0,0071
		Diferença				-0,0001	-0,0002	-0,0002	0,0107	0,0122	0,0097
	99	Saída A				0,7657	0,7717	0,7687	0,0073	0,0083	0,0066
		Saída B	0,01	0,01	0,01	0,7656	0,7718	0,7686	0,0085	0,0092	0,0075
		Diferença				0,0001	-0,0001	0,0001	0,0096	0,0109	0,0082
200	9.999	Saída A				0,7654	0,7593	0,7620	0,0305	0,0348	0,0284
		Saída B	**0,0320**	0,0001	0,0001	0,7652	0,7595	0,7620	0,0302	0,0348	0,0283
		Diferença				0,0002	-0,0002	<0,00001	0,0322	0,0365	0,0295
	999	Saída A				0,7648	0,7590	0,7616	0,0310	0,0346	0,0285
		Saída B	**0,029**	0,001	0,001	0,7654	0,7598	0,7622	0,0299	0,0348	0,0281
		Diferença				-0,0006	-0,0007	-0,0007	0,0330	0,0355	0,0290
	99	Saída A				0,7623	0,7562	0,7590	0,0310	0,0332	0,0285
		Saída B	**0,04**	0,01	0,01	0,7651	0,7552	0,7598	0,0280	0,0334	0,0261
		Diferença				-0,0028	0,0010	-0,0008	0,0332	0,0405	0,0322
25	9.999	Saída A				0,7665	0,7612	0,7612	0,0860	0,0945	0,0799
		Saída B	**0,4585**	**0,0946**	**0,1582**	0,7668	0,7618	0,7617	0,0860	0,0951	0,0804
		Diferença				-0,0003	-0,0006	-0,0005	0,0931	0,1047	0,0858
	999	Saída A				0,7637	0,7545	0,7563	0,0923	0,0967	0,0843
		Saída B	**0,438**	**0,094**	**0,149**	0,7645	0,7631	0,7609	0,0878	0,0958	0,0809
		Diferença				-0,0007	-0,0086	-0,0046	0,0916	0,1109	0,0877
	99	Saída A				0,7769	0,7645	0,7678	0,0778	0,0922	0,0742
		Saída B	**0,38**	**0,17**	**0,23**	0,7809	0,7636	0,7699	0,0832	0,0954	0,0812
		Diferença				0,0040	0,0010	-0,0021	0,0906	0,1123	0,0920

Tabela 5.4: Valores de p, médias e desvios-padrão das diferenças entre métricas das saídas A e B, para três subconjuntos de blocos e três valores de n_r.

5.4.1 Conclusões

O método de aleatorização parcial foi escolhida para a validação estatística dos resultados do HAREM. A sua adaptação ao HAREM precisou de resolver alguns problemas inerentes à metodologia adoptada por este, como lidar com a vagueza e com alinhamentos parcialmente correctos.

Para verificar se as colecções usadas no HAREM continham um tamanho suficiente para permitir discriminar os sistemas, repetiu-se a validação para ambas as avaliações HAREM, sobre cada colecção dourada e sobre ambas em conjunto. Os resultados finais foram idênticos, o que confirma que as colecções usadas são adequadas para exprimir diferenças com significado entre sistemas de REM.

A análise estatística mostra também que não é possível determinar o tamanho mínimo de tais colecções, pois este parâmetro varia com a diferença inicial observada entre saídas. Contudo, ela permite calcular em todos os casos a margem de erro associada à medição.

Um outro factor que influencia os valores finais da avaliação é a anotação manual das colecções. Como acontece com a maior parte das tarefas desempenhadas por seres humanos, há uma percentagem de EM que suscitam interpretações diferentes no seu reconhecimento por parte de anotadores diferentes. A validação estatística pode ser estendida de maneira a ter em conta a diferença que há na confiança entre as observações, adequando-

-se melhor ao ambiente de avaliação implementado. Um exemplo será usar a informação relativa às EM ambíguas e/ou vagas, atribuindo consequentemente um peso à respectiva observação no teste de aleatorização parcial.

Capítulo 6

O HAREM e a avaliação de sistemas para o reconhecimento de entidades geográficas em textos em língua portuguesa

Bruno Martins e Mário J. Silva

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 6, p. 79–86, 2007.

O HAREM focou uma tarefa genérica de REM em textos na língua portuguesa (Santos et al., 2006), sendo que os tipos de entidades considerados foram mais genéricos do que apenas locais (por exemplo, pessoas, organizações, valores ou abstracções). Tem-se ainda que no caso específico dos locais, não foi feita qualquer atribuição dos mesmos a coordenadas ou a conceitos numa ontologia, e portanto a tarefa de desambiguação não foi considerada. A classificação semântica atribuída às entidades era também bastante genérica (ver capítulo 16), dividindo-se em `CORREIO`, `ADMINISTRATIVO`, `GEOGRAFICO`, `VIRTUAL` e `ALARGADO`. Note-se que muitos destes tipos de "locais" não correspondem a entidades físicas (ou seja, locais com correspondência no mundo real), e portanto um sistema como o CaGE, especialmente desenhado para a tarefa do reconhecimento e desambiguação de referências geográficas em páginas *web* (descrito no capítulo 8), não estaria à partida interessado na extracção destas entidades.

Estas características levam-nos a considerar que a tarefa de avaliação no HAREM, tal como foi definida, não é adequada para a avaliação da totalidade um sistema como o CaGE. Sistemas de extracção de informação focados no problema de extracção de referências geográficas apenas podem fazer uso do HAREM num cenário selectivo bastante restrito, por forma a medir a eficácia no reconhecimento simples e sem classificação geográfica ou desambiguação dos locais reconhecidos. Parece-nos importante que uma futura edição do HAREM considere o caso das referências geográficas de uma forma diferente, através da utilização de anotações na colecção dourada que sejam mais abrangentes e que melhor reflictam a temática geográfica. Nesse sentido, este capítulo apresenta algumas propostas para futuras edições do HAREM, as quais assentam sobretudo em alterações às directivas de anotação (ver capítulo 16).

6.1 Conceitos e trabalhos relacionados

A extracção de diferentes tipos de entidades mencionadas em texto é uma tarefa básica em processamento da linguagem natural, e um dos componentes chave da MUC (Chinchor, 1998b). O problema foi automatizado com sucesso, sendo frequente obter-se um desempenho semelhante ao de um ser humano. No entanto, o caso específico das referências geográficas levanta algumas considerações adicionais:

- As nossas entidades (referências geográficas e a sua classificação em tipos tais como ruas, cidades ou países) são mais específicas do que os tipos básicos considerados no MUC (pessoas, organizações, locais).
- A especificação completa de uma localização geográfica pode necessitar de relações espaciais (por exemplo, distância, direcção, ou topologia). Expressões contendo este tipo de relações devem ser consideradas como referências geográficas.

- Mais que reconhecer referências geográficas, é necessário fazer também a correspondência com os conceitos numa ontologia, uma vez que o reconhecimento só por si não atribui um sentido às referências reconhecidas. O REM é estendido com classificação semântica por tipo geográfico e com a associação a conceitos numa ontologia, ambos problemas mais complexos do que o simples reconhecimento (Kornai e Sundheim, 2003).

- Por forma a processar grandes quantidades de texto em tempo útil, os documentos individuais devem ser processados num tempo razoável. Esta restrição afecta seriamente a escolha de heurísticas a considerar pelo sistema. Infelizmente, tem-se que as questões de desempenho tendem a ser ignoradas em estudos de avaliação de REM, e o evento HAREM não foi uma excepção.

A investigação na especialização geográfica da tarefa genérica do REM está agora apenas a começar, mas existem já resultados publicados sobre este problema em concreto (Li et al., 2002; Olligschlaeger e Hauptmann, 1999; Smith e Mann, 2003; Smith e Crane, 2001; Schilder et al., 2004; Manov et al., 2003; Nissim et al., 2004; Leidner et al., 2003; Rauch et al., 2003). Por exemplo, a *Workshop on the Analysis of Geographical References* focou tarefas mais complexas que o simples reconhecimento de entidades geográficas em texto (Kornai e Sundheim, 2003). Alguns dos sistemas apresentados lidavam com a classificação e o mapeamento das referências geográficas nas coordenadas geodésicas correspondentes, embora apenas tenham sido reportadas experiências iniciais. Várias heurísticas foram já testadas, mas os sistemas variam muito nos tipos de classificação e desambiguação que efectuam, sendo que os recursos usados para avaliação também não se encontram normalizados. Não existe até hoje uma solução geral para o problema, e não existe ainda nenhum recurso de avaliação do tipo "colecção dourada" para a avaliação de sistemas de REM focados em referências geográficas.

Pensamos que o HAREM pode ter um papel importante no desenvolvimento desta área, possibilitando a avaliação de sistemas de extracção de informação que tratem o problema das referências geográficas em texto de uma forma mais abrangente do que apenas limitando-os a uma tarefa de reconhecimento simples.

6.2 Proposta para futuras edições do HAREM

Tal como exposto atrás, a colecção dourada e as directivas de anotação utilizadas pelo HAREM não se adequam à avaliação de sistemas que lidem explicitamente com o problema das referências geográficas. No entanto, pensamos ser possível fazer uma re-anotação da colecção dourada por forma a torná-la mais útil a este problema, não sendo para isso necessário um grande dispêndio de esforço. A nossa proposta para futuras edições do

HAREM vai essencialmente no sentido de considerar a sub-tarefa do reconhecimento das referências geográficas a um maior nível de detalhe.

No que resta desta secção abordamos três aspectos que nos parecem de especial importância, nomeadamente a existência de uma classificação semântica refinada para as entidades de categoria LOCAL, a existência de anotações para ontologias geográficas padrão, e a possibilidade dos sistemas considerarem sub-anotações e anotações alternativas para uma entidade. É ainda descrito outro aspecto que, embora de menor importância, deveriam ser também levado em conta numa futura edição do HAREM, nomeadamente a consideração do desempenho computacional como uma métrica de avaliação.

6.2.1 Classificação semântica refinada para as EM de categoria LOCAL

Em primeiro lugar, achamos que os tipos considerados para a classificação semântica das EM de categoria LOCAL deveriam ser estendidos por forma a melhor reflectir a temática geográfica. As etiquetas propostas no HAREM tiveram por base necessidades genéricas em processamento de linguagem natural. Como tal, pensamos que as etiquetas recomendadas para anotação da referências geográficas estão distantes das necessidades deste domínio específico, e carecem de uma revisão para futuras edições. Os tipos GEOGRAFICO e ADMINISTRATIVO, tal como se encontram definidos nas directivas de anotação, poderiam ser estendidos com sub-tipos mais específicos, tais como *rio*, *montanha* no primeiro caso, e *país*, *cidade*, *município* ou *freguesia* no segundo.

A hierarquia de tipos a considerar poderia, por exemplo, ser baseada num almanaque ou ontologia geográfica já existente (vários encontram-se amplamente divulgados, tais como o GeoNET (Chaves et al., 2005), o Getty TGN (Harpring, 1997), a *geonames ontology* (Vatant, 2006) ou o almanaque do projecto Alexandria Digital Library (Hill et al., 1999; Hill, 2000). Desta forma, teríamos uma classificação semântica para as EM de categoria LOCAL inspirada em trabalhos conhecidos na área do processamento de informação geográfica. Sistemas de anotação que, no seu funcionamento interno, utilizem uma hierarquia de tipos geográficos diferente, devem à partida conseguir traduzir os tipos geográficos por eles considerados para os tipos definidos nestes recursos. Estas próprias ontologias e almanaques incluem uma definição precisa de quais os tipos geográficos que consideram (Hill, 2000).

6.2.2 Geração de anotações para ontologias geográficas padrão

Além de uma classificação semântica mais refinada para as EM de categoria LOCAL, pensamos que a colecção dourada deveria conter as referências geográficas associadas a alguma forma de identificação única, por forma a se poder também testar uma tarefa de desambiguação completa. Poder-se-ia, mais uma vez, recorrer a almanaques ou ontologias geográficas padrão listados anteriormente. Exceptuando a GeoNetPT, todos os restantes recursos

são de âmbito global, contendo na sua maioria nomes geográficos em inglês. Contudo, a associação de uma referência geográfica em texto com o conceito correspondente na ontologia não depende obrigatoriamente do nome, mas sim do conceito que se encontra referenciado. Todos os recursos anteriormente listados descrevem conceitos geográficos relativos a Portugal, apresentando ainda alguns nomes em português (por exemplo, nomes alternativos para regiões geográficas importantes).

A anotação de cada local na colecção dourada seria estendida por forma a incluir uma referência para os identificadores correspondentes a esse conceito geográfico numa das ontologias. Este campo poderia incluir vários identificadores, no caso do local subjacente se encontrar definido por vários conceitos na ontologia, ou mesmo ser deixado em branco caso o local não se encontre definido.

Embora a anotação da colecção dourada com identificadores numa qualquer ontologia levasse à necessidade de que todos os sistemas que desejem fazer anotações desta forma partilhem esse mesmo recurso de informação externo, poder-se-ia considerar um cenário em que as referências geográficas fossem anotadas com as coordenadas geodésicas correspondentes, em lugar de se fazer as anotações com os conceitos na ontologia. Desta forma, a avaliação da tarefa de desambiguação podia ser feita com base nas coordenadas físicas reais associadas ao local, em lugar de depender de informação externa, sendo que cada sistema ficava livre de usar diferentes recursos para fazer a anotação. Ontologias padrão como as mencionadas anteriormente contêm coordenadas geodésicas, ou mesmo informação poligonal, para a maioria dos conceitos que definem, sendo que fazer a anotação da colecção dourada desta forma não nos parece problemático. Note-se no entanto que caso se usem coordenadas, a tarefa de avaliação necessita de contabilizar questões de imprecisão nas coordenadas (por exemplo, definindo uma distância mínima), visto que diferentes sistemas podem associar coordenadas diferentes ao mesmo conceito (devido, por exemplo, a factores de precisão numérica).

6.2.3 Possibilidade de considerar sub-anotações e anotações alternativas

As directivas de anotação do HAREM, tal como se encontram definidas, consideram que os nomes de locais que são dados como parte do nome de uma entidade de outro tipo (por exemplo, uma organização) não devem ser reconhecidos como tal. Por exemplo em *Câmara Municipal de Braga*, a totalidade da expressão deveria ser anotada como uma organização, sem que *Braga* fosse anotado como um local (ver secção 16.7.2). Para mais, o HAREM considerou o facto de os nomes dos locais muitas vezes assumirem um papel semântico diferente, não devendo nestes casos ser anotados como locais. Por exemplo, na frase *Portugal apoia envio de missão da ONU*, o nome *Portugal* deverá ser anotado como uma organização. Ainda que o papel semântico das entidades seja nestes casos claramente diferente do de uma referência explícita a uma localização, é também claro que estas entidades

continuam a ter uma forte conotação geográfica.

No sentido de resolver as questões colocadas acima, pensamos que as regras de anotação deveriam ser estendidas de forma a considerar sub-anotações e anotações alternativas. Nos casos como *Panificadora de Lisboa*, a expressão completa poderia ser anotada como uma organização e a palavra *Lisboa* nela contida poderia ser anotada como um local. Em casos como o da frase *Portugal apoia envio de missão da ONU*, deveria ser possível anotar *Portugal* de acordo com o seu papel semântico de local e o seu papel semântico de organização, mantendo-se desta forma os vários papéis semânticos possíveis para a palavra. Pretendemos assim que o HAREM continue a potenciar o desenvolvimento de sistemas que lidem com tarefas de desambiguação semântica das entidades, sem no entanto penalizar os sistemas que se focam numa tarefa de reconhecimento mais simples à semelhante do MUC, ou mais especializada num determinado tipo de entidades.

O HAREM poderia, por exemplo, considerar um formato de anotação que permitisse associar várias propriedades (possivelmente até de ontologias ou hierarquias de classificação diferentes) ao mesmo conteúdo textual. Em lugar de se providenciarem as anotações juntamente com o texto, poderíamos ter um esquema semelhante ao que se apresenta de seguida, no qual as anotações são feitas independentemente do texto, desta forma possibilitando que várias anotações possam facilmente ser feitas ao mesmo bloco do texto, ou até mesmo que as anotações sejam estendidas ao longo do tempo com novas classes de informação.

```
<DOCUMENTO>
<TEXTO>
Portugal envia missão de apoio.
</TEXTO>
<ANOTACOES>
<EM morf="m,s" palavra_inicio="1" palavra_fim="1" />
<EM classe="local" tipo="administrativo" subtipo="país"
geoid="GEO_1" palavra_inicio="1" palavra_fim="1" />
<EM classe="organização" tipo="administração"
palavra_inicio="1" palavra_fim="1" />
</ANOTACOES>
</DOCUMENTO>
```

Este esquema é bastante semelhante ao usado na proposta inicial do Open Geospatial Consortium para um serviço de anotação de referências geográficas em textos (Lansing, 2001). No entanto, um esquema desta natureza pressupõe a existência de uma atomização comum (isto é, partilhada por todos os sistemas participantes), visto que cada anotação é feita com base num átomo de início e fim para a mesma. Anteriores eventos de avaliação conjunta, focados no problema do REM, foram já baseados em colecções douradas

previamente atomizadas (Sang e Meulder, 2003). Contudo, uma conclusão importante do HAREM foi que a tarefa da atomização de textos em português é relativamente complexa, sendo que diferentes sistemas podem optar por fazer a atomização de diferentes formas (ver capítulos 18 e 19). Idealmente, a tarefa de avaliação deverá ser tanto quanto possível independente da atomização usada pelos sistemas, pelo que o esquema de anotação anterior poderá não ser o mais indicado.

Note-se ainda que o esquema de anotações alternativas em que cada entidade pode ter mais do que um tipo semântico associado deverá ser diferente do considerado nas directivas do HAREM para o caso da vagueza na classificação semântica. Em vez da anotação típica do HAREM, a qual não obedece aos requisitos da linguagem XML, e que se encontra exemplificada em baixo:

```
<LOCAL|ORGANIZACAO tipo="ADMINISTRATIVO|ADMINISTRACAO"
MORF="M,S">Portugal</LOCAL|ORGANIZACAO> envia missão de apoio.
```

Fazemos duas propostas de melhoria da representação de anotações das entidades mencionadas. A primeira seria de uma forma semelhante ao seguinte exemplo:

```
<EM classe="local|organizacao" tipo="local:administrativo"
subtipo="local:administrativo:pais" tipo="organização:administração"
morf="m,s" geoid="GEO_1"> Portugal </EM> envia missão de apoio.
```

Embora o exemplo anterior já obedeça aos requisitos da linguagem XML, a interpretação dos valores associados aos atributos das anotações `<EM>` pode ainda obrigar à criação de código adicional para processamento dos valores dos atributos. A segunda proposta teria um formato de anotação que define diferentes atributos XML para cada um dos tipos de entidades e classificações possíveis:

```
<EM local organizacao masculino singular tipo-local="administrativo"
subtipo-local="pais" tipo-organizacao="administracao" geoid="GEO_1">
Portugal</EM> envia missão de apoio.
```

6.2.4 Desempenho computacional

Além dos pontos referidos atrás, que essencialmente se relacionam com a anotação da colecção dourada de uma forma mais abrangente, há dois outros pontos que achamos importante rever, nomeadamente a consideração do desempenho computacional como uma métrica de avaliação. Esta é, quanto a nós, uma variável importante que afecta o desenvolvimento de qualquer sistema de REM, sendo que muitas vezes os sistemas optam por usar heurísticas mais simples em troca de ganhos significativos em desempenho. Juntamente com o envio das saídas dos sistemas, os participantes deveriam ser encorajados a partilhar

com os restantes o tempo que os seus sistemas demoraram a proceder à anotação dos textos, assim como a plataforma de *hardware* onde a anotação foi executada. Embora de uma forma algo informal, estes dados já permitiriam efectuar uma comparação dos diferentes sistemas participantes no que diz respeito à variável desempenho.

6.3 Conclusões

Neste capítulo discutimos as limitações do HAREM no que diz respeito aos sistemas focados no tratamento de referências geográficas. Em futuras edições, gostaríamos de ver o cenário das referências geográficas tratado em maior detalhe, nomeadamente através da anotação da colecção dourada de uma forma mais abrangente.

Capítulo 7

Balanço do Primeiro HAREM e perspectivas de trabalho futuro

Diana Santos e Nuno Cardoso

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 7, p. 87–94, 2007.

Neste capítulo iremos analisar em pormenor algumas opções tomadas no início do HAREM e que, vendo agora em retrospectiva, constatamos que errámos ou que não escolhemos, pelo menos, a alternativa mais apropriada.

Globalmente, fazemos um balanço francamente positivo do HAREM, não só pela participação e entusiasmo da comunidade em relação à iniciativa, mas também por ter levado a bom porto a avaliação em REM idealizada pelo estudo preliminar descrito no capítulo 2. Questões como a vagueza, a anotação em contexto, a adopção de uma categorização semântica consensual, ou a utilização de textos de diferentes proveniências e variantes, foram pela primeira vez introduzidas em avaliações conjuntas de REM. Adicionalmente, fomentámos a discussão no seio da comunidade, em torno da melhor metodologia de avaliação dos seus sistemas, o que resultou em contribuições importantes e que, acreditamos, fará do HAREM uma referência importante para avaliações conjuntas futuras na área.

O capítulo começa com uma autocrítica ao HAREM, referindo alguns tópicos sobre os quais temos actualmente uma opinião diferente em relação ao que foi feito. Esta análise pretende garantir que essas opções sejam documentadas e corrigidas em próximas avaliações conjuntas no âmbito do HAREM, fomentando uma reflexão da comunidade em seu redor. De seguida, apresentamos algum trabalho que achamos que será da maior utilidade efectuar, com base naquilo que foi feito no HAREM e no Mini-HAREM, antes de começar a organizar novas rondas de avaliação conjunta, mesmo que tal implique algum atraso na organização da segunda edição. Só nessa altura fará sentido, na nossa opinião, escolher o caminho futuro a seguir como comunidade, sobre o qual fazemos alguns comentários na terceira parte.

7.1 Uma retrospectiva das opções tomadas

7.1.1 Uma dependência infeliz entre a classificação e a identificação

Uma das opções que hoje admitimos não ter sido feliz diz respeito à separação da tarefa de REM em dois passos: identificação e classificação. Esta modularidade, apesar de ser interessante e até ter permitido que outros participantes pudessem também aproveitar o HAREM para tarefas relacionadas, como foi o caso nas Morfolimpíadas (Santos et al., 2003; Costa et al., 2007), em que além de analisadores morfológicos também participaram radicalizadores e verificadores ortográficos, transmitiu infelizmente uma aparência de independência entre os passos que na realidade não existiu, se tomarmos em conta a forma como as categorias do HAREM foram concebidas.

Ou seja, ao termos considerado que a delimitação correcta de certo tipo (semântico) era `COISA` ou `VALOR`, estamos já, ao nível da tarefa da identificação, a pressupor (nesses casos) uma classificação implícita correcta para podermos atribuir uma identificação correcta, dado que definimos directivas de identificação separadas para essa categoria (ver

capítulo 18):

```
Comprei uma flauta <COISA TIPO="CLASSE">de Bisel</COISA>
Tem um comprimento de <VALOR TIPO="QUANTIDADE">60 metros</VALOR>.
```

Embora isto não aconteça na maioria dos casos[1], ou seja, para as outras categorias a independência é real, essa dependência invalida conceptualmente a separação.

7.1.2 Avaliação da identificação baseada em categorias de classificação

Um outro ponto em que foi nebulosa a contribuição do HAREM foi a nossa escolha de apresentar relatórios de desempenho cujas medidas de identificação se encontravam discriminadas por categoria (semântica), o que produziu uma confusão generalizada entre os participantes. A necessidade ou mesmo o interesse de efectuar e apresentar esse tipo de relatórios precisa assim de ser repensada.

A ideia subjacente à geração desses relatórios era a seguinte: em paralelo com a apresentação de resultados considerando, por exemplo, apenas texto literário, ou apenas da Web, ou apenas da variante brasileira, era possível também mostrar os resultados segundo os vários conjuntos de categorias: só pessoas, só obras, só coisas, etc, e fazer as mesmas medições. Isto é equivalente a um participante apenas escolher uma categoria para concorrer, aplicando-se um véu que retirava todas as outras categorias (ver secção 19.2.5).

O que não foi compreendido pela maioria dos participantes foi que isso não significava filtrar apenas os casos em que a CD continha EM classificadas como PESSOA, mas sim entrar em conta, também, com todos os casos erradamente marcados como PESSOA pelos sistemas (ou seja, EM espúrias), o que significa que, ao contrário dos casos da variante ou do género textual, usados por todos os sistemas (e discriminados depois nos relatórios de desempenho), as medições por categoria dependem da saída de cada sistema e podem portanto não ser uma forma fácil de comparar os sistemas entre si.

A Tabela 7.1 exemplifica, usando a categoria PESSOA, todos os casos que são levados em conta para as várias pontuações por categoria. Para a tarefa de **identificação** em relação à categoria PESSOA, os casos 1, 2 e 3 são considerados correctos, enquanto que os casos 6, 7 e 8 são considerados parcialmente correctos. Para a tarefa de **classificação** da categoria PESSOA, já apenas o caso 1 e (parcialmente) o caso 6 são correctos. Além disso, a diferença entre os cenários relativo e absoluto (como sempre), é que o primeiro não considera no denominador casos espúrios e em falta, como por exemplo os casos 4 e 5 (veja-se a explicação detalhada dos diferentes valores destas medidas no capítulo 18).

[1] Por exemplo, na primeira CD, num universo de 5086 EM, há 7 casos de COISA com o padrão acima, 4 distintas. Para VALOR, há 132 ocorrências tal como o padrão de cima para unidades tais como metros, kg, escudos ou bits, sendo 106 dessas ocorrências unidades temporais (anos, meses, dias ou minutos).

Caso	CD	Sistema	Comentário
1	PESSOA	PESSOA	identificação e classificação correctas
2	X	PESSOA	o sistema identifica uma EM como PESSOA que na CD é diferente
3	PESSOA	X	o sistema identifica uma EM PESSOA como outro tipo de EM
4		PESSOA	o sistema identifica uma EM espúria como PESSOA
5	PESSOA		o sistema não identifica como EM uma PESSOA na CD
6	PESSOA	PESSOA	apenas parcialmente identificada, e class. semântica correcta
7	X	PESSOA	apenas parcialmente identificada, e class. semântica espúria
8	PESSOA	X	apenas parcialmente identificada, e class. semântica em falta

Tabela 7.1: Todos os casos relacionados com a avaliação da identificação por categoria PESSOA. X significa o nome de uma categoria diferente de PESSOA.

7.1.3 Cenários relativos vistos por outra perspectiva

Outra questão pode ser levantada em geral em relação à pertinência de definir um cenário relativo: se, de facto, como constatámos acima, em alguns casos as duas tarefas (identificação e classificação) não são independentes, isso retira (pelo menos nesses casos) o sentido a tal cenário. Parece ser portanto mais correcto usar apenas o cenário absoluto para avaliar os sistemas, dado que as medidas relativas são de certa forma virtuais, e os sistemas na prática têm de efectuar ambas as decisões até à marcação final da EM (ou melhor, as decisões não são independentes).

Note-se, aliás, que se tornam aparentes mais duas desvantagens do cenário relativo: uma, foi talvez ter induzido os sistemas em erro devido à aparente independência conceptual entre as duas tarefas. Outra, foi a possibilidade de introduzir um elemento de "adaptação ao HAREM": um sistema com dúvida numa dada categoria teria melhores resultados no HAREM (cenário relativo) não a reconhecendo do que tentando classificá-la. Pensamos que ninguém se aproveitou desta característica, mas é uma indicação de que não há vantagem em definir artificialmente um cenário que não representa (e consequentemente mede) uma tarefa independente.

7.1.4 Inconsistência nas medidas usadas

Outra questão refere-se às medidas: Embora nos tenhamos concentrado na capacidade de discriminação dentro de cada categoria, entrando em conta com a quantidade de informação que cada **tipo** (ou conjunto de tipos) implicava, ficou por fazer uma medida que entrasse em conta com a capacidade de discriminação entre **categorias**, e que é claramente mais interessante do ponto de vista de medir a dificuldade da tarefa de REM em português.

Uma outra área com clara potencialidade de melhoria refere-se à classificação de EM com alternativas de delimitação e/ou de encaixe, com a respectiva classificação de parcialmente correcto. Embora tenhamos argumentado em Santos et al. (2006) a favor da existên-

cia da classificação parcialmente correcta em vez de um "tudo ou nada" como preconizado pelo MUC, é claro que há casos em que tal faz mais sentido do que outros. Ou seja, pode haver EM disparatadas que recebem no HAREM uma gratificação que não merecem, enquanto que outras são desvalorizadas (pelo tamanho) embora com muito mais significado intrínseco. Apresentamos um exemplo hipotético apenas para ilustrar esta questão:

> As Actas do ETNR do Departamento de Informática do Rio Azul/Brasil e as do PROPOR foram publicadas pela Springer.

Segundo as directivas do HAREM, o exemplo seria anotado da seguinte forma:

> As **<OBRA TIPO="REPRODUZIDA"> Actas do ETNR do Departamento de Informática do Rio Azul/Brasil </OBRA>** e as do **<ACONTECIMENTO TIPO="ORGANIZADO"> PROPOR </ACONTECIMENTO>** foram publicadas pela **<ORGANIZACAO TIPO="EMPRESA"> Springer </ORGANIZACAO>**.

Neste caso, os sistemas que produzissem EM como <EM> Azul/Brasil </EM>, <EM> Informática do Rio </EM> ou <EM> As Actas </EM> não deveriam receber qualquer pontuação, enquanto que aqueles que marcassem <EM> Actas do ETNR </EM> ou <EM> Departamento de Informática do Rio Azul/Brasil </EM> já nos parecem merecer uma pontuação parcial.

7.1.5 Tratamento dos problemas incluídos em texto real

Finalmente, uma questão muitas vezes referida mas que não foi tratada convenientemente refere-se à inclusão de texto real (por exemplo, com erros ortográficos ou com uso indevido de maiúsculas) na Colecção HAREM e na CD. Esses casos deveriam estar marcados, de forma a poderem ser automaticamente ignorados pelos módulos da avaliação. É muito importante sublinhar que consideramos que os sistemas devem ser alimentados com texto real; contudo, nos casos em que não é possível obter um consenso, não se deve favorecer ou prejudicar os sistemas através de uma decisão arbitrária, e por isso a avaliação destes não deve incluir erros ou problemas não resolvidos. Embora tal já tenha sido parcialmente feito através da etiqueta <OMITIDO> na CD (ver capítulo 19), ainda muitos casos ficaram por tratar.

7.2 Receitas para uma nova avaliação conjunta fundamentada

Antes de nos abalançarmos a organizar um novo HAREM, há vários estudos que precisam de ser realizados, de forma a que todo o processo possa ser melhor avaliado, e sabermos que escolhas vale a pena manter e quais as que podemos abandonar ou mudar.

No que se refere à validação estatística do método, já foi feito um trabalho importante (veja-se o capítulo 5 e Cardoso (2006a)); contudo, é ainda preciso esclarecer algumas outras questões conceptuais.

Em alguns casos, isto requer o enriquecimento ou verificação adicional da CD, por isso principiamos por listar o que pretendemos fazer como uma continuação lógica do trabalho de investigação sobre o REM em português:

- marcação da CD por mais investigadores independentes, de forma a medir a concordância inter-anotadores e refinar também a compreensão (e documentação) das directivas. A determinação da concordância inter-anotador permitirá calcular o erro da medição inerente ao erro humano (Will, 1993), e determinar com maior rigor o nível de confiança nos resultados das avaliações (Maynard et al. (2003a) comparam o MUC e o ACE a esse respeito).

- marcação sistemática dos casos problemáticos e com erros, de forma a não serem contados pela arquitectura de avaliação;

- marcação de todas as EM encaixadas;

- marcação com o tipo semântico pormenorizado (país, cidade, jornal, etc) e eventualmente traduzi-lo para um esquema MUC, em que, por exemplo, país e cidade são `LOCAL` independentemente do seu contexto, ou menções a jornais classificadas como `ORGANIZACAO` (ver os capítulos 4 e 3 para explicação detalhada das diferenças entre os tipos semânticos empregues);

- marcação da CD segundo as directivas do ACE;

- marcação de dependências anafóricas.

Talvez a tarefa mais importante que se nos depara é a medição da dificuldade das tarefas, quer através do recurso a um almanaque "ideal", quer através da simplicidade da atribuição de uma dada classificação – e para isto teremos não só que classificar os contextos sintácticos como a possibilidade de encaixe e/ou de ambiguidade das várias EM.

Parece-nos pois interessante estudar meios de realizar uma selecção automática das EM mais difíceis de reconhecer e/ou classificar, e realizar uma nova avaliação (usando os resultados já existentes dos sistemas) segundo este cenário de "elite". A principal intuição subjacente a esta proposta é a de que há tipos de EM (por exemplo, as expressões numéricas) que pouco contribuem para distinguir os sistemas, e que "diluem" os valores dos resultados finais. Ao usar um leque de EM difíceis como um novo Véu (ver secção 19.2.5),será mais fácil distinguir os melhores sistemas, eventualmente para tarefas diferentes.

Outra questão de interesse óbvio é investigar a relação entre a dificuldade de anotação para um sistema automático e para a anotação intelectual. Na pista dessa, e após reanotação da CD, será também preciso comparar, como sugerido no capítulo 4, a dificuldade do esquema MUC com a do esquema HAREM e quantificar, ao mesmo tempo, em quantos casos é que há sobreposição, ou seja, em que a diferença é apenas teórica.

Finalmente, esperamos que a disponibilização pública, quer das CD quer dos resultados dos sistemas, permita estudar métodos de análise sintáctico-semântica que indiquem o tipo ou categoria de forma a podermos compilar semi-automaticamente mais texto, usando por exemplo a Floresta Sintá(c)tica (Afonso et al., 2002; Bick et al., 2007) para texto jornalístico, o COMPARA (Frankenberg-Garcia e Santos, 2002) para texto literário e o BACO (Sarmento, 2006a) (marcado automaticamente com o SIEMÊS (Sarmento, 2006b)) para texto da Web. Estes métodos permitirão não só criar maiores colecções de texto, mais variadas, como também alcançar (se tal for considerado desejável) um determinado balanço entre os vários casos difíceis, em vez de prosseguir uma abordagem cega de apenas mais quantidade de material.

7.3 Alguns futuros possíveis

Esta secção descreve algumas propostas feitas no Encontro do HAREM, dando evidentemente crédito aos seus autores, mas tentando sobretudo fazer um ponto da situação sobre os vários futuros que a comunidade tem à sua frente, convencidos de que o futuro dependerá tanto de nós, organizadores, como da comunidade.

Martins et al. (2006) sugeriram que o significado (ou seja, o resultado da análise semântica), pelo menos das EM geográficas, fosse dado com mais detalhe, ou seja, que além de simplesmente `LOCAL` se indicassem, por exemplo, as coordenadas geográficas. Para uma `PESSOA`, poder-se-ia especificar a data de nascimento, ou até uma pequena biografia; para uma obra, o seu ISBN ou a data da primeira edição; e para uma empresa, o seu número fiscal, por exemplo. Isto tornaria a tarefa mais realista, embora consideravelmente mais específica, e exigiria que os sistemas fizessem uso de almanaques muito maiores.

Sarmento e Mota (2006) sugeriram uma pista robusta, em que as maiúsculas ou minúsculas não importassem (apropriada, por exemplo, à detecção de entidades em texto transcrito automaticamente). De notar que nesse caso estamos a aproximarmo-nos do ACE, em que não só nomes próprios mas quaisquer referências/menções a entidades devem ser marcadas.

Mais uma vez, e embora tal já tenha sido aflorado no capítulo 4, convém relembrar que Mota, Bick, Sarmento e Almeida mencionaram o interesse de fazer algo semelhante ao MUC para poder ser comparável entre línguas – dada a repetição de afirmações como "para o inglês, o problema está resolvido a 95%, para o português ainda vamos a 70%",

afirmações essas que não são rigorosas mas que têm sido repetidamente feitas, como já referido em Cardoso (2006a, p. 85-87).

Pensamos que todos estes futuros (excepto o primeiro) dependem dos resultados das medições mencionadas na secção anterior, que nos permitirão ajuizar: o trabalho necessário, o esforço de anotação envolvido, e a necessidade de reformular ou não a arquitectura de avaliação e de criação de recursos.

Notamos também que, se não nos afastarmos demasiado do que já foi feito, os participantes em edições seguintes de uma avaliação conjunta têm a possibilidade de reutilizar os recursos criados na primeira para o treino dos seus sistemas. Essa é uma consideração que deve ser tida em conta antes de modificações demasiado radicais.

Agradecimentos

Este capítulo foi escrito no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC 339/1.3/C/NAC.

Parte II

Capítulo 8

O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa

Bruno Martins, Mário J. Silva e Marcirio Silveira Chaves

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 8, p. 97–112, 2007.

Os documentos textuais (por exemplo os artigos publicados em jornais ou páginas *web*) são muitas vezes ricos em informação geográfica, e principalmente relevantes a uma dada comunidade local (como textos noticiosos sobre eventos num local específico, ou um página *web* sobre um comerciante local). A utilização de técnicas de prospecção de texto para extracção desta informação, por forma a oferecer capacidades de raciocínio geográfico a sistemas de recuperação de informação, é um problema interessante que tem vindo a ganhar notoriedade (Amitay et al., 2004; Gey et al., 2006; Jones et al., 2004; Kornai e Sundheim, 2003; Purves e Jones, 2004).

Ao contrário dos sistemas de informação geográfica (SIGs) tradicionais, que lidam com dados estruturados e geo-referenciados, a área da recuperação de informação geográfica foca o tratamento de informação não estruturada (documentos textuais, por exemplo). O reconhecimento e desambiguação de nomes de locais em texto torna-se portanto uma tarefa crucial na geo-referenciação destes recursos de informação (por exemplo, a anotação dos documentos com os âmbitos geográficos que lhes correspondem) (Amitay et al., 2004; Densham e Reid, 2003). Foram já vários os projectos de investigação que abordaram os problemas relacionados com a interpretação de terminologia geográfica em texto (Kornai e Sundheim, 2003; Li et al., 2002; Olligschlaeger e Hauptmann, 1999; Smith e Mann, 2003; Smith e Crane, 2001; Schilder et al., 2004; Manov et al., 2003; Nissim et al., 2004; Leidner et al., 2003; Rauch et al., 2003). Contudo, um problema na área é a não existência de corpora apropriados para a avaliação destes sistemas (Leidner, 2004; Martins et al., 2005), contendo as referências geográficas devidamente anotadas com coordenadas geodésicas ou com os conceitos correspondentes numa ontologia.

Embora o problema geral do REM seja uma tarefa conhecida em extracção de informação (EI), o caso particular do tratamento de referências geográficas apresenta ainda novos desafios (Sang e Meulder, 2003; Kornai e Sundheim, 2003). Mais do que anotar uma expressão de texto como uma localização, pretende-se que seja feita a anotação de forma a que a expressão geográfica seja inequivocamente descrita (Kornai e Sundheim, 2003; Leidner et al., 2003). A desambiguação completa requer que as referências geográficas sejam classificadas de acordo com o tipo (por exemplo, cidade ou país) e associadas explicitamente a conceitos numa ontologia geográfica. Esta informação (a ontologia mais os documentos anotados) pode então ser utilizada noutras tarefas, tais como a indexação e recuperação de documentos de acordo com os seus âmbitos geográficos (Jones et al., 2004).

No âmbito do desenvolvimento de um motor de busca geográfico para a *web* portuguesa, resultante da extensão do já existente `www.tumba.pt`, foi desenvolvido o CaGE (CaGE é acrónimo de *Capturing Geographic Entities*). Por desambiguação, entendemos o processo de fazer a associação entre as referências geográficas que são reconhecidas nos textos com conceitos numa ontologia geográfica.

A metodologia proposta no nosso sistema REM assenta na existência de uma ontologia geográfica contendo os nomes de locais e outros tipos de informação associados (por

exemplo, relações topológicas entre eles). Faz ainda uso de "regras de contexto" (as quais combinam pistas internas e externas, através da utilização dos nomes de locais, expressões com uma conotação geográfica, e presença de maiúsculas no texto) por forma a fazer o reconhecimento destas EM nos documentos. A abordagem tem a vantagem de ser relativamente simples (e como tal rápida, adaptando-se ao processamento de grandes volumes de texto da *web*) e de não requerer quaisquer dados de treino, os quais podem ser difíceis de obter para línguas como o português. Posteriormente, a desambiguação dos nomes geográficos reconhecidos é baseada em heurísticas adicionais, tais como a hipótese do "um referente por discurso", semelhante à proposta por Gale et al. (1992).

Estudos anteriores demonstraram que transformar ontologias ou dicionários existentes em sistemas REM úteis, ou por outro lado pegar num sistema REM e incorporar informação de uma ontologia, são ambos problemas não triviais (Cohen e Sarawagi, 2004). Esta foi a principal razão que nos levou a não adoptar à partida por um dos sistemas REM *open-source* existentes, tais como o GATE (Cunningham et al., 2002). Embora tomando como ponto de partida os trabalhos anteriores e as melhores práticas da área do REM, escolhemos abordar o problema através da construção de um novo sistema de raiz, focando nos aspectos particulares do tratamento das referências geográficas e do desempenho computacional. Este último é um aspecto crucial no processamento de colecções de documentos do tamanho da *web*.

Neste capítulo é descrito a participação do sistema CaGE no HAREM. Embora o HAREM não seja apropriado para a avaliação da totalidade um sistema como o CaGE (como argumentado no capítulo 6), considerámos ser interessante a participação num cenário selectivo, que nos permitisse medir a eficácia do sistema no reconhecimento simples (sem qualquer classificação semântica ou desambiguação dos locais reconhecidos) de referências geográficas em textos na língua portuguesa. São aqui apresentados os resultados obtidos, discutindo-se as adaptações feitas no sistema por forma a cumprir os requisitos da tarefa de avaliação.

8.1 Conceitos e trabalhos relacionados

Como descrito no capítulo 6, a extração de referências geográficas em páginas *web* portuguesas levanta algumas considerações adicionais. Os sistemas de REM tradicionais combinam recursos lexicais com uma cadeia de operações de processamento de complexidade variável (alguns sistemas utilizam etapas de anotação de morfossintáctica ou de desambiguação do sentido das palavras), consistindo de pelo menos um atomizador, listas de nomes de entidades, e regras de extracção. A atomização parte o texto em segmentos (tais como palavras, números e pontuação). As regras para o reconhecimentos de EM são a parte central do sistema, combinando os nomes presentes nos léxicos com elementos tais como a presença de maiúsculas na palavra e o contexto em que as entidades ocorrem.

Estas regras podem ser geradas à mão (a abordagem baseada em conhecimento) ou automaticamente (aprendizagem automática). O primeiro método requer um perito humano, enquanto que o último visa a obtenção automática de regras, através da análise de corpora anotados.

Os melhores métodos de aprendizagem automática para reconhecer entidades mencionadas são usualmente testados em textos jornalísticos, tendo sido reportados resultados acima dos 90% em termos da medida F na tarefa partilhada do CoNLL (Sang e Meulder, 2003). Contudo, estas abordagens requerem dados de treino balanceados e representativos, sendo que um problema ocorre quando estes dados não estão disponíveis ou são difíceis de obter. Este é geralmente o caso com línguas diferentes do inglês, ou em tarefas bastante específicas, tais como a do reconhecimento de referências geográficas.

O grau em que os léxicos ou ontologias ajudam na tarefa de REM também parece variar. Por exemplo, Malouf (2002) reportou que os léxicos não melhoraram o desempenho, enquanto que outros estudos reportam ganhos significativos usando recursos lexicais e expressões simples para o reconhecimento (Carreras et al., 2002). Mikheev et al. (1999) mostraram que um sistema de REM sem um léxico podia até comportar-se bem em muitos tipos de entidades, mas este não é caso quando se trata de entidades geográficas. 11 das 16 equipas que participaram na tarefa de REM do CoNLL-2003 integraram recursos lexicais nos seus sistemas, e todos reportaram ganhos de desempenho (Sang e Meulder, 2003). Uma conclusão importante da tarefa partilhada do CoNLL-2003 foi a de que a ambiguidade em referências geográficas é bi-direccional. O mesmo nome pode ser usado para mais do que um local (ambiguidade no referente), e o mesmo local pode ser referenciado por vários nomes (ambiguidade na referência). Este último tipo tem ainda a variante do mesmo nome poder ser usado como uma referência quer a um local, quer a outro tipo de entidades tais como pessoas ou empresas (ambiguidade na classe da referência).

8.2 Os recursos lexicais usados pelo sistema CaGE

Ao contrário de uma tarefa de REM convencional, onde a utilização de padrões de reconhecimento é muitas vezes suficiente, para reconhecer e desambiguar referências geográficas temos normalmente de nos basear num recurso de informação externo (como um léxico ou uma ontologia geográfica). Ao lidarmos com referências geográficas em texto, o nosso verdadeiro objectivo é a utilização das referências geográficas noutras tarefas de recuperação de informação, sendo que as referências devem obrigatoriamente estar associadas a uma representação única para o conceito geográfico subjacente.

No contexto dos sistemas de prospecção de texto, as ontologias são uma boa alternativa em relação aos léxicos simples, uma vez que estas modelam não só o vocabulário como também as relações entre conceitos geográficos. Estas relações podem fornecer pistas úteis para heurísticas de desambiguação.

Ontologia de Portugal		Ontologia Mundial	
Componente	Valor	Componente	Valor
Conceitos	418,743	Conceitos	12,654
Nomes	419,138	Nomes	15,405
Adjectivos	0	Adjectivos	400
Relações	419,072	Relações	24,570
Tipos de conceitos	58	Tipos de conceitos	14
Relações parte-de	419,115	Relações parte-de	13,268
Relações de adjacência	1,132	Relações de adjacência	11,302
Conceitos do tipo NUT1	3	Conceitos do tipo ISO-3166-1	239
Conceitos do tipo NUT2	7	Conceitos do tipo ISO-3166-2	3,976
Conceitos do tipo NUT3	30	Aglomerações Populacionais	751
Províncias	11	Locais	4,014
Distritos	18	Divisões Administrativas	3,111
Ilhas	11	Cidades Capitais	233
Municípios	308	Continentes	7
Freguesias	4,260	Oceanos	2
Zonas	3,594	Mares	3
Localidades	44,386		
Arruamentos	146,422		
Códigos Postais	219,691		
Conceitos com coordenadas	9,254	Conceitos com coordenadas	4,204
Conceitos com caixas limitadoras	0	Conceitos com caixas limitadoras	2,083
Conceitos com dados demográficos	308	Conceitos com dados demográficos	8,206
Conceitos com frequência do nome	0	Conceitos com frequência do nome	10,067

Tabela 8.1: Caracterização estatística das ontologias usadas no sistema CaGE.

No contexto do CaGE e do desenvolvimento de um motor de busca geográfico, duas ontologias foram criadas, para tal consolidando-se informação de diversas fontes de dados públicas. Uma das ontologias considera informação geográfica de âmbito global, enquanto que a outra foca o território português, a um maior nível de detalhe. Estes dois recursos influenciam claramente as experiências com o sistema, e deve portanto ser feita a sua caracterização. A informação considerada nas ontologias inclui nomes de locais e outros conceitos geográficos, adjectivos de local, tipos de locais (por exemplo, distrito, cidade ou rua), relações entre os conceitos geográficos (por exemplo, adjacente ou parte-de), dados demográficos, frequência em textos *web*, e coordenadas geográficas sob a forma de centróides e caixas limitadoras ("*bounding boxes*"). A Tabela 8.1 apresenta algumas estatísticas, sendo que em Chaves et al. (2005) é apresentada informação mais detalhada.

Cada conceito geográfico pode ser descrito por vários nomes. A Figura 8.1 ilustra a repetição de nomes geográficos nas duas ontologias. Para cada nome, são contados o número de conceitos diferentes que lhe correspondem. No caso da ontologia de Portugal,

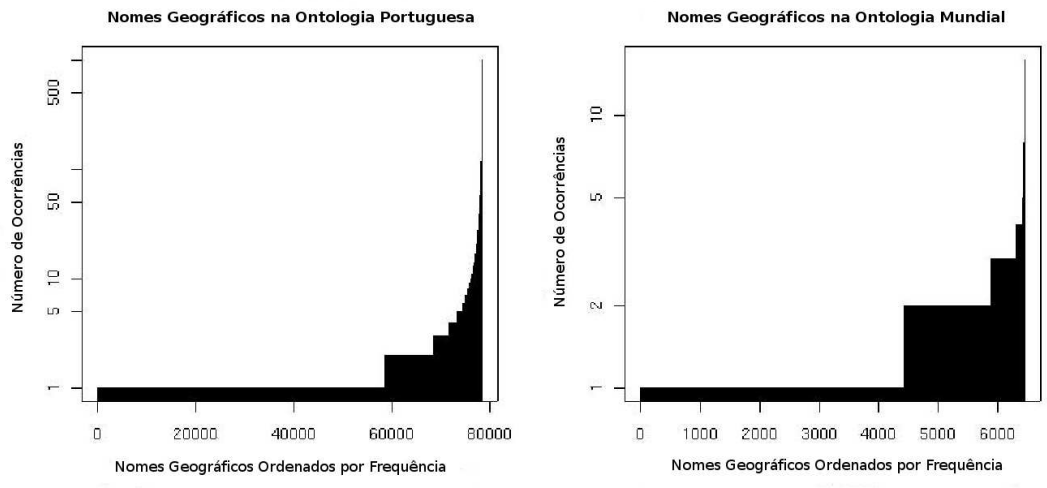


Figura 8.1: Frequência de repetição dos nomes geográficos nas ontologias.

os conceitos correspondentes a códigos postais não são apresentados, uma vez que eles já são por definição únicos e sem ambiguidade, e iriam confundir a interpretação do gráfico. As curvas apresentadas seguem a lei de Zipf (Zipf, 1949), como já notado em Li (1992), no sentido em que existe um número pequeno de nomes frequentes e uma longa lista de nomes pouco frequentes. Contudo, a Figura 8.1 também mostra que o número de nomes com múltiplas ocorrências (como a ambiguidade no referente) não é apenas um problema teórico, uma vez que eles correspondem a uma parte significativa dos nomes nas ontologias. A Tabela 8.2 apresenta exemplos de nomes geográficos comuns, correspondendo a vários conceitos.

A Figura 8.2 reforça as dificuldades associadas à utilização de nomes geográficos, desta feita mostrando a necessidade de considerar nomes compostos por múltiplas palavras. A figura separa a terminologia simples (ou seja, nomes geográficos compostos de apenas uma palavra), os nomes compostos (ou seja, nomes com várias palavras) e os casos difíceis (ou seja, nomes com hífen, abreviaturas e caracteres não alfa-numéricos). Mais uma vez, os códigos postais não são contabilizados, facilitando a interpretação do gráfico. Facilmente se pode observar que uma parte significativa dos nomes geográficos são compostos por mais do que uma palavra. As diferenças entre as duas ontologias advêm do facto da ontologia mundial conter apenas locais importantes (tais como países e cidades capitais), tendo portando um número maior de nomes simples.

Mesmo nos casos dos nomes simples podemos encontrar ambiguidade, visto que estes nomes também podem ser usados noutros contextos. Exemplos de palavras muito fre-

Ontologia de Portugal		Ontologia Mundial	
Nome do local	Número de locais	Nome do local	Número de locais
1 de Maio	618	Central	16
25 de Abril	881	Granada	10
Almada	15	Madrid	5
Bairro Alto	28	Portugal	4
Braga	11	Rio de Janeiro	4
Campo Grande	20	Roma	4
Lisboa	41	Taiwan	4
Seixal	42	Venezuela	4
Vila Franca	16	Washington	6

Tabela 8.2: Exemplos de nomes geográficos e o número de conceitos correspondentes nas ontologias portuguesa e mundial.

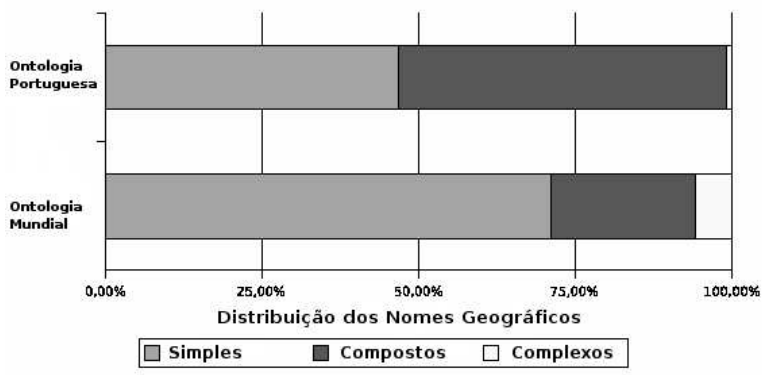


Figura 8.2: Distribuição dos nomes geográficos nas ontologias considerando a sua complexidade.

quentes que são também nomes geográficos são apresentados na Tabela 8.3. A mesma tabela mostra ainda que os nomes geográficos são muitas vezes homónimos com outros tipos de entidades, tais como pessoas (ou seja, ambiguidade na classe da referência). Por forma a lidar com este último tipo de ambiguidade, gerámos uma lista de excepções, com nomes que embora possam ter uma conotação geográfica, são muito mais frequentemente usados noutros contextos. Esta lista foi compilada através das nossas experiências (nomes que eram incorrectamente anotados foram colocados na lista), e através de um procedimento simples baseado em estatísticas num corpus da *web* (por exemplo, nomes que aparecem mais frequentemente escritos só em minúsculas do que com maiúsculas presentes foram adicionados à lista, seguindo a ideia que a detecção de letras maiúsculas pode distinguir entidades mencionadas).

Além da ontologia geográfica e da lista de excepções, a nossa técnica requer ainda

Palavras frequentes	Nomes de pessoas	
Homónimas com locais	Nome próprio	Nome de local
Central	Camilo Castelo Branco	Castelo Branco
Cruz	Cesária Évora	Évora
Direita	Teófilo Braga	Braga
Sol	Vergílio Ferreira	Ferreira
Nova	Irene Lisboa	Lisboa
Paz	Faria Guimarães	Guimarães
Casal	Almada Negreiros	Almada
Esta	Salgueiro Maia	Maia
Meio	Leonardo Coimbra	Coimbra

Tabela 8.3: Palavras frequentes e nomes de pessoas que incluem nomes de locais.

Tipo de expressão	Expressão
Identificadores	cidade, município, distrito, rua, avenida, rio, ilha, montanha, vale, país, continente, zona, região, condado, freguesia, deserto, província, povoado, aldeia, monte, vila, república, península
Localização	fora de, nos arredores de, dentro de, entre, em acima, ao longo, atrás, acima, ao lado, à esquerda, à direita
Distância Relativa	adjacente, longe de, perto de, próximo de
Orientação	este, norte, sul, oeste, oriente, ocidente, sudeste, sudoeste, nordeste, noroeste
Outras Expressões	"cidades como", "e outras cidades", "cidades, incluindo", "cidades, especialmente", "uma das cidades", "cidades tais como", padrões semelhantes para outros identificadores

Tabela 8.4: Expressões de contexto associadas a referências geográficas.

regras para efectuar o reconhecimento e desambiguação. Estas regras combinam pistas internas e externas, disparando quando um nome candidato está perto de uma expressão de contexto sugestiva. Estudos anteriores mostraram que as referências geográficas contêm muitas vezes informação sobre o tipo de locais a que se referem (por exemplo, *cidade de Lisboa*), sendo portanto passíveis de ser reconhecidas desta forma. As referências geográficas podem também conter expressões que denotem relações de distância ou de posicionamento relativo. A Tabela 8.4 exemplifica as expressões consideradas no desenvolvimento do CaGE, tendo essa lista sido baseada em trabalhos anteriores (Delboni, 2005; Kohler, 2003).

8.3 Reconhecimento e desambiguação de referências geográficas

A Figura 8.3 ilustra o procedimento utilizado pelo CaGE para identificar e desambiguar referências geográficas em texto, reflectindo os seus quatro estágios principais: pré-processamento, identificação, desambiguação e geração de anotações. O resto desta secção descreve cada um destes estágios em detalhe.

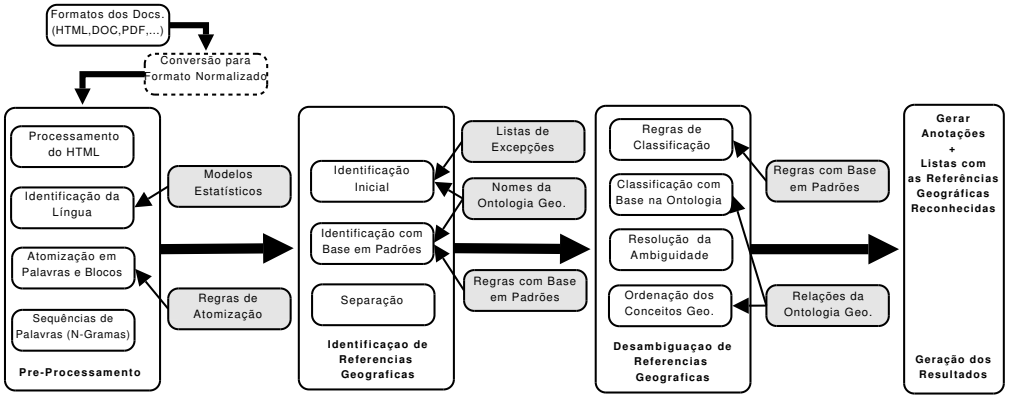


Figura 8.3: Arquitectura geral do sistema CaGE.

8.3.1 Operações de pré-processamento

A etapa de pré-processamento envolve as seguintes sub-etapas: conversão de formatos, processamento do HTML, classificação de língua, atomização e emparelhamento de *n*-gramas. As três primeiras são específicas do tratamento de textos provenientes da *web* no contexto do motor de busca geográfico. Estas foram desactivadas no contexto da produção de saídas para o HAREM, uma vez que apenas estávamos na presença de ficheiros de texto simples escritos na língua portuguesa.

A atomização das palavras e reconhecimento de frases é baseada numa tabela com os "pares de contexto" formados pelos caracteres que ocorrem antes e depois de uma dada posição no texto. Por exemplo, uma tabela para o reconhecimento de palavras coloca uma interrupção entre caracteres de pontuação e letras, mas não entre letras consecutivas ou entre caracteres de espaçamento consecutivos. As regras consideradas baseiam-se nas propostas pela Linguateca para o tratamento de corpora no projecto AC/DC (Santos e Sarmento, 2003), e descritas em `http://acdc.linguateca.pt/acesso/atomizacao.html`. Esta técnica lida com a grande maioria dos problemas de ambiguidade que ocorrem na atomização. É também simples de implementar, uma vez que a tabela de "pares de contexto" é simplesmente uma matriz de valores booleanos, em que cada linha e coluna correspondem

a um caracter ou grupo de caracteres. Um eixo representa o contexto anterior à posição, e o outro o contexto depois.

Depois do texto atomizado, as frases são divididas nos seus *n*-gramas constituintes. Isto é conseguido movendo uma janela sobre o texto de cada frase, tomando-se todas as possíveis sequências de *n* palavras consecutivas.

8.3.2 Identificação de referências geográficas

A etapa de identificação envolve a detecção de todas as sequências de *n*-gramas passíveis de constituir uma referência geográfica. Esta consiste de três sub-etapas, nomeadamente identificação inicial, identificação baseada em padrões e separação.

A identificação inicial envolve a aplicação de regras que combinam os nomes de locais na ontologia, expressões de contexto, e termos com a primeira letra em maiúsculas. As sequências de *n* palavras consecutivas identificadas na primeira etapa são inicialmente mapeadas nos nomes existentes na ontologia. Esta abordagem simples é suficiente para fazer a detecção de muitas referências, mas a ambiguidade pode conduzir a muitos erros. Por esta razão, apenas permitimos a detecção desta forma para certos tipos de conceitos geográficos na ontologia, particularmente os tipos que correspondem a regiões grandes e importantes (por exemplo, países e cidades om mais de 100.000 habitantes). Descartam-se ainda nesta fase de detecção simples os nomes geográficos presentes numa lista de excepções. Esta lista de exclusão tenta lidar com o problema de nomes muito frequentes que são usados noutros contextos que não o geográfico.

Dadas as limitações da identificação inicial, a sub-etapa seguinte usa regras para combinar os nomes geográficos com expressões de contexto e termos em maiúsculas. A Tabela 8.4 ilustra as expressões de contexto que foram consideradas. Algumas destas regras são relativamente complexas, combinando diferentes referências (por exemplo, *cidades tais como A, B ou C*) ou qualificando referências geográficas de acordo com critérios espaciais ou de posicionamento (por exemplo, *perto da cidade de X*). Contudo, o algoritmo de aplicação de regras, implementado por um autómato finito, é rápido. As regras são especificadas num ficheiro de texto, encontrando-se codificadas numa linguagem semelhante à das expressões regulares (as diferenças prendem-se com a utilização da informação de maiúsculas e dos nomes na ontologia).

É de notar que as regras consideradas na a geração de saídas para o HAREM têm algumas diferenças em relação às regras consideradas para a utilização normal do sistema. Em particular, fazemos para o HAREM um uso diferente do termos em maiúsculas, no sentido em que as directivas de anotação indicam que todas as entidades devem obrigatoriamente ter a primeira letra maiúscula[1], enquanto que no contexto das páginas *web* consideramos que os locais ocorrem muitas vezes em minúsculas. Têm-se ainda que no contexto do

[1] **Nota dos editores:** Com algumas pequenas excepções, documentadas na secção 16.1.4.

HAREM estamos interessados em reconhecer locais que não se encontrem descritos na ontologia (ou seja, reconhecidos apenas pela aplicação de regras), enquanto que nas aplicações normais do CaGE estamos apenas interessados em locais que possam ser mapeados em identificadores na ontologia, por forma a serem posteriormente usados noutras tarefas.

Finalmente, na sub-etapa de separação, os *n*-gramas passíveis de constituírem mais do que uma referência geográfica são detectados e os problemas de separação são resolvidos. Se um *n*-grama constitui uma referência, então todos os seus *n*-gramas constituintes são descartados, mantendo-se apenas a referência para o mais geral. As expressões complexas (por exemplo, *cidades tais como A, B, C*) são, neste caso, tratadas como uma excepção, mantendo-se cada referência independentemente.

8.3.3 Desambiguação de referências geográficas

Depois das referências geográficas terem sido identificadas, segue-se uma etapa de desambiguação. Esta envolve quatro sub-etapas, nomeadamente aplicação de regras de classificação, classificação baseada na ontologia, comparação das referências ambíguas com as que já se encontram desambiguadas e ordenação dos conceitos geográficos correspondentes. As regras de classificação são baseadas nas expressões de identificação usadas na etapa anterior, uma vez que muitas referências contêm palavras que podem ser usadas para inferir o tipo implícito ao conceito geográfico referenciado (por exemplo, em *cidade de Lisboa*, sabemos que a referência diz respeito à cidade e não a outro conceito).

A classificação baseada na ontologia usa as relações semânticas presentes na mesma para determinar o tipo correcto das referências. Pode-se dar o caso simples da uma referência, contendo ou não o tipo geográfico correspondente, poder ser mapeada num único conceito. Contudo, quando mais do que um conceito da ontologia está potencialmente a ser referenciado, usamos a hipótese de "um referente por discurso" para tentar a desambiguação. A hipótese diz que uma referência geográfica feita na mesma unidade de texto (ou seja, no mesmo parágrafo) refere-se ao mesmo local, ou a locais relacionados. Hipóteses semelhantes já foram usadas no passado no problema da desambiguação do sentido das palavras (Gale et al., 1992). A existência de uma relação entre dois conceitos é dada pela ontologia, sendo que consideramos os casos em que o nome ambíguo é um nome alternativo, uma região mais geral, uma região equivalente, ou uma região adjacente a um outro nome que já se encontre desambiguado.

O último estágio faz a comparação das referências ainda não desambiguadas com outras que já o tenham sido. Esta comparação é feita usando variações dos nomes das referências ambíguas, por forma a lidar com o problema de nomes truncados ou erros ortográficos. A comparação entre dois nomes é feita de acordo com as seguintes regras:

- Ambos os nomes devem ter o mesmo número de palavras.
- Maiúsculas, acentos e hífens são todos ignorados ao fazer a comparação.

- Palavras abreviadas são equivalentes (por exemplo, *Lis.* é dito equivalente a *Lisboa*).
- Palavras não abreviadas devem divergir no máximo em um caracter diferente, um caracter extra, ou um caracter a menos (por exemplo, *Lisboa* é dito equivalente a *Lusboa*).

Finalmente, nos casos não cobertos pelas heurísticas acima, mantemos a associação com todos os conceitos possíveis da ontologia. No entanto, ordenamos os conceitos possíveis de acordo com a importância do conceito geográfico referenciado, de acordo com as seguintes heurísticas:

- Regiões maiores (conceitos de topo na ontologia) são preferidas, uma vez que é mais provável que sejam mencionadas.
- Regiões com maior população são preferidas, pela mesma razão.

Em aplicações que requeiram a associação de cada referência a um único conceito, podemos usar estas heurísticas para escolher qual a referência mais provável, em lugar de manter a associação a todos os conceitos (Leidner et al., 2003).

8.3.4 Geração de anotações para a ontologia

A última etapa prende-se com a geração das saídas, mantendo-se cada referência geográfica associada com os conceitos correspondentes na ontologia. O formato usado pelo CaGE facilita o desenvolvimento de outras ferramentas de recuperação de informação, as quais usem as referências geográficas extraídas dos textos.

Sistemas anteriores optaram por associar a cada referência as coordenadas geodésicas correspondentes (Leidner et al., 2003), mas no CaGE optamos por associar as referências aos identificadores dos conceitos na ontologia. Isto traz algumas vantagens, nomeadamente ao permitir lidar com regiões imprecisas, ou no facto de não precisarmos de lidar com questões de precisão numérica associadas às coordenadas. Além de anotar cada referência com os conceitos na ontologia, mantemos ainda a associação com o tipo de conceito geográfico. O texto é anotado com etiquetas SGML correspondendo aos locais reconhecidos, tal como no seguinte exemplo:

```
O tempo de viagem entre a <PLACE type=administrative
subtype="city" geoid="GEO_146">cidade de Lisboa</PLACE> e a
<PLACE type=administrative subtype="city" geoid="GEO_238">cidade
do Porto</PLACE> é de duas horas e meia.
```

Além das anotações SGML, há ainda a possibilidade de gerar uma lista com todos os identificadores da ontologia reconhecidos no texto, assim como a frequência de ocorrência correspondente. Esta lista será a preferencialmente usada por outras ferramentas de recuperação de informação que façam uso das referências geográficas.

Para o HAREM foi necessário converter o formato SGML do nosso sistema no formato aceite pelo evento (ver capítulo 16). Para o mesmo exemplo fornecido acima, a anotação HAREM é a seguinte:

```
O tempo de viagem entre a cidade de <LOCAL>Lisboa</LOCAL> e
a cidade do <LOCAL>Porto</LOCAL> é de duas horas e meia.
```

Note-se que os tipos considerados pelo HAREM para a classificação semântica dos locais não se mapeavam directamente na nossa ontologia. Não foi tentado nenhum mapeamento dos nossos tipos de classificação para os considerados pelo HAREM, pelo que apenas participamos num cenário selectivo de identificação de EM de categoria LOCAL, sem qualquer classificação semântica. Outra das adaptações necessárias prende-se com o facto de as directivas para a anotação do HAREM especificarem que não se deve incluir os prefixos em minúsculas (tal como *cidade de*) como parte das anotações HAREM.

8.4 Experiências de avaliação no Mini-HAREM

Tal como descrito anteriormente, a nossa participação no HAREM limitou-se a num cenário selectivo de identificação de EM de categoria LOCAL, visto a colecção dourada e as directivas de anotação não considerarem a classificação semântica das entidades geográficas de acordo com os tipos geográficos usados no nosso sistema, nem muito menos a associação das mesmas com os conceitos geográficos da nossa ontologia.

Participámos na primeira edição do HAREM com uma versão inicial do sistema, mas neste capítulo apenas descrevemos os resultados obtidos na segunda edição do evento (o Mini-HAREM), onde os resultados obtidos com uma versão do sistema significativamente melhorada foram consistentemente melhores.

Para o Mini-HAREM foram geradas duas saídas. Uma delas corresponde à utilização da ontologia portuguesa, tal como descrita na secção 8.2, e a outra corresponde à utilização de uma ontologia conjugando as ontologias portuguesa e mundial. Aquando da primeira edição no HAREM, e por inspecção da colecção dourada usada como recurso de avaliação, verificámos que muitos dos locais anotados correspondiam a países e cidades internacionais importantes. Como o nosso sistema está fortemente dependente da ontologia, pensamos que a ontologia portuguesa seria insuficiente para um bom desempenho do sistema. Nas Tabela 8.5 e 8.6 é feito um resumo dos resultados obtidos por cada uma das saídas. A Tabela 8.6 apresenta ainda os melhores resultados obtidos no evento de acordo com as várias medidas de avaliação consideradas.

Da análise das tabelas ressalta que os resultados obtidos são aceitáveis em termos de precisão e abrangência no reconhecimento simples de EM de categoria LOCAL. Observa-se ainda que a segunda saída, gerada com uma ontologia com nomes de locais estrangeiros, é consistentemente melhor.

	Total	Identificados	Correctos	Parcialmente Correctos	Espúrias	Em Falta
Saída 1	893	686	469 (52,5%)	50 (5,6%)	169 (18,9%)	379 (42,4%)
Saída 2	893	696	486 (54,4%)	49 (5,5%)	163 (18,2%)	363 (40,6%)

Tabela 8.5: Número de EM de categoria *LOCAL* reconhecidos nas saídas para o Mini-HAREM.

	Precisão	Abrangência	Medida F	Erro Combinado	Sobre-geração	Sub-geração
Saída 1	69,78%	53,61%	0,6063	0,5514	0,2464	0,4244
Saída 2	71,17%	55,47%	0,6235	0,5331	0,2342	0,4065
Melhor resultado	92,07%	73,91%	0,7085	0,4398	0	0,2290

Tabela 8.6: Resultados obtidos no Mini-HAREM.

No que diz respeito ao desempenho computacional, e usando um PC Intel Pentium 4 com o sistema operativo Linux e 2 GB de RAM, o CaGE procedeu à anotação do texto a um débito de sensivelmente 50 KB de texto por segundo.

Embora o sistema CaGE tenha ficado ligeiramente aquém dos melhores resultados, importa frisar que a tarefa proposta pelo HAREM é ligeiramente diferente da tarefa de anotação executada pelo CaGE[2]. Em primeiro lugar, as EM na colecção dourada anotadas como `<LOCAL TIPO="CORREIO">` e correspondentes a moradas completas (por exemplo, a morada *Rua 25 de Abril, 77 R/C ESQ - Cruz de Pau - 2840 Seixal*) eram apenas parcialmente reconhecidos pelo nosso sistema (ou seja, este reconhece as entidades *Rua 25 de Abril*, *Cruz de Pau* e *Seixal* separadamente). A tarefa de reconhecimento de moradas completas não foi considerada durante o desenvolvimento do CaGE. Existe muita variabilidade nas expressões deste tipo, levando a um elevado custo computacional para a execução da tarefa.

Em segundo lugar, as EM anotadas na colecção dourada como `<LOCAL TIPO="VIRTUAL">` não eram reconhecidos pelo nosso sistema, visto estes muitas vezes não corresponderem a qualquer localização física. Os locais de tipo virtual podem dizer respeito a endereços electrónicos ou a sítios abstractos com função de alojamento de conteúdos, tais como jornais ou programas de televisão. Uma vez que estes locais não têm interesse no contexto da utilização num motor de busca geográfico, o sistema CaGE nunca foi concebido para reconhecer este tipo de entidades.

Em terceiro lugar, as EM anotadas na colecção dourada como `<LOCAL TIPO="ALARGADO">` também não eram reconhecidos pelo nosso sistema. De acordo com as directivas de anotação, estes locais correspondem a edificações ou pontos de referência tais como bares, hotéis ou centros de congressos. Este caso particular, e visto

[2] **Nota dos editores:** O facto de três subtipos de `LOCAL` contemplados no HAREM não interessarem ao CaGE teria sido razão para que este concorresse ao HAREM apenas no cenário selectivo `LOCAL` (`ADMINSTRATIVO; GEOGRAFICO`).

que estes locais têm uma correspondência física, trata-se de uma limitação do nosso sistema, sendo que numa versão futura pretendemos também fazer o reconhecimento e desambiguação destes casos.

Num cenário selectivo correspondente apenas à anotação de entidades do tipo <LOCAL TIPO="ADMINISTRATIVO"> e <LOCAL TIPO="GEOGRAFICO">, a melhor saída do CaGE teria obtido uma precisão e abrangência de 67,1% e 66,5%, respectivamente. É ainda de salientar que o CaGE teria detectado um total de 27 ocorrências apenas parcialmente correctas, apesar de neste cenário não estarem a ser considerados locais do tipo ALARGADO ou CORREIO. Num mesmo cenário, o melhor sistema a concurso no HAREM teria obtido uma precisão e abrangência de 82,8% e 61,6%, respectivamente. Estas diferenças entre os dois sistemas estão relacionadas quer com limitações do sistema CaGE no reconhecimento de algumas entidades, quer com o facto de as directivas de anotação do HAREM diferenciarem os nomes de locais que assumem no texto um papel semântico diferente.

Pelas razões apresentadas, parece-nos importante que uma futura edição do HAREM considere o caso das referências geográficas de uma forma diferente, através da utilização de anotações na colecção dourada que sejam mais precisas e que melhor reflictam a temática geográfica. Este tema foi já desenvolvido no capítulo 6, por isso não o repetiremos aqui.

8.5 Conclusões

Este capítulo descreveu o sistema CaGE para o reconhecimento, classificação e desambiguação de referências geográficas em textos na língua portuguesa. O mesmo foi desenhado segundo métodos rápidos e simples, por forma lidar de forma robusta com grandes quantidades de documentos. O reconhecimento de referências geográficas é apenas um meio para outras utilizações em ferramentas de recuperação de informação conscientes da geografia. A abordagem aqui descrita é parte de um projecto de âmbito mais largo, visando a construção de um motor de busca geográfico para a *web* portuguesa, baseado na atribuição de âmbitos geográficos aos documentos. Este motor de busca, e consequentemente a abordagem descrita neste capítulo, foi usado no contexto das edições de 2005 e 2006 do GeoCLEF, uma avaliação conjunta semelhante ao TREC dedicada aos sistemas de recuperação de informação geográficos (Gey et al., 2006; Martins et al., 2007).

Para o evento de avaliação HAREM foram feitas algumas adaptações ao sistema, por forma a testar o desempenho do mesmo num cenário selectivo de reconhecimento simples de EM de categoria LOCAL. Neste capítulo apresentamos os resultados obtidos pelo nosso sistema no Mini-HAREM, sendo ainda discutidas as limitações no evento no que diz respeito à avaliação de sistemas focados no tratamento de referências geográficas. Em futuras edições do HAREM, gostaríamos de ver o cenário das referências geográficas tratado

em maior profundidade, nomeadamente através da anotação da colecção dourada de uma forma mais precisa.

A nossa participação no HAREM indicou resultados aceitáveis em termos de precisão e abrangência no reconhecimento de referências geográficas, embora exista ainda lugar para diversos melhoramentos. Estudos adicionais com outras colecções de documentos, maiores e devidamente anotadas com referências geográficas, são quanto a nós necessários para se tirarem mais conclusões.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia, através do projecto com referência POSI/SRI/40193/2001 e da bolsa de doutoramento com referência SFRH/BD/10757/2002.

Capítulo 9

O Cortex e a sua participação no HAREM

Christian Nunes Aranha

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 9, p. 113–122, 2007.

O Cortex é um sistema de inteligência artificial desenvolvido a partir de minha tese de doutorado na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Em minha tese desenvolvi o esboço teórico e implementei a primeira versão a qual participou do HAREM, e hoje já se encontra em sua versão 3.0.

O Cortex nasceu com a ambição de simular as faculdades cognitivas de PLN. Isto significa dizer que seu maior objetivo é a eficiente manipulação da linguagem humana, tanto na leitura, codificação e interpretação de textos como na produção. Acreditamos que se nos aproximarmos cada vez mais do processo cognitivo humano, teremos cada vez melhores resultados.

Nós, da Cortex, entendemos que a produção eficiente tem como pré-requisito uma boa leitura. Sendo assim, não trabalhamos com produção ainda (apenas de resumos). Da mesma forma, para uma boa leitura, é necessário um bom conhecimento das palavras, dos seus significados e da gramática de uma língua, em princípio nesta ordem. Logo, o Cortex é um processador dependente da língua, o que está alinhado com nossos objetivos finais, já que, nós, seres humanos também somos dependentes da língua, porém, com capacidade de aprender novas. Assim como deve ser o Cortex.

9.1 Filosofia

Em psicologia do desenvolvimento humano vemos que bebês/crianças manifestam espontaneamente a capacidade de adquirir (e não aprender) a linguagem sozinhas, simplesmente ao ouvir frases e pequenos textos falados provenientes em grande parte de seus pais. Mais tarde, utilizando essa linguagem "adquirida", irão então, não adquirir, mas "aprender" (por exemplo, na escola) a língua escrita. Aprender porque precisam de um professor para ensinar. Seres humanos não costumam ter a capacidade espontânea de ler e escrever.

Adicionalmente, parece que a explicação natural para a ordem do áudio-visual, ou seja assimilar primeiro o som e só depois a imagem, está contida no domínio biológico já que existe uma conversão quase que direta entre uma mensagem falada e uma escrita. Isso nos leva a crer que, se existe um processo para adquirir a fala, há de haver um para adquirir textos também.

Inspirado nestas observações empíricas, o sistema Cortex surge, então, para responder à seguinte pergunta: Que "programa" haveriam estes bebês de processar para adquirir a linguagem através do som? E mais, que programa seria rodado, para com isso adquirir novas palavras?

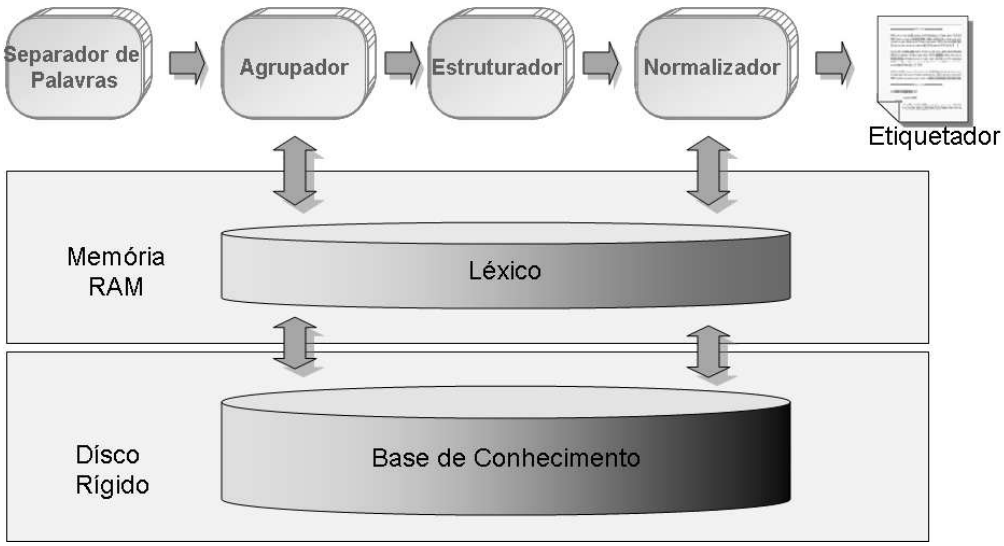


Figura 9.1: Etapas de processamento do texto no Cortex.

9.2 Classificação de entidades mencionadas no Cortex

O Cortex é um sistema computacional para o processamento da língua, cujo algoritmo reproduz alguns comportamentos lingüísticos de um falante, como sua adaptabilidade, flexibilidade, e capacidade de antecipar, pressupor e rever suas hipóteses.

Dessa maneira, o processamento do Cortex é feito em várias etapas, como mostra a Figura 9.1. Cada etapa é capaz de rever os passos anteriores e influir sobre os subseqüentes. Após a separação inicial das palavras, a etapa seguinte consiste em reconhecer as entidades que possam ser constituídas por mais de uma palavra. Substantivos compostos e locuções são descobertos nesse momento. O processo de reconhecimento dos termos é feito com o auxílio de um autômato escrito para identificar padrões de formação de entidades compostas com base num repertório de regras. O resultado dessa etapa é adicionada ao conhecimento existente no léxico, e posteriormente à base de dados.

O próximo passo constitui na classificação dos termos previamente extraídos. Sabendo-se que a criatividade lingüística é de suma importância na produção textual, o Cortex recorre a um banco de informações lexicais com certa parcimônia. As informações armazenadas sobre uma palavra (sua classe, significado, etc.) são tomadas apenas como um dado *a priori*, que pode ser questionado e reavaliado por outras circunstâncias a que esta palavra se vê envolvida no texto. O resultado disso é que o Cortex se torna um mecanismo provido de experiência, ou seja, quanto mais texto processa mais conhecimento lingüístico ele acumula e mais poder de inferência ganha para processar novas informações/textos.

Além disso, o Cortex obtém as informações de quatro fontes de dados, como mostra a

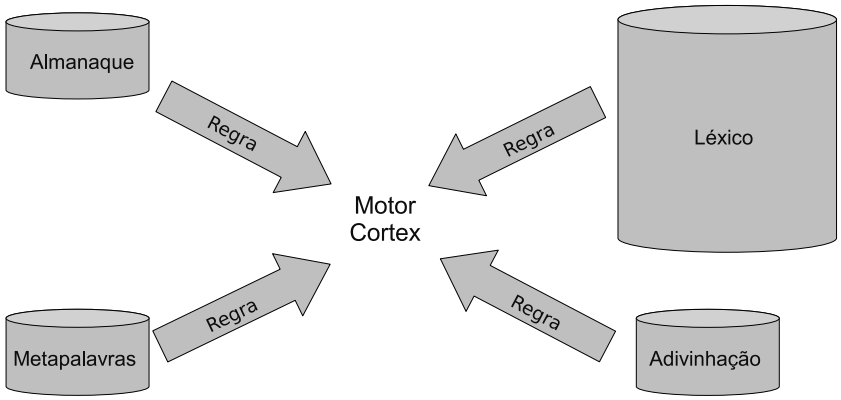


Figura 9.2: Fontes de dados do Cortex.

Figura 9.2: o Almanaque, que contém uma lista de entidades de uma determinada categoria provenientes de uma fonte enciclopédica; o Metapalavras, constituído por uma lista de termos que aparecem nas vizinhanças das entidades, por exemplo, *pianista*, *jogador*; a Adivinhação, que contém um conjunto de termos que constituem as entidades mencionadas, por exemplo, *Prof.*, *Dr.*, *Presidente*; e o Léxico, que armazena todo o conhecimento aprendido pelos textos já processados pelo Cortex.

Cada uma das fontes infuencia a tomada de decisão do Cortex quanto à identificação e classificação de EM. Cada regra traz consigo uma probabilidade associada, que é usada pelo Motor Cortex. Em paralelo a esse sistema existem máquinas de estimação de novas regras e probabilidades. Exemplos de aplicação das quatro fontes de dados são:

Categoria Pessoa

```
Entrada: O acordeonista Miguel Sá(...)
Saída:   O acordeonista <PESSOA TIPO="INDIVIDUAL">Miguel Sá</PESSOA>(...)
```

onde *acordeonista* é um termo obtido da fonte de dados Metapalavras, associado à pessoa *Miguel Sá*.

```
Entrada: Na pesquisa do Dr. Lewis(...)
Saída:   Na pesquisa do <PESSOA TIPO="INDIVIDUAL">Dr. Lewis</PESSOA>(...)
```

onde *Dr.* é uma evidência obtida através da lista Adivinhação que indica probabilidade para nome de pessoa. No modelo original do Cortex, *Dr.* não faz parte da EM. A entidade final é *Vernard Lewis* obtida pela regra de co-referência. Especialmente para o HAREM adicionamos um novo conjunto de regras que juntava `TITLE + NOME` e produzia a etiqueta SGML final.

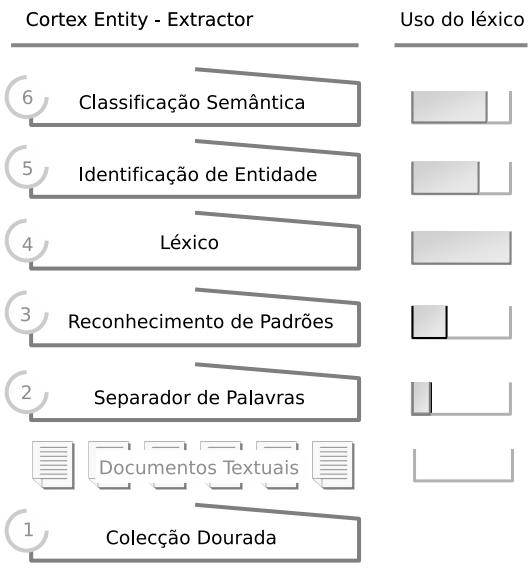


Figura 9.3: Etapas de avaliação de documentos no Cortex.

Categoria Local

```
Entrada: (...)Pela primeira vez no Haiti um padre foi assassinado
         por motivos políticos(...)
Saída:   (...)Pela primeira vez no <LOCAL TIPO="ADMINISTRATIVO">Haiti</LOCAL>
         um padre foi assassinado por motivos políticos(...)
```

onde *Haiti* pode ser primeiramente aprendido pela fonte Almanaque e depois passa para a fonte Léxico.

A Figura 9.3 apresenta todas as etapas as quais os documentos são submetidos ao Cortex, em particular o corpus Coleção HAREM, para se obter sua classificação. Na coluna à direita da figura é apresentado o percentual de uso do Léxico nas diferentes etapas.

O Cortex é composto pelo Separador de Palavras, que identifica cada termo (simples ou composto) como uma palavra; Reconhecimento de Padrões, que reconhece categorias ou classes de termos; o Léxico, que armazena as informações lingüísticas de cada termo; Identificador de Entidades, que identifica os limítes de cada entidade mencionada; o Classificador de Entidades, que finaliza o processo de reconhecimento da entidade atribuindo a ela um rótulo semântico dentro de uma ontologia pré-definida, gerando uma etiqueta SGML correspondente como formato de saída.

Medida	Cenário: TOTAL 1º Lugar	Cenário: SELECTIVO 1º Lugar	Cenário: TOTAL Resultados Cortex
Precisão	ELLE (80,64%)	PALAVRAS (78,50%)	CORTEX_NO (65,57%)
Abrangência	SIEMES1 (84,35%)	SIEMES1 (84,35%)	CORTEX_NO (86,69%)
Medida F	PALAVRAS (0,8061)	PALAVRAS (0,8061)	CORTEX_NO (0,7466)

Tabela 9.1: Vencedores da tarefa de identificação do HAREM (considerando apenas saídas oficiais), e resultados da saída não-oficial do Cortex.

9.3 A participação do Cortex no HAREM

O Cortex foi submetido à avaliação do HAREM nas seguintes tarefas e categorias:

- Tarefas efetuadas: identificação e classificação semântica de EM.
- Cenário seletivo: PESSOA, ORGANIZACAO, LOCAL, TEMPO, ACONTECIMENTO e VALOR.

O principal erro cometido foi, conjuntamente, a baixa flexibilidade do formato de saída de nosso sistema e a má interpretação das regras do HAREM. Não tinhamos muito tempo, começamos a estudar e trabalhar na avaliação poucos dias antes. Foi quando nos deparamos com a diferença entre a saída de nosso sistema e o formato padrão do HAREM.

O Cortex se aproximava da versão 1.0 e não tinha flexibilidade nenhuma de configuração das etiquetas de saída. A solução foi improvisar uma transformação do arquivo através de uma substituição manual, o que ocupava um tempo bastante grande. O Cortex imprimia a saída como PESSOA, se a entidade fosse classificada como pessoa, GEOGRAFIA, se a entidade fosse LOCAL, e ORGANIZAÇÃO idem, mas se não conseguiu classificar imprimia apenas NOME. Achávamos que só poderiamos concorrer nas tarefas de identificação e classificação semântica, e NOME não existia nas directivas do HAREM, sendo assim, optamos por retirar as entidades com marcação NOME e não marcar nada. No dia seguinte, lendo as regras com mais calma descobrimos a existência da etiqueta <EM>. Fizemos tudo novamente e entramos na avaliação não-oficial.

O prejuízo no resultado oficial foi grande porque nosso sistema de identificação estava razoável para a época, porém, nosso sistema de classificação tinha uma abrangência muito fraca e eliminou várias entidades que poderiam ter sido identificadas. Enfim, fazendo as contas considerando nosso resultado não-oficial, não ficaríamos em primeiro lugar total da medida F por outros problemas que explicarei a seguir, mas pelo menos ganharíamos o primeiro lugar em termos de abrangência no cenário seletivo, com 86,69% (acima de 84,35%, como mostra a Tabela 9.1).

Quanto ao desempenho por Género, apenas nos textos *correio eletrônico* teríamos obtido primeiro lugar na medida F. Em média, teríamos ficado em quarto lugar geral com nossa saída não-oficial.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: TOTAL Forma: RELATIVO 1º Lugar	Cenário: SELECTIVO Forma: ABSOLUTO 1º Lugar	Cenário: SELECTIVO Forma: RELATIVO 1º Lugar
Categorias	PALAVRAS (0,6301)	CORTEX2 (0,7171)	PALAVRAS (0,6301)	CAGE3 (0,8161)
Tipos	-	ELLE (0,8497)	-	NOOJ1 (0,8917)
Combinada	PALAVRAS (0,5829)	ELLE (0,6812)	PALAVRAS (0,5829)	ELLE (0,7327)
Plana	PALAVRAS (0,5293)	ELLE (0,6548)	ELLE (0,5487)	ELLE (0,7044)

Tabela 9.2: Vencedores para tarefa de classificação semântica do HAREM.

O resultado para a classificação semântica (Tabela 9.2) nos mostrou que a classificação tinha uma boa precisão, obtendo o primeiro lugar no cenário total relativo. Os outros problemas de padronização da saída que tivemos foi com relação aos números por extenso que não apresentam letra maiúscula são marcados como entidade do tipo valor pelo Cortex e não pelo HAREM, assim como as referência a tempo (por exemplo, *ontem* e *segunda-feira*). Em contrapartida perdemos muitos pontos pela identificação de *R:* nos textos de gênero *entrevista* que foi marcado porque tinha letra maiúscula, e de fato não faz sentido ser entidade. Finalmente, a titulação das pessoas como por exemplo *Sr.*, *Dom* ou *Dr.* são excluídas da entidade pessoa pelo Cortex, já que esses lexemas são classificados como metapalavras e não fazem parte da entidade, uma mera questão de configuração de saída, e foram consideradas pelo HAREM como parte da pessoa. Veja o exemplo:

```
HAREM:  Na pesquisa do <PESSOA TIPO="INDIVIDUAL">Dr. Lewis</PESSOA>(...)
CORTEX: Na pesquisa do Dr. <PESSOA TIPO="INDIVIDUAL">Lewis</PESSOA>(...)
```

Conclusão, o sistema como estava implementado, sem flexibilidade de configuração, seria impossível fazer essas modificações para o HAREM. Sendo assim, deu-se início ao trabalho do refatoramento para construir a versão 2.0.

9.4 A participação do Cortex no Mini-HAREM

A participação do Cortex no Mini-HAREM contou com a versão 2.0 de nosso sistema, onde havia principalmente flexibilidade de configuração para adequar a saida aos padrões do HAREM. Com isso conseguimos reduzir enormemente os erros de sobre-geração que tanto nos penalizou na primeira edição.

Para implementar a segunda versão e as seguintes foi necessário, não só o refatoramento da primeira versão, como o apoio de mais três membros.

Além disso, a versão 2.0 contava com um sistema de classificação bem mais evoluído, com mais estratégias cognitivas e também mais conhecimento lexical, dado que o sistema Cortex acumula o conhecimento a cada documento novo lido.

O Cortex foi então submetido à avaliação do Mini-HAREM nas seguintes tarefas e categorias:

Medida	TOTAL 1º Lugar	SELECTIVO 1º Lugar
Precisão	Cortex2CEM (87,33%)	Cortex2CEM (83,87%)
Abrangência	Cortex1REM (87,00%)	Cortex1REM (88,93%)
Medida F	Cortex1REM (0,8323)	Cortex1REM (0,7662)

Tabela 9.3: Vencedores da tarefa de identificação no Mini-HAREM.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: SELECT. Forma: ABSOLUTO 1º Lugar
Categorias	Cortex2CEM (0,6157)	Cortex2CEM (0,6839)
Tipos	-	-
Combinada	Cortex2CEM (0,5855)	Cortex2CEM (0,6501)
Plana	Cortex2CEM (0.5525)	Cortex2CEM (0,6145)

Tabela 9.4: Vencedores da tarefa de classificação semântica no Mini-HAREM.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: SELECT. Forma: ABSOLUTO 1º Lugar
HAREM	PALAVRAS (0,8061)	PALAVRAS (0,8061)
Mini-HAREM	Cortex1REM (0,8323)	Cortex1REM (0,7662)

Tabela 9.5: Comparação dos resultados HAREM e do Mini-HAREM para a tarefa de identificação.

Medida F	Cenário: TOTAL Forma: ABSOLUTO 1º Lugar	Cenário: SELECT. Forma: ABSOLUTO 1º Lugar
HAREM	PALAVRAS (0,6301)	PALAVRAS (0,6301)
Mini-HAREM	Cortex2CEM (0,6157)	Cortex2CEM (0,6839)

Tabela 9.6: Comparação dos resultados HAREM e do Mini-HAREM para a tarefa de classificação semântica, medida por categorias.

- Tarefas efetuadas: identificação e classificação semântica de EM.
- Cenário seletivo: `PESSOA`, `ORGANIZACAO`, `LOCAL`, `TEMPO`, `ACONTECIMENTO` e `VALOR`.

E obteve os resultados mostrados pelas Tabelas 9.3 e 9.4 para as avaliações de identificação e classificação respectivamente das quais participou.

Comparando os resultados do Mini-HAREM e os do HAREM, podemos fazer um *ranking* total, com todos os participantes (embora esta seja uma comparação bastante artificial,

Gênero	Precisão	Abrangência	Medida F
web	76,26%	81,97%	0,7901
correio eletrônico	64,80%	81,50%	0,7220
literário	79,29%	87,12%	0,8302
político	90,83%	90,83%	0,9083
expositivo	90,76%	91,59%	0,9117
técnico	38,81%	69,67%	0,4985
entrevista	93,40%	93,79%	0,9359
jornalístico	90,52%	94,24%	0,9234

Tabela 9.7: Comparativo dos resultados do Cortex segmentado por gênero.

Saída	Precisão (%)	Abrangência (%)	Medida F	Erro Combinado	Sobre-geração	Sub-geração
cortex3	57,12	73,54	0,6430	0,4969	0,3492	0,1743
cortex2cem	57,12	73,54	0,6430	0,4969	0,3492	0,1743

Tabela 9.8: Resultado para categoria `QUANTIDADE`.

porque compara desempenho sobre textos diferentes, de diferentes versões dos mesmos sistemas). Mas admitindo que essa comparação é válida, os resultados das Tabelas 9.5 e 9.6 mostram que o sistema Cortex obteve o primeiro lugar no cenário total absoluto para a tarefa de identificação, e o primeiro lugar no cenário selectivo absoluto para a tarefa de classificação semântica.

Nessa seção analisaremos os pontos críticos apontados pelos relatórios disponibilizados pela Linguateca. Esses serão os pontos de melhora para as próximas versões na intenção de aumentar a medida F.

O primeiro ponto crítico que vale a pena ressaltar foi o desempenho do Cortex no gênero *técnico*. A Tabela 9.7 mostra como o desempenho foi bem inferior aos demais.

Isso se deu em grande parte pelo reconhecimento dos subtítulos como entidades. Além de nomes de teorias e pessoas que acabaram dificultando a tarefa.

O segundo ponto crítico foi o desempenho semântico do Cortex na categoria `VALOR`, mostrado na Tabela 9.8. Analisando o arquivo de alinhamento, descobrimos que o Cortex considera *80 anos* (por exemplo) como `TEMPO` e não como `VALOR TIPO="QUANTIDADE"`, o que ocasionou uma baixa significativa na medida F.

Além desses pontos, vale destacar que o Cortex é treinado na língua portuguesa do Brasil e portanto, diversos verbos diferentes foram encontrados no início de frase, provocando uma confusão com uma entidade desconhecida.

Finalmente, cargos em letra maiúscula também foram descartados e serão configurados como `GRUPOCARGO` a partir de agora e números referentes a artigos que foram considerados como número e irão pra categoria `OBRA` para a próxima edição do HAREM.

9.5 Cortex 3.0

Os últimos resultados levam-nos a pensar que a utilização de almanaques é bastante interessante e útil no início do aprendizado do sistema, porém, conforme ele vai adquirindo inteligência gramatical, a utilização destes descresce bastante, e algumas vezes, acaba por prejudicar a precisão do sistema.

Por esse motivo, o foco do sistema Cortex é cada vez mais em cima das informações presentes no texto, ontologias e conhecimento enciclopédico. Procuramos atualmente um modelo de representação para o conhecimento abstrato extraído dos textos e que seje o mais interpretável possível de modo a aumentar o poder de gerenciamento do conhecimento acumulado.

9.6 Conclusões

Este capítulo descreve o sistema Cortex, um sistema baseado em inteligência artificial para o aprendizado, aquisição, reconhecimento e classificação de, não só entidades como também verbos, substantivos e adjetivos. Para as duas primeiras edições do HAREM, trabalhamos principalmente com em textos na língua portuguesa do Brasil.

O sistema foi projetado para integração com mecanismos de indexação, o que o torna completamente escalável para mineração de textos em grandes quantidades de documentos. A abordagem aqui descrita faz parte de um projeto maior de estruturação de dados não-estruturados. Isso significa extrair um modelo de representação semântico para ser usado em domínios como a Web Semântica. Esse mesmo sistema é usado na plataforma de inteligência competitiva da empresa Cortex Intelligence[1].

Para o HAREM foram feitas algumas adaptações ao sistema para atender a especificação da ontologia da avaliação, que difere em parte da utilizada por nós. Mesmo criando um módulo mais sofisticado de configuração da ontologia para o Mini-HAREM, vimos que ainda cometemos erros de transdução.

Os relatórios produzidos pela Linguateca ajudaram em muito o aperfeiçoamento de nosso sistema. Apontando detalhes que nos passavam desapercebidos, mostrando novos domínios de informação a serem explorados, assim como um panorama mundial do tratamento da língua portuguesa. Além, é claro, na produção de um corpus de treinamento para as próximas edições.

Estamos em constante melhoramento de nosso sistema, ainda temos muito a caminhar, principalmente para outras línguas. Em futuras edições do HAREM, gostaríamos de ver avaliações de anáforas e fatos.

[1] www.cortex-intelligence.com

Capítulo 10

MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish

Thamar Solorio

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 10, p. 123–136, 2007.

Due to the many potential uses of named entities (NE) in higher level NLP tasks, a lot of work has been devoted to developing accurate NE recognizers. Earlier approaches were primarily based on hand-coded knowledge, lists of gazetteers, and trigger words (Appelt et al., 1995; Krupka e Hausman, 1998; Black et al., 1998; Téllez et al., 2005). More recently, as machine learning has increased its popularity within the NLP community, NER systems are taking advantage of machine learning algorithms (Arévalo et al., 2002; Bikel et al., 1997, 1999; Borthwick, 1999; Carreras et al., 2002, 2003b; Madrigal et al., 2003; Petasis et al., 2000; Sekine et al., 2002; Zhou e Su, 2002). However, lists of trigger words and gazetteers remain a key component of these systems.

Newer approaches try to avoid limitations of language dependency by tackling NER on a multilingual setting (Carreras et al., 2003a; Curran e Clark, 2003; Florian et al., 2003; Maynard et al., 2003b), and although it is very unlikely that a general NER system performing well across all languages will exist in the near future, recent systems have successfully achieved higher portability than that of previous attempts. The main goal of this research work is to provide a representation of the learning task that increases coverage of a hand-coded NE tagger and evaluate its effectiveness and portability to different collections and languages. Our approach needs to be flexible and easy to port so that an average user can adapt the system to a particular collection or language. In a previous work we presented results of extending the coverage of a hand-coded tagger for Spanish to different texts (Solorio, 2005). Here we show how the same representation can be used to perform NE extraction in Portuguese without needing to adapt the task to Portuguese. Results presented here show that it is possible to perform NE extraction on both languages, Spanish and Portuguese, using the same design for the learning task.

The next section describes our framework for NE extraction. Section 10.2 presents the results of performing NE extraction on Portuguese using the framework previously described. The paper concludes by summarizing our findings.

10.1 The MALINCHE System

Similar to the strategy used by other researchers in previous approaches, we divide the NER problem into two sub-tasks that are solved sequentially:

1. We first determine which words, or sequences of words, are likely to be NEs. This task is called Named Entity Delimitation (NED).

2. Once we have extracted possible NEs from documents, we then try to categorize each NE into one of the following classes: `PERSON`, `ORGANIZATION`, `LOCATION` and `MISCELLANEOUS`. This task is called Named Entity Classification (NEC).

We decided to divide the problem in this way considering the unbalanced distribution of data. Normally, in a given document around 10%, or at most 18%, of words are NE.

This unbalanced distribution can cause trivial classifiers to achieve accuracies of up to 85% by tagging all words in the document as non-NE. We can circumvent this problem by carefully selecting the learning algorithm, or by assigning a cost matrix to the classification errors. Some authors, working with classification problems with similar conditions, have used the solution of selecting the training instances in an attempt to give the learner a well balanced training set. This can be achieved by means of *over-sampling*, where instances of the ill-represented classes are randomly selected and added to the training set (Ling e Li, 1998), or *under-sampling*, where random instances of the over represented class are removed to balance the distribution (Zhang e Mani, 2003). Whatever the alternative taken, we can not be certain that the bias for selecting the most frequent tag can be completely removed. Moreover, according to a study performed by Japkowicz (2003), when class imbalances cause low classification accuracies it is best to tackle the small disjunct problem (Holte et al., 1989) than to attempt to rectify the imbalances. Thus, even though for some works this condition does not seem to be a problem, for example (Borthwick, 1999), we opted for the strategy of performing NED first and then NEC. This separation of tasks will allow for different attributes for each task, and thus, we can tackle each subproblem using a different strategy.

The methods we developed for NED and NEC are very similar in spirit. In both cases we take advantage of the tags assigned by the hand-coded tagger[1] and use them together with some lexical features to train a learning algorithm. Our goal is to allow the classifier to take advantage of the knowledge the hand-coded tagger has about the NER task. Going a step further, we want the classifier to learn from the hand-coded tagger mistakes. This is why a key component in our method is precisely the output of the hand-coded tagger, because we believe it provides valuable information. In the following sections we describe in more detail the NED and NEC methods.

10.1.1 Named Entity Delimitation

As mentioned earlier, in this task we are concerned with extracting from documents the words, or sequences of words, that are believed to be NE. This extraction process can be performed by means of classifying each word in the document with a tag that discriminates NE. In our classification setting we use the BIO scheme, where each word is labelled with one of three possible tags, according to the following criteria:

- The B tag is for words that are the beginning of a NE.
- The I tag is assigned to words that belong to an NE, but they are not at its beginning.
- The O tag is for all other words that do not satisfy any of the previous two conditions. All words not belonging to NE are assigned the O tag.

[1] The hand-coded system used in this work was developed at the TALP research center by Carreras e Padró (2002).

Word	BIO Class
Monaco	B
was	O
in	O
mourning	O
for	O
the	O
death	O
of	O
Prince	B
Rainier	B
III	I

Table 10.1: An example of NED using the BIO classification scheme

Let D_R be the set of labelled documents that will be used for training
Let D_T be the set of test documents
TRAINING
1. Label D_R with PoS and NE tags using the hand-coded tagger
2. Transform the PoS tags assigned by the hand-coded tagger to the compact set of tags
3. Transform the NE tags from the output of the hand-coded tagger to BIO format
4. Build the training instances adding to the output of the hand-coded tagger the training attributes
5. Give the learning algorithm the training instances and perform training

TESTING
1. Label D_T with PoS and NE tags using the hand-coded tagger
2. Transform the PoS tags assigned by the hand-coded tagger to the compact set of tags
3. Build the test set adding to the output of the hand-coded tagger the training attributes
4. Transform the NE tags from the output of the hand-coded tagger to BIO format
5. Let the trained classifier label the test instances

Table 10.2: The NED algorithm.

An example of a possible output for this classification setting is shown in Table 10.1. Here we present a sentence where each word is classified under the BIO scheme.

The algorithm for NED is summarized in Table 10.2. As we can see, the only processing we need to perform are two transformations of the output of the hand-coded system. One postprocessing step was needed in order to reduce the set of PoS tags. The hand-coded tagger has a set of tags that gives detailed information about each word. That is, in addition to giving the word category, it also gives information concerning the type, mode, time, person, gender, and number, whenever possible. Then, for the category verb there are around 600 possible tags. We decided to eliminate some of this information and retain only what we consider most relevant. For all categories we kept only information regarding their main PoS category, a detailed description of the reduced list can be found

Word	Hand-coded tag	BIO tag
La	O	O
Comisión	ORG	B
Nacional	ORG	I
del	ORG	I
Agua	ORG	I
alertó	O	O
el	O	O
desbordamiento	O	O
del	O	O
río	O	O
Cazones	LOC	B

Table 10.3: An example of how the tags assigned by the hand-coded tagger to the sentence are translated to the BIO scheme.

in Solorio (2005). The other postprocessing step is required to map the NE tags from the hand-coded tagger to the BIO tags; the hand-coded tagger does not assign BIO tags, instead it recognizes the NE in the documents and classifies them according to our predefined set. A very simple program analyzes these tags and translates them to the BIO scheme. Table 10.3 shows an example, where the hand-coded tagger tags are translated to the BIO scheme for the sentence *La Comisión Nacional del Agua alertó el desbordamiento del río Cazones* translated to English as *The National Commission of Water warned the flooding of the Cazones river*.

This NED algorithm is independent of the learning algorithm used to build the classifier. We can use the algorithm of our preference, provided it is well suited for this kind of learning task. In our evaluation we have used as learning algorithm Support Vector Machines (SVM) (Vapnik, 1995; Stitson et al., 1996). We give a brief description of this learning strategy on Subsection 10.1.4.

10.1.2 The features

The representation of instances of the learning concept is one of the most important considerations when designing a learning classification task. Each instance is represented by a vector of attribute values. For our problem, each word w_i is described by a vector of five attributes, $\langle a_1, a_2, \ldots, a_5 \rangle$, where a_1 to a_3 are what we call internal, or lexical, features: the word w_i, the orthographic information, and the position of the word in the sentence, respectively. Attributes a_4 and a_5 are the PoS tag and the BIO tag, both assigned by the hand--coded tagger. These two attributes are considered as external features, given that they are acquired from external sources, while the internal features are automatically gathered from the documents. In addition to this, we use for each word w_i the attributes of the two words surrounding w_i; that is, the attributes for words w_{i-2}, w_{i-1}, w_{i+1} and w_{i+2}. The final

feature vector for a given word w_i is the following:

$$w_i = [a1_{w_{i-2}}, ..., a5_{w_{i-2}}, a1_{w_{i-1}}, ..., a5_{w_{i-1}}, a1_{w_i}, ..., a5_{w_i}, a1_{w_{i+1}}, ..., a5_{w_{i+1}}, a1_{w_{i+2}}, ..., a5_{w_{i+2}}, c_i] \quad (10.1)$$

where c_i is the real class for word w_i.

To illustrate this, consider the sentence *El Ejército Mexicano puso en marcha el Plan DN-III*, the attribute vector for word *Mexicano* is the following:

$w_{Mexicano}$ = [*El*, 3, 1, DA, O,
Ejército, 3, 2, N C, B,
Mexicano, 3, 3, N C, I,
puso, 2, 4, V M, O,
en, 2, 5, SP, O,
I]

Within the orthographic information we consider 6 possible states of a word. A value of 1 in this attribute means that the letters in the word are all capitalized. A value of 2 corresponds to words where all letters are lower case. Value 3 is for words that have the initial letter capitalized. A 4 means the word has digits, 5 is for punctuation marks and 6 refers to marks representing the beginning and end of sentences.

Note that the attributes $a5_{w_i}$ and c_i will differ only when the base hand-coded tagger misclassifies a named entity, whereas by erroneously mixing the *B* and *I* tags; or by failing to recognize a word as an NE, in this case tags *B* and *I* will be misclassified by the hand-coded tagger as *O*. Intuitively, we may consider the incorrectly classified instances as noisy. However, we believe that by having the correct NE classes available in the training corpus, the learner will succeed in generalizing error patterns that will be used to assign the correct NE. If this assumption holds, that learning from other's mistakes is helpful, the learner will end up outperforming the initial hand-coded tagger.

The idea of the BIO labelling scheme, which uses three tags: *B*, *I* and *O*, for delimiting NE follows the work by Carreras et al. (2003a,b). The differences between their approach and the one proposed here lie in the representation of the learning task and the classification process. Concerning the attributes in the representation of problem instances, Carreras et al. include chunk tags of window words, chunk patterns of NE, trigger words, affixes and gazetteer features, none of them were used in our work. Their classification process is performed by selecting the highest confidence prediction from three binary AdaBoost classifiers, one for each class. In contrast, our classifier is a multi class adaptation of SVM.

10.1.3 Named Entity Classification

NE Classification is considered to be a more complex problem than NED. This may be due to the fact that orthographic features are less helpful for discriminating among NE classes.

Internal Features			External Features		Real class
Word	Caps	Position	POS tag	NEC tag	
El	3	1	DA	O	O
Ejército	3	2	NC	ORG	ORG
Mexicano	3	3	NC	ORG	ORG
puso	2	4	VM	O	O
en	2	5	SP	O	O
marcha	2	6	NC	O	O
el	2	7	DA	O	O
Plan	3	8	NC	O	MISC
DN-III	1	9	NC	ORG	MISC

Table 10.4: An example of the attributes used in the learning setting for NEC in Spanish for the sentence *El Ejército Mexicano puso en marcha el Plan DN-III* (The Mexican Army launched the DN-III plan).

The majority of NE seem to have very similar surface characteristics, and as a consequence envisioning good attributes for the task becomes more challenging. A common strategy to achieve good accuracy on NEC is to use linguistic resources such as word lists, dictionaries, gazetteers or trigger words. These resources are very helpful, and many of them are easily built because they are available in machine-readable format. However for most languages these resources have not been created yet, plus they can become obsolete quite rapidly. In this work, we try to use features without restricted availability, so we restrained the source of features to the information in the documents themselves.

The final set of features used in the NEC task includes all the attributes described in the NED task. Originally we thought it would be necessary to add other attributes for this task, as NEC poses a greater challenge to the learner. It turned out that the original set of features was good enough, and we will discuss this in more detail in the following section. Then, for a given word w we have as internal features the word itself (attribute a_1), orthographic information, (a_2), and the position in the sentence of word w (a_3). The external features also remained unchanged for the NEC task. We use the PoS tags and the NE tags from the hand-coded tagger. In Table 10.4 we present the features that describe each instance in this NEC task.

A summary of the NEC algorithm is presented in Table 10.5. Note, however, that concerning the output of the hand-coded tagger, the NE tags remain unchanged for this task.

10.1.4 The machine learning algorithm

The methods proposed in this work to solve the NER problem are used in combination with a machine learning algorithm. Note, however, that they are not designed to work with a specific learning algorithm. Rather, we can select the most appropriate algorithm considering the type of the learning task, the computing resources, namely CPU and me-

Let D_R be the set of labelled documents that will be used for training
Let D_T be the set of test documents
TRAINING
1. Label D_R with PoS and NE tags using the hand-coded tagger
2. Transform the PoS tags assigned by the hand-coded tagger to the compact set of tags
3. Build the training instances adding to the output of the hand-coded tagger the internal attributes
4. Give the learning algorithm the training instances and perform training

TESTING
1. Label D_T with PoS and NE tags using the hand-coded tagger
2. Transform the PoS tags assigned by the hand-coded tagger to the compact set of tags
3. Build the test set adding to the output of the hand-coded tagger the internal attributes
4. Let the trained classifier label the test instances

Table 10.5: The NEC algorithm

mory, and the amount of time we are willing to spend on the training and testing of the algorithm.

In this work we selected for our experiments Support Vector Machines as the learning strategy. However it is worth mentioning that due to computer resources constraints we did not carry out experiments with other learning schemes. For instance, ensemble methods are a promising alternative, as it is well known that they are a powerful learning strategy that usually outperforms the individual classifiers that make up the ensemble (Dietterich, 2000). Our main concern in this work is not to find the best learning algorithm for NER, but come up with a good representation of the learning problem that could be exploited in conjunction with any powerful learning algorithm. Thus, we selected the best algorithm that we could afford experimenting with and we consider the results reported throughout this document as a lower bound on classification measures. With a more powerful learning strategy, such as ensembles, and a larger training set, results could be improved considerably.

Support Vector Machines

Given that Support Vector Machines have proven to perform well over high dimensionality data, they have been successfully used in many natural language related applications, such as text classification (Joachims, 1999, 2002; Tong e Koller, 2001) and NER (Mitsumori et al., 2004). This technique uses geometrical properties in order to compute the hyperplane that best separates a set of training examples (Stitson et al., 1996). When the input space is not linearly separable SVM can map, by using a kernel function, the original input space to a high-dimensional feature space where the optimal separable hyperplane can be easily calculated. This is a very powerful feature, because it allows SVM to overcome the limitations of linear boundaries. They also can avoid the over-fitting problems of neural

Class	Instances
B	648
I	293
O	7,610

Table 10.6: Distribution of examples in the Portuguese corpus for the NED task.

networks as they are based on the structural risk minimization principle. The foundations of these machines were developed by Vapnik, and for more information about this algorithm we refer the reader to Vapnik (1995) and Schölkopf e Smola (2002).

In our work, the optimization algorithm used for training the support vector classifier is an implementation of Platt's sequential minimal optimization algorithm (Platt, 1999). The kernel function used for mapping the input space was a polynomial of exponent one. We used the implementation of SVM included in the WEKA environment (Witten e Frank, 1999).

10.2 Named Entity Recognition in Portuguese

We believe that the portability of our method is very important, even though we know that our method will not be completely language independent. There are important differences across languages that do not allow for a general NLP tool to be built, and the same applies to an NE tagger. We can aim at developing tools that will be useful for similar languages, which is a reasonable and practical goal, and is one of our goals in this research work. We are not expecting that our method will perform well on languages such as English or German, but we can expect it to be useful for other languages similar to those used in the current study, such as Italian, Portuguese or even Romanian. Considering that our method is based on an existing tagger for Spanish, it is reasonable to expect better results for Spanish than for any other language. However, if our method is capable of achieving good results for a different language, then we can claim it is a portable method, and it can be exploited to perform NER on several languages without any modifications.

In this Section, we evaluate the classification performance of our method on Portuguese. For this we used the training corpus provided by HAREM (see Chapter 1)[2]. This corpus contains documents of various literary genres. The corpus has 8,551 words with 648 NE. The following sections present our experimental results.

				Attributes								
	Hand-coded tagger			Internal			External			Internal & External		
Class	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*
B	60.0	68.8	0.641	**82.4**	85.8	0.841	75.9	81.0	0.784	82.1	**87.8**	**0.849**
I	64.5	73.3	0.686	80.1	76.8	0.784	73.8	70.3	0.720	**80.9**	**77.8**	**0.793**
O	97.2	95.5	0.964	98.7	98.5	0.986	98.1	97.7	0.979	**98.8**	98.4	0.986
Overall	73.9	79.2	0.763	87.0	87.0	0.870	82.6	83.0	0.827	87.2	88.0	0.876

Table 10.7: Experimental results for NED in Portuguese.

10.2.1 Results on NED

In this section we report our results of NED in Portuguese. We describe the distribution of instances over classes for the Portuguese corpus in Table 10.6. As the goal is to explore to what extent our method can be applied to similar languages, we did not make any particular changes to our system. The method is applied in the same way as it was applied previously to Spanish, results for Spanish can be found at Solorio (2005). Experimental results on NED are presented in Table 10.7. These results are averaged using a 10-fold cross validation[3]. We can observe that the hand-coded tagger achieved surprisingly high classification measures, it reached an F measure of 0.763. We believe that these results reveal that the two languages share some characteristics, among them the orthographic features: in Portuguese it is also conventional to write proper names with the first letter in uppercase. On the other hand, note also that the behavior of the two types of features differs greatly from that observed for Spanish. The internal features have better results than the external, for Spanish we observed that external features achieved better results than the internal ones. A plausible explanation to this is that, given that the hand-coded tagger misclassified more instances in the Portuguese case, then it is harder for the SVM, trained with the output of the hand-coded tagger, to learn the task in this somehow noisier setting. Nonetheless, SVM did improve the accuracy of the hand-coded tagger, and even more relevant for us, the combination of the two types of features yielded the best results. In this setting, our method is still the best option to achieve higher precision and recall on NED in Portuguese.

10.2.2 Results on NEC in Portuguese

We have shown that our proposed solution works well for Portuguese NED, now we need to evaluate how well this solution works for NEC in Portuguese. In this case the classifi-

[2] **Editors' note.** Note that the author does not apply in the chapter the measures used for HAREM elsewhere in this book, but rather defines her own, such as accuracy per word. Also she uses a small subset of the first golden collection, not the full golden collection.

[3] Since this is a classification task where we need to assign to every word one out of three possible classes, we measure per word accuracies.

Class	Instances
PESSOA	237
COISA	4
VALOR	68
ACONTECIMENTO	14
ORGANIZACAO	195
OBRA	56
LOCAL	187
TEMPO	112
ABSTRACCAO	55
VARIADO	13

Table 10.8: Distribution of examples in the Portuguese corpus for the NEC task.

cation task is more difficult due to several factors, among them, those we have discussed previously (Subsection 10.1.3). Another relevant factor is that the Portuguese corpus has a different set of NE classes than that of the hand-coded tagger. This Spanish tagger discriminates only among four different classes, namely PERSON, ORGANIZATION, LOCATION and MISCELLANEOUS. For the Portuguese set the classifier needs to assign NE tags from a set of 10 classes, these are PESSOA (person), COISA (object), VALOR (quantity), ACONTECIMENTO (event), ORGANIZACAO (organization), OBRA (artifact), LOCAL (location), TEMPO (date/time expression), ABSTRACCAO (abstraction) and VARIADO (miscellaneous). This will require the SVM to discover a function for mapping from the reduced set of classes to the larger set. Yet another complicating factor is the distribution of examples in the Portuguese set, which is shown in Table 10.8. We can observe that there are several classes for which we have very few examples, then there is little information for the classifier to learn these classes well. The following experimental results will show that these are not issues to be concerned of, the classifier does learn this type of target function. However, it is evident that more examples of the poorly represented classes can make a considerable difference in the classification performance.

Table 10.9 presents the results of NEC in Portuguese. Here again, we compared the four sets of results: the hand-coded tagger for Spanish, the internal features only, the external features only and the combination of both features. Similarly as in the NEC experiments we measured per word accuracies, but independently from the NED task[4]. The hand-coded tagger performed poorly, the overall F measure barely reaches the 0.10, and naturally it has an F measure of 0 on all the instances belonging to the classes not included on its set of classes. However, the hand-coded tagger has also an F measure of 0 for the VARIADO (miscellaneous) class even though for Spanish the hand-coded tagger was able to label corrECTLY some of the instances in this class.

[4] These results are optimistic since we are assuming a perfect classification on the NED task. On a real scenario the errors on NED classification would be carried on to the NEC task, degrading the performance of the NEC task.

				Attributes								
	Hand-coded tagger			Internal			External			Internal & External		
Category	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*
PESSOA	34.8	72.5	0.466	87.7	92.9	0.9023	47.7	74.0	0.58	83.3	89.6	0.864
COISA	0	0	0	0	0	0	50.0	25.0	0.333	0	0	0
VALOR	0	0	0	89.0	79.9	0.842	76.9	78.4	0.777	87.1	89.7	0.884
ACONTECIMENTO	0	0	0	1	76.2	0.864	83.3	9.5	0.169	1	38.1	0.550
ORGANIZACAO	41.4	38.4	0.393	83.4	88.9	0.849	46.5	48.0	0.472	79.7	85.5	0.825
OBRA	0	0	0	94.0	91.4	0.927	57.0	21.2	0.309	92.3	82.1	0.869
LOCAL	52.5	16.5	0.248	79.8	80.8	0.803	53.8	46.2	0.497	75.9	77.6	0.767
TEMPO	0	0	0	85.2	88.0	0.866	85.5	81.3	0.833	87.7	87.7	0.877
ABSTRACCAO	0	0	0	86.9	71.0	0.782	26.3	4.4	0.075	81.8	67.9	0.742
VARIADO	0	0	0	63.9	18.2	0.280	0	0	0	33.3	3.03	0.056
Overall	12.8	12.7	0.110	77.0	68.7	0.712	52.7	38.8	0.404	72.1	62.1	0.643

Table 10.9: NEC performance on the Portuguese set.

SVM trained with only the external features achieved impressive improvements, it is surprising to see how good this classifier performs, especially on the classes where the hand-coded tagger had errors of 100%. Consider for example, the case of the classes COISA and OBRA, the error reductions of these classes are quite large, external features achieved F measures of over 0.30, we were able to reduce the classification errors for more than 30%. We consider this an excellent achievement of this method.

On the other hand, internal features helped SVM to outperform the results of external ones, reaching F measures as high as 0.927 on the OBRA class. The set of results attained by the internal features are the best overall, leaving the SVM classifier combining both internal and external features as the second best. It is interesting to observe how, the internal features helped boost classification performance of the SVM trained with the external features, when both are combined. Regarding the performance of the SVM with internal features, we cannot assert the same, given that in this case the internal features performed better than the combination. It seems that, for Portuguese, combining both types of features was beneficial only in one direction.

As we mentioned at the beginning of this section, the hand-coded tagger classifies NE only into four categories. Considering this, it might be a little unfair to compare our method against the performance of the hand-coded tagger, as presented on Table 10.9. However, we believe that this comparison is important to show the flexibility of our method. We performed a different experiment in order to present a comparison with equal conditions for both taggers. In this experiment, we transformed the Portuguese corpus so that it fits the classification setting of the hand-coded tagger. First, we removed from the corpus instances belonging to classes VALOR and TEMPO. These classes were removed because the hand-coded tagger does not consider them as NE. Then, instances from classes

Class	Transformation	Description
PESSOA	PESSOA → PESSOA	remains unchanged
COISA	COISA → VARIADO	relabelled as VARIADO
VALOR	VALOR → ∅	eliminated from corpus
ACONTECIMENTO	ACONTECIMENTO → VARIADO	relabelled as VARIADO
ORGANIZACAO	ORGANIZACAO → ORGANIZACAO	remains unchanged
OBRA	OBRA → VARIADO	relabelled as VARIADO
LOCAL	LOCAL → LOCAL	remains unchanged
TEMPO	TEMPO → ∅	eliminated from corpus
ABSTRACCAO	ABSTRACCAO → VARIADO	relabelled as VARIADO
VARIADO	VARIADO → VARIADO	remains unchanged

Table 10.10: Modifications of the Portuguese corpus to fit the classification setting of the hand-coded tagger.

				Attributes								
	Hand-coded tagger			Internal			External			Internal & External		
Category	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*	*P* (%)	*R* (%)	*F*
PESSOA	35.6	72.3	0.477	86.7	91.0	0.888	48.9	72.3	0.583	87.3	91.0	0.891
ORGANIZACAO	41.8	37.8	0.397	84.4	89.4	0.868	47.3	44.5	0.459	82.2	87.0	0.845
LOCAL	68.0	17.2	0.274	85.4	82.7	0.840	56.3	51.2	0.536	79.9	79.9	0.799
VARIADO	0	0	0	90.0	77.3	0.832	31.7	12.6	0.180	83.6	70.7	0.766
Overall	36.3	31.8	0.287	86.6	85.1	0.857	46.0	45.1	0.440	83.3	82.1	0.825

Table 10.11: NEC performance on the modified Portuguese set.

COISA,ACONTECIMENTO,OBRA and ABSTRACCAO were relabelled as VARIADO, which is equivalent to class MISC. The remaining instances, belonging to classes PESSOA, ORGANIZACAO and LOCAL, were left unchanged.

In Table 10.10 we summarize the transformation process. Classification results of this experiment are presented in Table 10.11. These results are similar to those on Table 10.9. The hand-coded tagger had the lowest classification measures, reaching an F measure of 0.287; despite this poor behavior of the hand-coded tagger, we were able to improve NEC performance by a large margin, a combination of features yielded an F measure of 0.825. SVM trained on internal features attained the best results overall, although for class PESSOA the combination of internal and external features outperformed SVM trained only with internal features.

10.3 Final remarks

We are pleased to see the outcome of these experiments. Although the test set is small, we still consider these results very promising. We posed this problem as a machine learning task, then we trained a learning algorithm with the data available. Thus, a reasonable ex-

pectation of having more data available is that of expecting the classifier to learn better the target function, since for a learning algorithm the more data the better they will perform, provided the new data is not noisy.

We were able to reach excellent results on both NE tasks showing that our method can be applied to the task of NER on Portuguese and achieve high accuracies. We succeeded on our goal of increasing the coverage of a hand-coded named entity tagger in a different domain. The hand-coded system was developed for Spanish, then its coverage on Portuguese texts was very low. Nevertheless, by using our representation of the learning task, the coverage was increased tremendously, in some cases error reductions were as high as 80%; see classification measures for classes `VALOR`, `TEMPO` and `ABSTRACCAO` on Table 10.9. It is not surprising that internal features deliver better results in the majority of the cases, however the combination of features deliver competitive results. The important contribution from this work is that we can have the same method, using exactly the same representation, to perform NER on Spanish and Portuguese, without any manual tuning.

Our system entered the HAREM evaluation contest and it ranked #12 from 22 runs on the global results, and as high as #8 on the literary genre for NED.

Our design of the learning task has shown that it is possible to build good NE taggers without the need of complex and language-dependent features that are commonly used for NER. The method is flexible that we do not even need the hand-coded tagger: the internal features proved to be sufficient by themselves, leaving the use of a hand-coded tagger as optional.

An important characteristic of our method is its flexibility. We showed results proving that the method can be applied to a language other than Spanish with excellent results. Additionally, the method performed equally well on simulated speech transcripts, thus it is very flexible. Moreover, the method is flexible also regarding the classification setting of NE. Recall that the hand-coded tagger can only classify NE into a set of four categories. However, as the Portuguese data set has 10 different categories, it was unclear, at first, if this wider classification represented a problem for our method. This turned out not to be a problem, as it achieved impressively high accuracies. We can conclude that the method is not restricted in this respect, it can be applied to different categorizations of NE, regardless of the ones determined by the hand-coded tagger.

Acknowledgements

We would like to thank the different reviewers of this chapter for their thoughtful comments and suggestions. We would also like to thank Nuno Cardoso and Diana Santos for their great job on this book.

This work was done while the first author was at the National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico.

Capítulo 11

Tackling HAREM's Portuguese Named Entity Recognition task with Spanish resources

Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Rafael Muñoz e Andrés Montoyo

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 11, p. 137–144, 2007.

This chapter presents our participation in the HAREM evaluation contest. This is a challenge regarding the identification and classification of named entities in Portuguese. Our NER system, initially designed for Spanish, combines several classifiers in order to resolve the classification of the entities. Besides, a rule-based module has been used to deal with entity types easily recognized by applying knowledge resources such as regular expressions (e.g. `TEMPO:DATA`).

The rest of this chapter is organized as follows. The next section introduces our system and the modules it is made of. The carried out experiments are explained and discussed in Section 11.2. Finally, Section 11.3 outlines our conclusions.

11.1 System Description

For our participation in HAREM (Santos et al., 2006), we have used the architecture of our system NERUA (Ferrández et al., 2005; Kozareva et al., 2007). This is a NER system that was developed combining three classifiers by means of a voting strategy. This system carries out the recognition of entities in two phases: detection[1] of entities and classification of the detected entities. The three classifiers integrated in NERUA use the following algorithms: Hidden Markov Models (HMM) (Schröer, 2002), Maximum Entrophy (ME) (Suárez e Palomar, 2002) and Memory Based Learning (TiMBL) (Daelemans et al., 2003). The outputs of the classifiers are combined using a weighted voting strategy which consists of assigning different weights to the models corresponding to the correct class they determine. An overview of our system is depicted in Figure 11.1.

The first stage starts with the feature extraction for the entity detection (FEM). The text, enriched with feature values corresponding to each word, is passed to the HMM and TiMBL classifiers. Due to its high processing time, ME was not used in the detection phase, but its absence is not crucial, as entity delimitation is considered to be easier than entity classification. Classifiers' outputs are then combined through a voting scheme.

The second stage has as starting point the text with the identified named entities. Therefore, only entities that have been previously detected are going to be classified and features for the classification of these entities will be extracted. The performance of the second stage is obviously influenced by the results of the first one. The classifiers involved at this stage are: HMM, TiMBL and ME. Each one of them uses labeled training examples in order to predict the class of the unseen example. The final outcome is the result of the voting scheme. This second stage yields all the identified NE together with the class each entity belongs to.

Our voting approach regarding both the identification and the classification phases has been already evaluated in Ferrández et al. (2005) and Kozareva et al. (2007). TiMBL is the classifier that obtains the best results for identification, while ME is the one reaching the

[1] **Editors' note.** As in the previous chapter, the authors use *detection* to mean what we dubbed *identification* in HAREM.

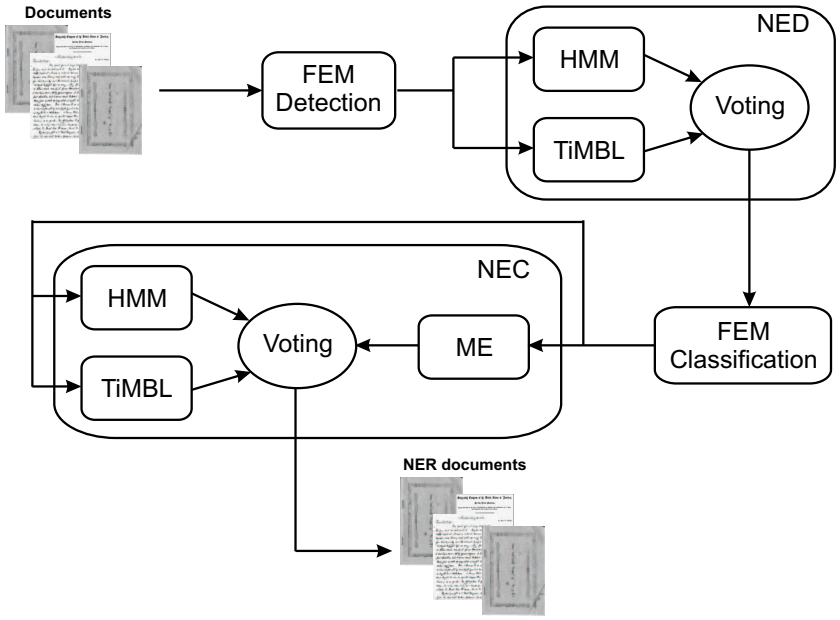


Figure 11.1: The NERUA architecture.

best score for the classification. The voting strategy meaningfully increases the final score above the results provided separately by the algorithms.

Due to the small size of tagged corpora available for Portuguese and the facts that our NER system was initially designed for Spanish and Spanish and Portuguese are close-related languages, we decided to merge the Spanish and Portuguese training corpora in order to train our system. The Spanish training corpus we used was provided for the CoNLL-2002 shared task (Sang, 2002). As in CoNLL-2002 only four kind of entities were considered (`PERSON`, `ORGANIZATION`, `LOCATION` and `MISCELLANEOUS`) we have focused in the following HAREM correspondent types: `PESSOA`, `ORGANIZACAO` and `LOCAL`.

By studying the entity taxonomy of HAREM (Santos et al., 2006), we saw that for some of the NE types, a knowledge-based approach could perform better. Entities such as `TEMPO:DATA` or `VALOR:QUANTIDADE`, have regular and a priori known structure, therefore they can be tackled more efficiently by using regular expressions and dictionaries.

Therefore, apart from the machine-learning system, we used a knowledge-based one which classifies the following entity types: `LOCAL:VIRTUAL`, `TEMPO:DATA`, `TEMPO:CICLICO`, `TEMPO:HORA`, `VALOR:MOEDA` and `VALOR:QUANTIDADE`. The system we used is called DRAMNERI (Toral, 2005). This system is a NER application belonging to the knowledge paradigm and adaptable to different domains and languages. In this research, this system has been adapted to recognize the aforementioned types of entities by hand-coding the correspondent dictionaries and rules.

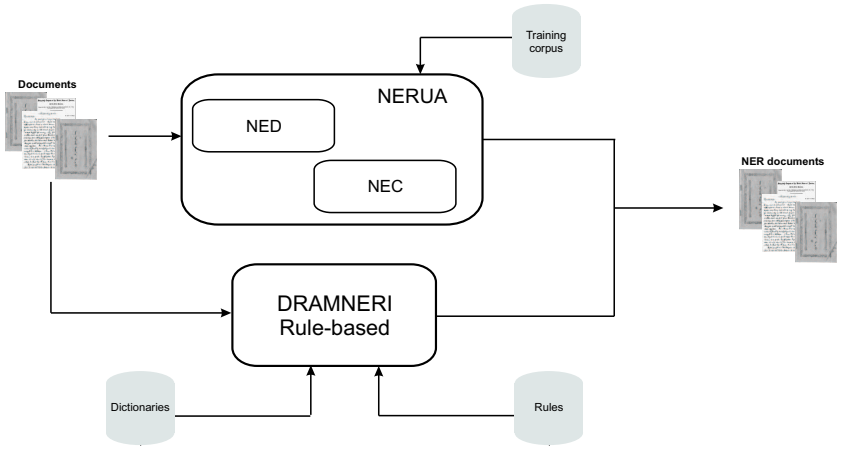


Figure 11.2: System description.

For this purposes, DRAMNERI uses 32 rules (4 for `LOCAL:VIRTUAL`, 21 for `TEMPO:DATA`, 1 for `TEMPO:CICLICO`, 2 for `TEMPO:HORA`, 3 for `VALOR:MOEDA` and 1 for `VALOR:QUANTIDADE`). The applied dictionaries contain 80 tokens. These resources were adapted from the Spanish resources. The adaptation consisted only of translating the language dependent strings in the dictionaries and in the rules (e.g. January (*Enero* to *Janeiro*)). In other words, the rules' structure was not modified.

Figure 11.2 depicts our system using both the machine learning and rule-based NER sub-systems. Both NER sub-systems are applied to the input-text in a parallel way. Afterwards, a postprocessing module receives both tagged texts and composes a final tagged text. If a snippet is tagged as an entity by both modules then the rule-based one is given precedence, i.e., the entity tagged by this latter NER system would be the one preserved[2].

11.1.1 Feature sets

To improve the performance of the classifiers, a large number of features were extracted from the training corpus to get a pool of potentially useful features (this procedure is shown in detail in Ferrández et al. (2006)). Many of these features are acquired from the best performing NER systems such as Carreras et al. (2002) and Florian et al. (2003). We have divided our features into several groups: orthographic (about the orthography of the word), contextual (about the context of the word), morphological (about morphological characteristics), statistic (about statistical characteristics) and handcrafted-list (test whether or not the word is contained in some handcrafted list of general entities obtained from several web pages). Below, we describe the features in detail:

[2] This case rarely happens, since the systems were designed to classify different kind of entities.

- **Orthographic**
  - **a**: anchor word (e.g. the word to be classified)
  - **cap**: capitalization of the word and context
  - **allcap**: whole word and context are in upper case
  - **lower**: whole word and context are in lower case
  - **internal**: word and context have internal upper case letters
  - **digits**: word and context are only made up of digits
  - **contdig**: word and context contain digits
  - **ispunct**: word and context are punctuation marks
  - **contpunct**: word and context contain punctuation marks
  - **hyphen**: word and context are hyphenated
  - **initial**: word and context are initials (e.g. B.O.E. or D.O.G.V.)
  - **url**: word and context represent an URL
  - **prefix**: the first three and four characters of the word and context
  - **suffix**: the last three and four characters of the word and context
  - **middle**: half substring of the word and context
  - **firstword**: first word of the whole entity
  - **secondword**: second word of the whole entity
  - **clx**: words within the entity are upper-cased (c), lower-cased (l) or made up of other symbols (x), e.g. *Charles de Gaulle*: clc
- **Contextual**
  - **cntxt**: word context at position ± 1, ± 2, ± 3
  - **verbword**: the nearest verb that comes with the entity
- **Morphological**
  - **postag**: PoS tag of the word and context
  - **lemma**: lemma of the word and context
  - **stem**: stem of the word and context
- **Metrical**
  - **length**: number of characters of the word and context
  - **firstpos**: word is the first word of the sentence

- **Handcrafted list**
  - **stopword**: word and context are stop-words
  - **dict**: word and context are in handcrafted dictionaries of entities (locations, persons and organizations)
  - **trigg**: word and context are in handcrafted dictionaries of trigger words
  - **connec**: context is contained in a dictionary of connectives
  - **WNword**: the WordNet semantic prime of the word from the Spanish WordNet

Since in HAREM we did not have enough training resources for the target language (Portuguese), we have considered only sets containing features that do not depend on a language-specific tool (called IDL sets) (Ferrández et al., 2006). In order to select the most meaningful features, we have followed a bottom-up strategy. This strategy iteratively adds one feature at a time and checks the effect of this feature in the results according to the information gain of this feature. The feature sets used for HAREM were:

- IDL sets for the detection phase
  - IDL1d: a, cntxt, cap, allcap[3], firstpos, url[3], ispunct[3], contpunct[3], digits[3], contdig[3], internal[3], ishyphen[3], lower[3].
  - IDL2d: IDL1 + prefix[3], suffix[3], middle[3].
- IDL sets for the classification phase
  - IDL1c: a, cntxt, firstpos, firstword, secondword, clx, url[3], ispunct[3], cont-punct[3], digits[3], contdig[3], internal[3], ishyphen[3], lower[3].

11.2 Experiments and discussion

This section presents the experiments carried out for our participation in HAREM. We show the obtained results and briefly discuss them. The aim of our study is to evaluate the recognition of entities with resources for a close-related language.

We have carried out three runs: one for the identification (*r_detection*) and the remaining two for the semantic classification. Regarding the two classification runs, one (*r_clas_total*) deals with all the entity types that we have considered while the other one (*r_clas_partial*) treats the ones that we thought the system could obtain better results (all categories but OBRA and ABSTRACCAO).

Table 11.2 shows the results obtained for the identification phase in HAREM. Table 11.2 presents the results for the semantic classification task according to CSC (combined) measure (Santos et al., 2006).

[3] only the word (not the context)

Category	Run	Total scenario			Selective scenario		
		Precision	Recall	F measure	Precision	Recall	F measure
all	r_detection	56.93%	64.39%	0.6043	-	-	-
	r_clas_partial	59.43%	64.39%	0.6181	52.25%	65.43%	0.5810
	r_clas_total	57.19%	63.51%	0.6019	-	-	-

Table 11.1: Results of the identification task, for the total and selective scenarios.

Category	Run	Absolute scenario			Relative scenario		
		Precision	Recall	F measure	Precision	Recall	F measure
PESSOA	r_clas_partial	26.93%	16.44%	0.2042	84.37%	49.86%	0.6268
	r_clas_total	19.59%	26.67%	0.2259	79.15%	79.62%	0.7938
ORGANIZACAO	r_clas_partial	27.35%	21.44%	0.2404	76.63%	46.36%	0.5777
	r_clas_total	25.57%	27.61%	0.2655	65.56%	68.44%	0.6697
LOCAL	r_clas_partial	40.13%	19.27%	0.2603	89.72%	52.37%	0.6614
	r_clas_total	32.90%	29.78%	0.3126	82.38%	83.50%	0.8294
TEMPO	r_clas_partial	75.26%	65.36%	0.6996	91.58%	91.88%	0.9173
	r_clas_total	53.58%	66.57%	0.5937	91.22%	91.80%	0.9151
VALOR	r_clas_partial	35.23%	71.12%	0.4712	77.42%	79.22%	0.7831
	r_clas_total	34.72%	72.26%	0.4690	77.61%	79.39%	0.7849
ABSTRACCAO	r_clas_total	15.14%	6.72%	0.0931	58.52%	59.66%	0.5908
OBRA	r_clas_total	6.62%	5.36%	0.0592	60.74%	52.98%	0.5660
VARIADO	r_clas_partial	1.28%	21.96%	0.0241	85.64%	85.64%	0.8564

Table 11.2: Results of the semantic classification task according to the CSC (combined) measure, for the selective scenario (runs *r_clas_partial*) and for the total scenario (*r_clas_total*).

Regarding identification (see Table 11.2), even if we have not made an extensive use of Portuguese specific resources, we have reached the 5th best score in F measure. Considering the small effort realised in order to adapt our system to Portuguese, the overall results are promising. It should be noted as well that the result for the selective scenario is worst (see r_class_partial) than that for the total scenario. This is due to the fact that for the selective scenario the categories ABSTRACCAO and OBRA are not considered but they might be detected by our system although afterwards they will not be classified (this is why the results for the selective scenario in the semantic classification (see Table 11.2) are better than for the total scenario).

As to the entity classification (see Table 11.2), our system obtains quite high scores for TEMPO (F measure of 0.9173) and LOCAL (F measure of 0.8294). This is due to the fact that, in the first case, temporal expressions can be appropriately tackled with regular expressions and, in the second case, local entities do not depend that much on the specific language.

11.3 Conclusions

In this paper we have presented our participation in HAREM. In order to recognize named entities in Portuguese, we decided to apply our previously developed NER system for Spanish. We have merged our already available Spanish corpus with the Portuguese one because of the lack of sufficient training data. The feature sets developed for Spanish were directly ported to detect and classify Portuguese NE. This was possible due to the proximity and the common characteristics of the two languages. Apart from this, we treated some entities (VALOR, TEMPO, LOCAL:VIRTUAL) with a knowledge-based approach.

NERUA came on fifth position in the NE identification task in the first HAREM. It obtained better results in the identification task compared to the classification one. This is due to the lack of annotated resources for Portuguese and the fact that we have focused on the recognition of a subset of entities. In this contest, we showed that our NER system, initially designed and developed for Spanish, was adapted with little effort to Portuguese and achieved promising results.

Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01.

Capítulo 12

Functional aspects on Portuguese NER

Eckhard Bick

This chapter is republished, with kind permission from Springer-Verlag, from Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 2006. Proceedings*, LNAI series, Vol. 3960, pp. 80-89. ISBN-10: 3-540-34045-9.

Therefore, we restrained from doing any changes to the original text, even notational conventions, adding instead editors' notes commenting on possible mismatches.

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM, *a primeira avaliação conjunta na área*, Capítulo 12, p. 145–155, 2007.

The PALAVRAS-NER parser is a progressive-level Constraint Grammar (CG) system, treating Named Entity Recognition (NER) as an integrated task of grammatical tagging. The original version, presented at the PROPOR 2003 (Bick, 2003) and also used for Linguateca's avalia-SREC task 2003, implemented a basic tag set of 6 NER categories (person, organisation, place, event, semantic products and objects) with about 20 subcategories, following the guidelines of a joint Scandinavian NER project (Nomen Nescio (Johannessen et al., 2005)). Category tag candidates were added at three levels, and subsequently disambiguated by CG-rules:

a) known lexical entries and gazeteer lists (about 17.000 entries)

b) pattern-based name type prediction (morphological module)

c) context-based name type inference for unknown words

Since PALAVRAS originally was conceived primarily as a syntactic parser (Bick, 2000), it fuses fixed expressions with non-compositional syntactic-semantic function into multi--word expressions (MWEs), creating complex tokens and in the process making life easier for the token-based syntactic CG-rules as well as avoiding arbitrary descriptive decisions as to the internal structure of such MWE[1]. Names, too, are treated as MWEs, and semantic NER-classes are assigned to the whole, not the parts.

12.1 Recognizing MWE name chains

Identification of names, as a sequence of atomic tokens, was a separate task in the HAREM joined NER evaluation (www.linguateca.pt), and the PALAVRAS-system performed best, with an F-Score of 80.61%, in both the selective and total measures. Single-token names, with the exception of sentence-initial position, are clearly marked by upper case - therefore, since multi-token names can't be identified without chaining them into MWEs first, and since very few other (non-NE) cases involve productive MWE-chaining, the NE identification task is to a large degree identical to an MWE-recognition task[2]. The 2003 PALAVRAS--NER system (in this text, PAL-1), taking a more static approach, tried to fix MWE names *before* running the system's grammars – either by simple lexicon-lookup or by pattern-recognition in the preprocessor – and the only allowed post-grammar token alteration was fusion of adjacent name chains. This technique was replaced by a more dynamic, grammar based tokenisation approach in the new, 2005 system (henceforth, PAL-2), used for HAREM. Here, preprocessor-generated name candidate MWEs that cannot be verified in

[1] For corpus-users with a blank-space based token definition, MWEs can be unfolded and assigned an internal analysis by an add-on filter-program.

[2] Strictly speaking, the HAREM annotation and metrics did not employ MWEs per se, but rather XML-tags marking the start end end of name expressions. These XML tags were automatically added to PALAVRAS output before evaluation, at the same time turning semantic category tags into XML attributes.

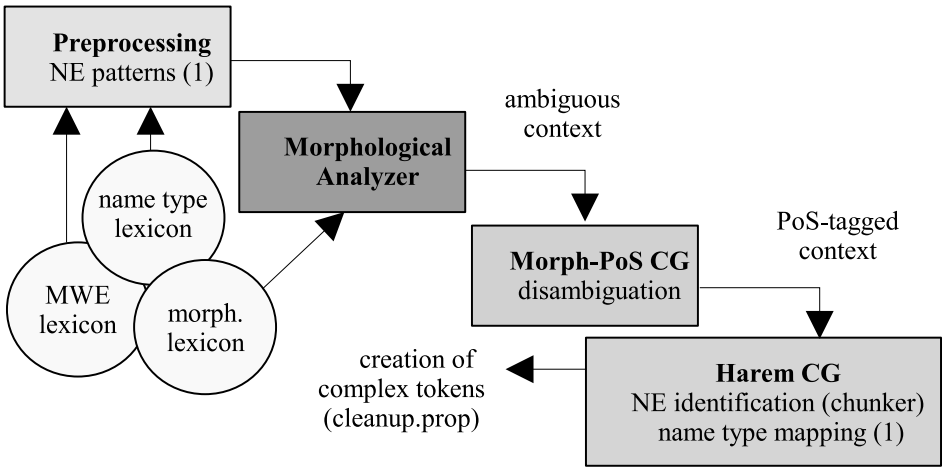


Figure 12.1: Name chain identification modules

the lexicon as either known names or non-name polylexicals, are fed to the morphological analyser not as a whole, but in individual token parts, with < and > tags indicating start and stop of name MWE candidates. Thus, parts of unknown name candidates will be individually tagged for word class, inflexion and - not least - semantic prototype class. In addition, each part is tagged either @prop1 (leftmost part) or @prop2 (middle and rightmost parts). This technique has two obvious advantages over the old approach:

1. It allows the morphological disambiguation grammar to establish the gender and number of names from their constituents, as well as internal morphological features, name-internal pp-constructions etc.

2. A specialized, newly-written name grammar can *change* the very composition of a name MWE, by removing, adding or replacing @prop1 start and @prop2 continuation tags.

For instance, the grammar can decide contextually whether sentence initial upper case is to be treated as a part of a name or not. Thus, certain word classes (prepositions, adverbs, conjunctions, finite verbs) can be recognized and tagged as no-name even with another upper case word to the right. Though a simple preprocessor might have sufficed to check for the closed classes, this is problematic due to ambiguity, and certainly not true of finite verbs, which are both open-class and often ambiguous with nouns, so the task has to be done after morphological analysis and disambiguation (illustration 12.1).

The name-chunker part of the Harem CG can progressively increase the length of a half-recognized chunk in a grammatically founded and context-sensitive way, for instance by adding conjuncts (e.g. the last two tokens in ... *Doenças Infecciosas e Parasitárias*, a1) or

PPs (e.g. the last part of *a Câmara Municipal de Leiria*, a2). Since the parts of name chains at this stage are "perspicuous" as nouns or other word classes, valency potential may be exploited directly (a3). In the rules below, the MAP operator adds information (tags) to a TARGET for a given context (1 meaning "one word to the right", -2 "two words to the left" etc.). BARRIER conditions can block a context if the barrier tag is found between the target and the context tag in question, while LINK conditions add secondary context conditions to an already instantiated context.

```
(a1)
MAP (@prop2) TARGET (KC) (-1 <prop2> LINK 0 ATTR) (1 <*> LINK 0 ATTR)
MAP (@prop2) TARGET <*> (0 ATTR) (-1 KC) (-2 <prop2> LINK 0 ATTR) ;
where <*> = upper case, KC = coordinator, ATTR = attribute

(a2)
MAP (@x @prop2) TARGET PRP-DE (*-1 N-INST BARRIER NON-ATTR LINK
0 <prop1>) (1PROP LINK 0 <civ> OR <top>)
MAP (@x @prop2) TARGET PROP (0 <civ> OR <top>) (-1 PRP-DE) (*-2 N-INST
BARRIER NON-ATTR LINK 0 <prop1>); where PROP = (atomic) proper noun, N-
INST = nouns with a semantic-prototype tag of institution, <civ> = known
civitas names, <top> = known place names, <prop1> = preprocessor-proposed
start of name chunk.

(a3)
MAP (@prop1) TARGET <*> (0 <+a>) (1 PRP-A) (NOT -1 >>>) ; where <+a> =
noun's or participle's binding potential for the preposition a, >>> =
sentence start
```

Not all name-part mapping rules are unambiguous - (a2), for instance, includes @x, meaning "wrongly assumed name part", along with @prop2, meaning "second part of name". Ultimately, a set of REMOVE and SELECT rules decides for each name part candidate if it is valid in context and if it is a first or later part of the chain. For instance, definite articles or the preposition *de* cannot be part of a name chain, if the token immediately to the right is not a second part candidate, or has been stripped of its name tag by another, earlier, rule:

```
REMOVE (@prop2) (0 <artd> OR PRP-DE LINK 0 @y) (NOT 1 @prop2)
```

The result, an unambiguous tag (@prop1=first part, @prop2=later part, @x=ex-name, @y=confirmed no-name) is implemented by a filter program, *cleanup.prop*, such that later programs and grammars will see only ready-made complex name tokens.

12.2 Semantic typing of name tokens: Lexematic versus functional NE categories

The next task, after identifying the name chain tokens, was to assign them a semantic category and subtype. The original PAL-1 did subdivide the 6 *Nomen Nescio* supercategories into subcategories, but recognized only about 17 partly experimental categories, while the new PAL-2 had to accommodate for HAREM's 9 categories and 41 subcategories[3]. This meant more than doubling the category inventory, and category matching was in many cases complicated by the fact that matches were not one-to-many, but many-to--many. This difference was not, however, the most important one. Far more crucial, both linguistically (i.e. in terms of descriptive meaning) and application ally (i.e. in terms of parsing grammars), was the treatment of metonymy. For many name types, metonymy is a systematic, productive and frequent phenomenon – thus, author names may be used to represent their works, city names may denote soccer clubs and a country name may be substituted for its government. Here, PAL-1 subscribed to a lexeme based definition of name categories, while HAREM used a function-based category definition. In the former tradition, a given name would have one, unchanging lexematic category, while in the latter it would change category according to context. Thus, the name of a country would always be `<civ>` (civitas) in PAL-1, a hybrid category of place and organisation, allowing, for instance, both +HUM subject-hood, and BE-IN-LOC-adverbiality. According to the HAREM guidelines, however, hybrid categories were not allowed[4] , and simply turning `<civ>` into `<top>` (place) would result in a considerable error rate in those cases, where the country--name *functions* as an organisation or a humanoid group, i.e. where it announces, suffers or goes to war. Likewise, institutions `<inst>` can be seen as both places and organisations, while the erstwhile `<media>` category implies a function-split between a newspaper being read (semantic product), burned (object) or sued in court (company). On the other hand, HAREM also introduced some distinctions that *were* lexematic rather than functional, for instance the split between the (money-making) *company* subtype and the non-profit institution subtype of the organisation category.

In order to handle the lexeme-function difference, PAL-2 had not only to increase its category inventory, but treat lexicon-, morphology- and pattern-derived categories as "potentialities" to a much higher degree than PAL-1 had done. 5 levels can be distinguished for such lexicon-dependence or -independence of name tagging:

1. lexicon-entered names that have a reasonably unambiguous name category (e.g. Christian names, to a lesser degree surnames, which can denote styles or an artist's

[3] **Editors' note.** There are 10 categories in HAREM; the author is here ignoring the `VARIADO` category.

[4] **Editors' note.** A little precision is in order here: Since no system at the First HAREM reported that it would use the OR notation (in this case, `LOCAL | ORGANIZACAO`) in its output, "hybrid" categories were only used in the golden collection. In fact, the PALAVRAS-NER system could have used them, but then it would still not fare well in the cases where the golden resource had only `LOCAL` or `ORGANIZACAO`, which we believe to be Eckhard Bick's main message in this context.

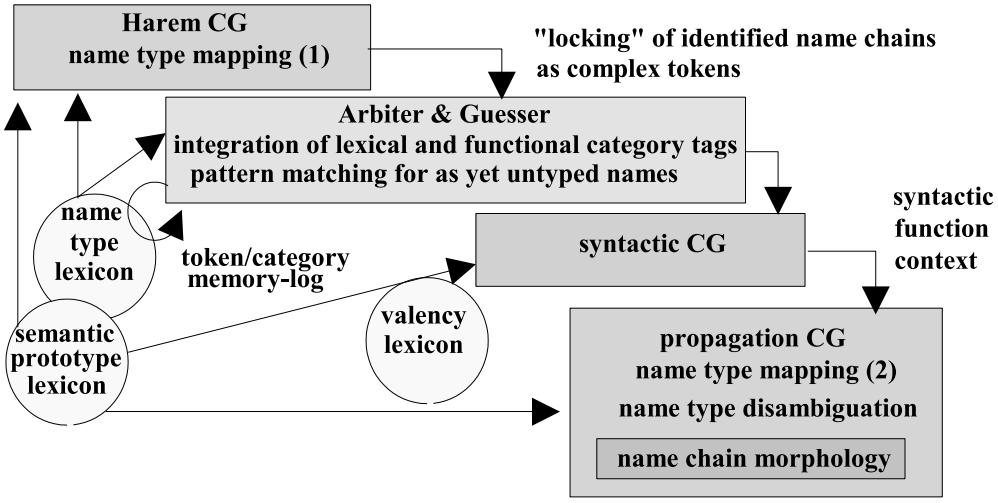


Figure 12.2: Name typing modules

collected work)

2. lexicon-entered names with semantically hybrid categories (<civ>, <media>, <inst>) or with systematic metaphoring (<brand> as <object>)

3. pattern/morphology-matched names of type (1)

4. pattern/morphology-matched names of type (2)

5. names recognized as such (upper case, name chaining), but without a lexicon entry or a category-specific pattern/morphology-match

Even in the PAL-1 evaluation (Bick, 2003), where hybrid categories did not have to be resolved and where only few, strong rules were allowed to override lexicon- or gazeteer--supported name-readings (1. and 2.), this group had an error rate of 5%, indicating that for many names, ambiguity is not merely functional, but already hard-wired in the lexicon (e.g. *Washington* as person or place name). In PAL-2, lexicon-derived categories were treated as contextual indications only, and the names carrying them were submitted to the same rule set as "unknown" names (3. - 5.), opening up for considerably more ambiguity and a correspondingly higher error risk.

Illustration 12.2 shows the distributed nature of PAL-2 and the interaction of its different name typing modules. An essential cut, the "locking"of identified name chains into complex tokens, is made between the (new) Harem CG on the one hand and the (modified) syntactic module and propagation CG on the other. While the former (*micromapping*) works on minimal tokens (name-part words) and can exploit their PoS, semantics and

morphology, this is not any longer possible for the latter, which is geared for syntactic clarity and therefore works on whole name chunks, and uses syntactic function and structure to "propagate" information from the rest of the sentence onto nouns (*macromapping*).

12.2.1 Micromapping: Name type rules based on name parts and patterns

Many of the micromapper's rules map chunking information at the same time as classifier tags, like in the following rule which types known towns or countries (<civ>) or typical *noun parts* (N-CIVITAS) of unknown towns or countries as "administrative", if they fill the subject slot of a human-agent or experiencer verb (V-HUM).

```
MAP (@admin @prop1) TARGET <*> (0 <civ> OR N-CIVITAS) (*1 V-NONAD
BARRIER CLB LINK 0 V-HUM) (NOT 0 <prop2>)
```

It is the first part of a complex name (@prop1) that will carry the classifier tag (@admin), and both tag types may be mapped ambiguously for later rule based disambiguation. Once output from the micromapper CG has been "frozen"into name chunks, the Arbiter module checks the result against lexical data and morphological patterns, adding pattern based classifier tags where no category has been mapped, or where tags are marked as unsafe (e.g. <hum?>) by the pre-CG inflexion and derivation analyzer. The Arbiter is the only part of the system that has a text-level memory - logging identified names and their types to resolve the classification of name abbreviations and the gender of person names. Thus, on a small scale, entity disambiguation is used for NE typing as suggested by Blume (2005).

The Pal-1 based morphological analyzer only treats numbers as NE material if they are part of a larger NE, e.g. time and place units, not when occurring as mere quantifiers, as in the HAREM categories[5] of QUANTIDADE, CLASSIFICACAO and MOEDA. In PAL-2, it is the Arbiter's pattern-matching module, not the "character-blind" CG, who has to recognize such number expressions as names, as well as pre-classify them for later treatment in the CG macromapper.

12.2.2 Macromapping: Name type rules based on syntactic propagation

Macromapping is an adapted PAL-1 module that adds name type tags to already-identified name chains by using a number of syntactic "propagation"techniques (described in Bick (2003)), exploiting semantic information elsewhere in the sentence:

1. *Cross-nominal prototype transfer*: Postnominal or predicative names (NE @N<, PRP @N< + NE @P<, @SC, @OC) inherit the semantic type through of their noun-head

[5] **Editors' note.** We used the denomination "categories" for what the author refers as "major categories" elsewhere in this text, and "types" for "subcategories". So, in this case, the author is referring to HAREM types, and not categories.

2. Coordination based type inference: Types are propagated between conjuncts, if one has been determined, the other(s) inherit the same type.

3. Selection restrictions: Types are selected according to semantic argument restrictions, i.e. +HUM for (name) subjects of speech- and cognitive verbs, +TIME is selected after temporal prepositions etc.

In Constraint Grammar terms, macromapping is as much a mapping technique as a disambiguation technique, as becomes particularly clear from method (3), where many rules discard whole sets of name type categories by targeting an atomic semantic feature (+HUM or +TIME) shared by the whole group.

12.3 Evaluation

The complete HAREM evaluation computed a number of other metrics, such as text type dependent performance. PAL-2 came out on top for both European and Brazilian Portuguese, but in spite of its Brazilian-optimized lexicon and syntactic parser, it achieved a higher F-Score for the latter (60.3% vs. 54.7%), possibly reflecting sociolinguistic factors like the higher variation of person names in a traditional immigration country like Brazil, its Tupi-based place names etc. all of which hamper regular pattern/morphology-based name type recognition[6]. HAREM also had separate selective scores, where systems were allowed to compete only for certain categories and skip others. However, since PAL-2 competed globally in all areas, selective scores equaled total scores.

Another HAREM measure not presented in the overview table were relative performance, defined as category recognition measure separately for only those NEs that were correctly identified. Since this was not done by presenting systems with a ready-chunked ("gold-chunk-") corpus, but by measuring only against NEs correctly recognized by the system itself, PAL-2 had the relative disadvantage of being the best identifier and thus having to cope also with a larger proportion of difficult names than other systems, resulting in suboptimal rank performance.

For a direct performance comparison between PAL-1 and PAL-2, only the per-category scores are relevant, since even if subcategory scores had been available for PAL-1, score differences might simply reflect the difference in type set size. Even so, however, scores neither matched nor differed systematically. Of the major categories, *person* and *place* scored better in PAL-2/HAREM than what was published for the lexeme-based approach in PAL-1 (Bick 2003), while *organisation* and *event* had lower scores. Interestingly, the major categories (person, organisation, place) even *ranked* differently, with *person* higher (lowest in PAL-1) and organisation lowest (second in PAL-1). The reason for this may reside in the

[6] Alas, since all HAREM participants but the winner were anonymous, and different code names were used for the Brazilian and Lusitan evaluation, this pattern could not at the time of writing be verified as either general or system-specific.

PALAVRAS Subtype	Category (incidence)	HAREM Subtype	F-Score (precision - recall) cat total	cat/types total	identification
hum official member grupoind groupofficial grouporg	**hum** PESSOA 20.5%	INDIVIDUAL CARGO MEMBRO GRUPOIND GRUPOCARGO GRUPOMEMBRO	**67.4** 61.1-75.2 rank 1	**65.6** 59.3-73.4 rank 1	**65.0** 58.6-72.7 rank 1
admin inst, party org suborg	**org** ORGANIZACAO 19.1%	ADMINISTR. INSTITUICAO EMPRESA SUB	**58.7** 53.3-65.4 rank 1	**50.0** 45.3-55.9 rank 1	**56.3** 51.0-62.7 rank 1
date hour period cyclic	TEMPO 8.6%	DATA HORA PERIODO CICLICO	**75.5** 79.8-71.7 rank 1	**72.2** 76.1-68.7 rank 1	**73.5** 77.7-69.8 rank 1
address admin top virtual site	**top** LOCAL 24.8%	CORREIO ADMINISTR. GEOGRAFICO VIRTUAL ALARGADO	**69.6** 75.1-64.8 rank 3	**64.3** 69.4-59.9 rank 4	**68.6** 74.1-63.9 rank 3
product, V copy, tit artwork pub	**tit** OBRA 4.3%	PRODUTO REPRODUZIDO ARTE PUBLICACAO	**21.3** 22.3-20.4 rank 1	**16.5** 17.3-15.8 rank 2	**19.7** 20.6-18.9 rank 1
history occ event	**event** ACONTECIMENTO 2.4%	EFEMERIDE ORGANIZADO EVENTO	**36.2** 28.9-48.6 rank 4	**30.8** 24.6-41.3 rank 4	**32.7** 26.0-43.8 rank 4
genre,brand, disease,idea, school,plan, author,abs-n	**brand** ABSTRACCAO 9.2%	DISCIPLINA,MARCA, ESTADO,IDEIA, ESCOLA,PLANO, OBRA,NOME	**43.1** 47.3-39.6 rank 1	**39.6** 43.3-36.4 rank 1	**41.4** 45.4-38.0 rank 1
object mat class,plant	**object** COISA 1.6%	OBJECTO SUBSTANCIA CLASSE	**31.3** 25.4-40.7 rank 1	**31.2** 25.5-40.3 rank 1	**31.3** 25.4-40.7 rank 1
prednum quantity currency	VALOR 9.5%	CLASSIFICADO QUANTIDADE MOEDA	**84.3** 87.0-81.7 rank 1	**82.5** 84.8-80.2 rank 1	**82.2** 84.8-79.7 rank 1

Table 12.1: Global HAREM results for PALAVRAS-NER, semantic classification absolute/total (i.e. all NE, identified or not) combined metric for 9 categories and 41 subcategories (types)

HAREM Category	combined		per category		PAL-1 F-Score
	Precision -recall	F-Score (rank)	Precision -recall	F-Score (rank)	
PESSOA	90.1-91.9	91.0 (3)	92.7-94.0	93.4 (3)	92.5
ORGANIZACAO	77.0-79.0	78.0 (5)	91.1-92.4	91.8 (7)	94.3
LOCAL	87.7-89.3	88.5 (7)	96.1-95.5	95.8 (5)	95.1
OBRA (tit, brand, V)	58.5-59.5	59.0 (3)	75.3-76.6	76.0 (3)	ABSTRACT 84.3 (tit, genre, ling) OBJECT: 57.1 (brand, V, mat)
ABSTRACCAO (genre, ling)	82.6-85.6	84.1 (1)	90.5-93.2	91.8 (1)	
COISA (brand, V, mat)	98.8-98.8	98.8 (1)	100-100	100 (1)	
ACONTECIMENTO	69.6-72.6	71.1 (5)	81.9-85.4	83.6 (5)	88.7
TEMPO	91.5-91.5	91.5 (4)	96.8-95.5	95.8 (5)	-
VALOR	94.2-95.8	95.0 (1)	96.6-97.6	97.1 (1)	-

Table 12.2: Relative HAREM performance of PAL-2.

fact that the function of human names is much more likely to stick to its lexeme category, while organisations frequently *function* as either human agents or place names[7]. The abstract and object categories of PAL-1 were not directly comparable to the ABSTRACCAO and COISA categories of HAREM, since the latter also had OBRA, drawing (book etc.) titles from PAL-1's *abstract* category and brands (unless *functioning* as objects) from the *object* category, with a number of minor subcategories and function distinctions further complicating this 2-to-3 category match.

12.4 Conclusion: Comparison with other systems

Though state-of-the-art NER systems make often make use of lexical and grammatical information, as well as extra-textual gazetteer knowledge, most do so in a framework of data-driven statistical learning, using techniques such as HMM, Maximum Entropy, Memory or Transformation-based Learning. The statistical learning approach has obvious advantages where language independence is desired, as in the CoNLL2002 and CoNLL2003 shared tasks (Sang, 2002; Sang e Meulder, 2003), but language-specific systems or subsystems may profit from explicit linguistic knowledge (hand-written rules or lexica), as e.g. in a number of Scandinavian NER systems (Bick (2004) and Johannessen et al. (2005)). Petasis et al. (2004) describes a 4-language NERC system with hybrid methodology, where the French section relies on human modification of rules machine-learned from an human-annotated corpus. PALAVRAS-NER stands out by being entirely based on hand-written rules, both locally (morphological pattern recognition) and globally (sentence context) - not only in assigning the grammatical tags used as context by the NER-system, but also within the latter itself. However, though PAL-2's rule based method worked best in the Portuguese HAREM context, with overall F-Scores of 80.6 for identification and 63.0/68.3 for abso-

[7] the *commercial* vs. *administrative* distinction also increases PAL-2's error risk

lute/relative category classification, it is difficult to compare results to those achieved for other languages, due to differences in metrics and category set size. In the CoNLL shared tasks on newspaper-text, the best absolute F-scores were 88.8 (English), 81.4 (Spanish), 77.1 (Dutch) and 72.4 (German) for a 3-way category distinction: *person*, *organisation*, *place* (plus *miscellaneous*), and given PALAVRAS-NER's high *relative* scores for these categories (93.4, 91.8 and 95.8), its lower total scores may well be due to suboptimal identification, reflecting either shortcomings of the PAL-2 rule system in this respect or linguistic-descriptive differences between the gold-standard CD and PALAVRAS-NER[8]. However, it is not at all clear how the CoNLL systems would have performed on a large (41) subcategory set and HAREM style mixed-genre data[9]. On the other hand, HAREM's category-specific and relative rank scores clearly show that there is much room for improvement in Pal-2, especially for the place and event categories, where it didn't rank highest (Table 12.1). Also, Pal-2 appears to be *relatively* better at name chunk identification than at classification, since it ranked lower in the relative scores (on correct chunks only) than in the absolute scores (identification task included). However, improvements do not necessarily have to be Pal-2-internal: Given an integrated research environment and a modular perspective (for instance, a cgi-integrated web-interface), a joined Portuguese HAREM system could act on these findings by delegating the identification and classification tasks to different systems and by applying weighted votings to exploit the individual strengths of specific systems, thus seamlessly integrating rule based and statistical systems.

Acknowledgments

The authors would like to thank the Linguateca team for planning, preparing, organising and documenting HAREM, and for making available a multitude of evaluation metrics in a clear and accessible format.

[8] Such differences are particularly relevant for a system built by hand, not from training data. Thus, PAL-1 made far fewer chunking errors when evaluated internally (Bick, 2003).

[9] The MUC-7 MENE-system (Borthwick et al., 1998), for instance, experienced an F-Score drop from 92.2 to 84.2 even within the same (newspaper) genre, when measured not on the training topic domain, but in a cross-topic test.

Capítulo 13

RENA - reconhecedor de entidades

José João Dias de Almeida

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 13, p. 157–172, 2007.

O RENA (Alves e Almeida, 2006) é um protótipo de sistema de extracção/marcação de entidades mencionadas construído por Edgar Alves sob supervisão de J.J. Almeida no âmbito do projecto IKF.

O projecto IKF (Information + Knowledge + Fusion) (Silva, 2004; Oliveira e Ribeiro, 2003; Tettamanzi, 2003) foi um projecto Eureka (E!2235) envolvendo participantes universitários e industriais de seis países, cuja finalidade básica foi o desenvolvimento de uma infraestrutura distribuída baseada em ontologias para o manuseamento inteligente de conhecimento – contemplando um ambiente documental multifonte e distribuído.

O IKF *framework* baseia-se num modelo de representação de conhecimento sofisticado (baseado em ontologias, facetas, lógica vaga (*fuzzy*), informação incompleta, e raciocínio temporal) (Silva, 2004), e é constituído por um conjunto de módulos envolvendo, entre outros:

1. Extractores básicos – extracção de conhecimento a partir de documentos heterogéneos de modo a construir um sistema de assimilação documental:
   - organização de um conjunto de ficheiros de modo a construir uma base documental
   - extracção de informação (rica) a partir desse conjunto de documentos
   - classificação facetada, *fuzzy* vaga e parcial de documentos e da informação neles contida
   - fusão da informação extraída dos vários documentos
2. Renovador de conhecimento (*Knowledge Renovator*) (Oliveira e Ribeiro, 2003) – ligada à evolução (temporal ou não) da informação e do conhecimento.
3. Enfermaria do Conhecimento – ligado a sistemas legados, e à reparação de inconsistências por razões variadas.
4. Navegadores – um conjunto de navegadores sobre a base de conhecimento e a base documental.

A título de exemplo de aplicação considere-se o caso da assimilação documental de caixas de correio electrónico: ao extrair e fundir conhecimento, pretende-se obter informação capaz de responder a perguntas como:

- quem é a pessoa F?
- qual a lista dos amigos de F? quais os parceiros de X?
- qual o conjunto de áreas de interesses de Y?
- que documentos são relevantes acerca de Z?

Tendo em vista estes objectivos, para além das tarefas principais (as tarefas estruturais ligadas ao projecto), foi realizado um conjunto de pequenas tarefas/experiências exploratórias, envolvendo recursos muito limitados e frequentemente envolvendo alunos finalistas.

É neste contexto que surge o protótipo RENA que, não fazendo directamente parte do projecto IKF, foi desenhado como um caso de estudo com a intenção de fazer extracção de conhecimento simples – extracção de uma base de entidades:

$$Rena : Ficheiro^* * BaseEnt \longrightarrow BaseDoc * BaseEnt * \ldots$$

13.1 Descrição do RENA

Na sequência do enquadramento anteriormente descrito, o protótipo RENA tem como intenção uma extracção tão rica quanto possível de informação, com vista a ser usada por sistemas de processamento e fusão de conhecimento (e em particular no projecto IKF).

À medida que a ferramenta RENA foi sendo projectada, decidiu-se que era importante que pudesse ser usada por um conjunto menos restritivo de aplicações – ou seja, que pudesse ser usada em modelos semânticos menos sofisticados (um Micro-IKF).

Dum modo resumido o RENA é um sistema REM constituído por:

- Uma biblioteca Perl:
  1. baseada num conjunto de ficheiros de **configuração** alteráveis,
  2. com funcionalidade para **extrair a lista das entidades** a partir de conjuntos de textos,
  3. ou, em alternativa, **marcar entidades** num conjunto de texto.
- Um conjunto de programas para fazer processamento de entidades.

Muita da capacidade de extracção depende de um conjunto de ficheiros e de regras – elementos de configuração – que descrevem conhecimento geral e regras de contexto usados na extracção.

Pretendeu-se desde o início que esses elementos de configuração fossem *externos* ao RENA, de modo a que o utilizador os pudesse adaptar à sua visão do mundo e à sua intenção concreta de utilização. Assim, foi requisito dos elementos de configuração que fossem legíveis, expressivos e compactos.

13.1.1 Estrutura interna do RENA

Do ponto de vista algorítmico, o RENA:

1. começa por procurar entidades e construir uma sequência de textos simples e entidades: $(texto \times entidade)^*$
2. seguidamente, esse objecto é processado por uma série de filtros com assinatura

$$f : (texto \times entidade)^* \rightarrow (texto \times entidade)^*$$

que vão processar os pares texto-entidades, enriquecendo a informação, alterando fronteiras e unindo zonas, com base nos recursos de configuração atrás referidos e utilizando ferramentas internas ou externas (como por exemplo o analisador morfológico jspell (Simões e Almeida, 2002; Almeida e Pinto, 1995)).

3. no final, de acordo com a saída pretendida, é criado:
   - um texto com as entidades anotadas
   - um resumo das entidades presentes

   O formato final pretendido pode ser:
   - *XML*, contendo uma versão do texto original onde são anotadas todas as referências a entidades encontradas.
   - *YAML* (Ben-Kiki et al., 2005, 2006), descrevendo todas as entidades com alguma referência no texto, bem como todas as classificações atribuídas.

Os filtros que gerem texto nos formato acimas referidos, que, aliás, podem ser desactivados, fazem tarefas como:

- tratamento de entidades com elementos de uma única letra,
- tratamento de aspas ligado às entidades
- remoção de entidades entre aspas (este filtro só deverá ser usado se se pretender ignorar este tipo de ocorrências).
- tratamento de entidades com traços interiores (por exemplo, *Benfica-Sporting*)
- tratamento de entidades em início de frase
- enriquecimento por análise de regras de contexto
- enriquecimento por análise do almanaque de nomes
- enriquecimento por análise do almanaque de cultura geral
- tratamento de acrónimos
- reconhecimento e unificação de entidades iguais (ou abreviadas) e criação de atributos de ligação entre as várias ocorrências da mesma entidade.

13.1.2 Ficheiros de configuração

A configuração de base do RENA é constituída por um conjunto de recursos:

1. Ontologia de classes – que estabelece relações (hierárquicas) entre os tipos de entidades existentes;
2. Tabela de contextos – com regras para deduzir qual o tipo das entidades com base no contexto esquerdo;
3. Almanaque de cultura geral – onde se registam termos/conceitos geográficos, culturais, patrimoniais, cultura geral;
4. Sistema de tratamento de nomes – em que se guardam alguns dos nomes/apelidos mais comuns e regras para determinar se um nome próprio se refere a pessoas;
5. Tabela de conversão/adaptação de nomes;
6. Tabela de contextos atributivos (em fase de construção).

Vários destes recursos são definidos usando linguagens de domínio específico (DSL) construídas com a intenção de conseguir uma descrição eficaz dessa informação.

Seguidamente vamos detalhar alguns destes recursos e apresentar alguns exemplos.

Ontologia de classes

A ontologia de classes armazena os tipos de entidades e respectivas relações. A definição dos tipos de entidades e dos seus relacionamentos é uma actividade delicada, sensível: corresponde a uma descrição do nosso modo de ver o mundo. Há zonas desta ontologia que são facilmente reutilizáveis, outras que são dependentes do projecto concreto.

Normalmente é importante ter controlo total sobre esta ontologia pelo que ela deve ser construída manualmente. No entanto, alguma zonas podem ser obtidas por aprendizagem automática.

No nossos exemplos vimos que pode haver utilidade em usar (pequenos) extractos de ontologias como o CDU, o tesauro da Unesco, o tesauro da Biblioteca de Alexandria, ou outros sistemas classificativos.

A existência deste recurso é crucial para se conseguir:

- fazer inferência parcial de tipos de entidades,
- facilitar a fusão de análises complementares,
- obter uma maior adaptabilidade da informação extraída.

```
- pessoa:
   - advogado
   - arquitecto
   - atleta:
      - futebolista
      - nadador
   - escritor:
      - poeta
   - jornalista
   - militar:
      - general
      - almirante
      - brigadeiro
      - sargento
      - tenente
      - capitão
   - músico:
      - compositor
      - pianista
      - trompetista
   - político:
      - presidente da república
      - deputado
```

Figura 13.1: Extracto da ontologia de classes.

Sempre que possível pretende-se que esta ontologia tenha um grão fino de modo a poder registar toda a informação extraída, mas ao mesmo tempo deseja-se que permita uma posterior abstracção/síntese.

A dimensão e conteúdo da ontologia de classes deverá ter em conta a pragmática e o conteúdo e dimensão do conjunto documental em análise. No caso concreto, utilizamos uma ontologia exemplo com cerca de 120 classes. Na Figura 13.1 representa-se um extracto da ontologia de classes (visto como uma taxonomia para mais fácil visualização).

Saliente-se mais uma vez que a ontologia para descrever as classes difere conforme a intenção e o conjunto de documentos em análise. Por exemplo, embora haja muitas coisas comuns, há uma enorme diferença entre o conjunto das classes referentes a um arquivo de biologia, a um arquivo de etnomusicologia, ou a um arquivo de *software* de PLN.

Tabela de contextos

A tabela de contextos permite que de um modo compacto se possa definir uma associação entre uma **expressão de contexto** esquerdo e uma classe (ver Figura 13.2).

```
  cidade (de do da)     => cidade !lctx
  freguesia (de do da)  => freguesia
  distrito (de do da)   => distrito
  concelho (de do da)   => concelho/90
  estado (de do da)     => estado
  capital               => cidade !lctx

  (Rio Oceano Lago Mar Serra Cordilheira) => $_
  Cabo (do de da)       => cabo
  Golfo (do de da)      => golfo

  (Lugar Largo Lg. Praça Rua R. Avenida) (de da do das dos)?  => lugar
  (Travessa Beco Quinta Viela Rotunda) (de da do das dos)?    => lugar

# Monumentos $

  (Convento Mosteiro Igreja Ig. Palácio Museu Sé) (de da)? => monumento
```

Figura 13.2: Extracto da tabela de contextos.

Note-se que:

- as regras podem ter valores de confiança, de modo a permitir distinguir entre indícios mais fortes e indícios mais fracos,
- a grafia maiúscula é usada para indicar se o termo de contexto esquerdo deverá ou não ser incluído na entidade,
- os padrões das regras podem incluir variantes alternativas, elementos opcionais, comentários, etc.

Embora esta tabela possa ser construída, consolidada e revista manualmente, uma boa base de início pode ser obtida através da extracção dos bigramas de palavras do contexto direito e do início de entidade (das entidades antes ou depois de classificadas) – podendo ser usadas técnicas de *bootstraping* habituais em situações idênticas[1].

Muitas regras são gerais; no entanto, no caso geral, esta tabela depende do problema concreto.

Almanaque de cultura geral

Conforme atrás se referiu, o almanaque de cultura geral pretende guardar alguma informação de cultura geral de índole diversa.

[1] No estado actual do RENA, há apenas um esqueleto de ferramentas de ajuda à construção dessa tabela segundo o método referido.

```
Rio Douro =
 rio Douro
        IOF =>  rio
        AFLUENTES =>    rio Mau,
                rio Sousa,
                rio Varosa,
                rio Tâmega,
                rio Pinhão,
                ....
                rio Torto,
                rio Távora,
                rio Esla,
                rio Tua
        COMPRIMENTO =>  927
        FOZ =>  Porto
        IN =>   Portugal,
                Espanha
        NASCE =>        serra do Urbião
        PASSA_EM =>     barragem do Pocinho,
                barragem de Miranda,
                barragem de Crestuma,
                Miranda do Douro,
                barragem do Carrapatelo,
                Régua,
                barragem da Bemposta
```

Figura 13.3: Extracto da informação existente no almanaque de cultura geral.

Presentemente este almanaque tem por base informação criada no âmbito do projecto T_2O (Almeida e Simões, 2006a,b), e a informação associada a cada entidade é por vezes rica (ainda que heterogénea): além duma classe de base, pretende-se armazenar um conjunto de atributos e ligações tão rico quanto possível.

Simplificadamente este almanaque corresponde a uma vista sobre a projecção de uma ontologia T_2O, seleccionando-se os termos por exemplo referentes a geografia, personagens famosas, ou eventos.

Na Figura 13.3 mostra-se um extracto da informação existente no almanaque associada a **Rio Douro**, demonstrando a intenção de dispor de um conjunto de dados de base rico e estruturado que permita processamento posterior (interactivo ou não).

Sistema de tratamento de nomes

A intenção subjacente ao **sistema de tratatamento de nomes**, demonstrado na Figura 13.4, é permitir dispor de dados para determinar se certos identificadores constituem (ou não)

```
26.62287        Maria   nome
13.70273        Ana     nome
6.85846         José    nome
5.16030         Silva   apelido
4.90977         António nome
3.95357         Carla   nome
3.51606         Manuel  nome
3.50263         João    nome
...
0.02148         Dinis   misto
```

Figura 13.4: Extracto do sistema de tratamento de nomes.

prováveis nomes de pessoas (quando não houver fortes indícios noutro sentido).

De um modo simplificado, guarda-se uma tabela que indica a taxa de ocorrência (por milhão de palavras) de determinada palavra, indicando ainda se o seu uso é preferencialmente nome, apelido ou misto. Esta tabela tem por base uma lista de 150.000 nomes completos, de várias proveniências.

Tabela de conversão/adaptação de nomes

Dado que há necessidade de poder usar ontologias de classes e tabelas de contextos adaptadas a cada projecto concreto, temos necessidade de criar mecanismos para conversão de classes.

Esta tabela pretende criar um grau de indirecção de modo a permitir uma mais fácil alteração da estrutura da ontologia de classes, criando alguma independência entre a ontologia de classes, o almanaque e a tabela de contextos.

Tabela de contextos atributivos

O objectivo da tabela de contextos atributivos é, para além de eventualmente inferir classes, ajudar a inferir mais atributos, factos e informações acerca das entidades – numa palavra, informação mais rica.

Considere-se o seguinte extracto exemplo:

```
a atleta portuguesa A :: atleta(A), nacionalidade(A,portuguesa)
X , no norte de Y     :: geo(X), geo(Y), norte(X,Y)
o francês Z           :: pessoa(Z), nacionalidade(Z,francês)
```

Quando for encontrada uma ocorrência do tipo **...a atleta portuguesa Rosa Mota ...** é feita a inferência de que Rosa Mota é uma atleta (e portanto uma pessoa, etc), e que o atributo nacionalidade da entidade em causa é preenchido com o valor **portuguesa**.

Esta tabela é crucial para aumentar a riqueza da informação extraída. Até ao momento, ela tem sido construída manualmente, no entanto há planos para a construção de ferramentas que proponham regras e extraem pistas a partir de textos anotados.

13.2 Participação no HAREM

A participação no HAREM foi muito importante e produtiva para nós já que:

- envolveu discutir e trocar impressões com os outros participantes e com a organização,
- envolveu lidar com um problema para o qual o RENA não tinha sido pensado,
- levantou uma série de questões que nunca nos tinham ocorrido referentes à necessidade de criação de camadas de adaptação de notações e de adaptação de estruturas classificativas.

Há, no entanto, alguma diferença entre o tipo de avaliação que pretendíamos (mais ligada a um uso de extracção de informação enciclopédica) e a avaliação feita no HAREM.

Os resultados finais ficaram aquém do que seria possível por várias razões:

- um dos autores do RENA (Edgar Alves) não participou (por ter já deixado a universidade)
- houve decisões do RENA que não seguem as propostas do HAREM e das quais não quisemos prescindir,
- com o pouco tempo que nos foi possível dedicar ao RENA, optámos por melhorar alguns módulos que, não sendo os mais importantes para a avaliação no HAREM, são cruciais para o RENA.

Genericamente a identificação de entidades foi bem conseguida apesar de termos optado por não marcar valores numéricos em geral por nos parecer menos interessante para o RENA.

Os maiores problemas resultaram de uma diferente filosofia no que diz respeito às classes – diferente filosofia semântica. Enquanto que o HAREM pretende marcar a ocorrência específica em contexto específico, o RENA está menos preocupado com a ocorrência concreta mas com a entidade referida; está mais preocupado com a extracção de informação rica de cariz enciclopédico.

Considere-se o seguinte exemplo concreto:

```
(...) os diários "<OBRA TIPO="PRODUTO" >Jornal Tribuna de Macau</OBRA>" e
<OBRA TIPO="PRODUTO">Macau Hoje</OBRA> (...)
```

De acordo com a nossa intenção de extracção de informação enciclopédica, afirmar que o *Jornal Tribuna de Macau* é uma `OBRA:PRODUTO` seria completamente inaceitável: a resposta útil para o RENA (independentemente de o termos conseguido extrair) é **Jornal** ou **Jornal diário**.

Do mesmo modo demos preferência a **monumentos** em relação aos `LOCAL:ALARGADO` ou às `OBRA:ARTE`.

A participação do RENA na tarefa de classificação semântica foi feita da seguinte forma:

1. extrair a informação e usar apenas a classificação geral de acordo com a ontologia RENA,
2. traduzir (de acordo com uma tabela de tradução escrita manualmente) cada classificador RENA num par categoria:tipo do HAREM.

Esta abordagem também introduziu erros adicionais. Por exemplo, algumas classes, como monumento, acabaram por não encontrar um classificador natural na estrutura classificativa do HAREM.

Optámos por não fazer a tarefa de classificação morfológica por não nos parecer tão relevante para a nossa ferramenta específica e para não dispersar (e congratulamo-nos com a versatilidade do sistema HAREM de poder aceitar marcações parciais).

No próxima secção apresentamos mais alguns exemplos e situações em que os modelos HAREM e RENA divergiram.

13.3 Subsídio para a discussão sobre futuras edições

A organização e planeamento do HAREM foi muito boa. No entanto e tendo em conta futuras organizações vou enunciar algumas coisas que me parece ser vantajosas.

Em resumo, as propostas para futuras versões são:

1. uso de documentos seguindo (totalmente) a norma XML
2. uso claro e extensível de metadados nas colecções

$$coleccao = (MetaData \times Texto)^*$$

3. migração de taxonomia 2 níveis para uma ontologia de classes multi-nível
4. uso de etiquetagem mais versátil.

13.3.1 Uso de documentos seguindo XML

A migração para documentos XML, torna mais fácil tirar partido de um conjunto de ferramentas no sentido de:

- permitir verificar se os documentos (colecções e submissões) são bem-formados e se são válidos,
- ser claro e definido qual o sistema de encoding usado,
- poder obter mais facilmente uma variedade de vistas (*pretty-printers*), resumos, e reordenações dos documentos, de modo a se adaptar a diversas finalidades. (Usando CSS, XSL, etc.),
- ser trivial o cálculo de um conjunto de estatísticas e pesquisas (Usando XPath e afins).

13.3.2 Uso claro e expansível de metadados nas colecções

A existência de metadados nas colecções foi algo que a organização teve em conta. Existe, por exemplo, um elemento `<DOC>`, com metadados variante linguística e género textual.

```
<DOC>
  <DOCID>HAREM-871-07800</DOCID>
  <GENERO>Web</GENERO>
  <ORIGEM>PT</ORIGEM>
...
```

Por um lado, parece-me que os valores do atributo género cobrem mais que uma faceta: um documento *político* (conteúdo temático) poderá ser também uma *entrevista*, ou estar disponível (suporte) em *Web, CorreioElectrónico*. Ou seja, seria útil múltiplas ocorrências de géneros, ou separar esta informação em mais do que um campo.

Por outro lado, gostaria de ver um elemento `META` que agrupasse todos os metadados do documento de modo a permitir que possa haver mais fácil enriquecimento (por parte do HAREM ou de outro qualquer uso futuro).

13.3.3 Questões ligadas à estrutura classificativa usada

Cada entidade marcada está a ser classificada "semanticamente".

Originalmente o MUC propôs um sistema classificativo com 3 categorias e 7 tipos. O HAREM propôs subir a fasquia para uma categorização com 10 categorias e 41 tipos. A meu ver essa decisão foi necessária e acertada.[2]. Havendo uma taxonomia a dois níveis, há naturalmente a hipótese de participações parciais:

- nível 0 –> marcar apenas as entidades
- nível 1 –> apresentar apenas as classificações do primeiro nível

[2] Genericamente subir a fasquia é bom quando houver pelo menos um atleta que a transponha...

- nível 2 –> apresentar a classificação completa.
- ou ainda escolher uma subárvore da taxonomia em causa.

Por outro lado, foi construída uma função de conversão

$$harem2muc : Charem \longrightarrow Cmuc$$

que mapeia classificações HAREM em classificações MUC. – tornando possível a comparações de resultados (medidas de acerto) entre as duas competições[3]. Esta função de mapeamento entre os dois sistemas para a maioria dos casos é simples e natural, havendo no entanto zonas da estrutura HAREM que são difíceis de mapear em MUC (o que não surpreende nem impede a leitura dos valores após conversão).

Dum modo semelhante parece-me que há zonas da taxonomia HAREM que são pouco naturais e claras – vistas pelo prisma de representação de conhecimento. Constatou-se naturalmente dificuldades em arranjar consenso entre os participantes em relação ao referido sistema de classificação do HAREM, o que é natural e habitual nestas actividades, e que me parece não ter constituído obstáculo importante ao funcionamento.

Genericamente, a marcação combinada tem o seguinte aspecto:

```
<Nivel1  tipo="Nivel2">Entidade encontrada</Nivel1>
```

No que diz respeito à estrutura classificativa, os problemas com que deparamos são:

1. Apesar de existir uma etiqueta de alternativa (`<ALT></ALT>`) para descrever alternativas de que sequências de palavras compõem a entidade (vagueza na identificação textual), uma notação (`|`) para vagueza/indefinição das classes semântica e ainda uma classe especial `outra` para situações *duvidosas*, não vejo claramente como descrever ao nível da marcação:
   - **ignorância total** (ex: *o X é interessante* – não sei nada acerca de X). Um humano normalmente saberá classificar uma ocorrência mas é frequente um ferramenta não o saber; nessa situação pretendemos anotar essa ignorância.
   - dúvida (ex: *o Porto é imprevisível*: ou é uma cidade ou um clube de futebol mas não as duas ao mesmo tempo – só consegui concluir alguma informação parcial),
   - classificação múltipla (*na Biblioteca da Universidade de Coimbra encontramos o espírito barroco* – acho válidas duas ou mais classificações: Obra de arte, Local Biblioteca, ...)

[3] **Nota do editor:** A comparação entre os resultados do HAREM e os do MUC e a conversão das respectivas etiquetas não é um assunto trivial, contudo, , como é discutido nos capítulos 3 e 4.

ou seja:

```
<nivel1 tipo="não faço ideia">e1</nivel1>
<nivel1 tipo="das duas uma:A ou B mas tenho dúvidas qual">e1<nivel1>
<nivel1 tipo="tanto A como B são tipos de">e1<nivel1>
```

Estou convicto de que o nível de ambiguidades/ignorâncias aparece mais na resposta dos sistemas do que na resposta de humanos.

2. Há situações (ao fazer a "formatação"a dois níveis) em que certas sub-árvores são facetas (quase independentes) levando a que faça sentido duas classificações, e que por vezes a solução oficial "perca"certas facetas e aspectos cruciais à caracterização da entidade em causa.

   Considere-se o seguinte exemplo da colecção dourada:

   ```
   <LOCAL|OBRA TIPO="ALARGADO|ARTE">Biblioteca Pública</LOCAL|OBRA>
   ```

   A referida biblioteca é um lugar, um edifício ou semelhante mas simultaneamente é património artístico, (é uma obra de arte). De certo modo, ser ou não obra de arte é uma faceta que poderemos querer aplicar a edifícios, livros, cidades e outras classes. Portanto constitui uma informação que deveria poder coexistir com a informação da classe a que se refere. Ou seja aquela biblioteca é simultaneamente um edifício e uma obra de arte[4].

3. genericamente a existência de herança múltipla complica certas zonas da estrutura classificativas.

   Considere-se o seguinte exemplo teórico. Se a minha maneira de ver o mundo considerar que:

   ```
   palácio    é uma subclasse   obras de arte
   palácio    é uma subclasse   edifícios
   ```

   (ou seja palácio tem dois pais, ou tem herança múltipla dessas duas classes) uma marcação em taxonomia a dois níveis (e já agora usando uma notação semelhante à do HAREM) tenderá a ver uma ambibuidade artifical entre

   ```
   <ObraDeArte tipo="palácio">...
   <Edifício   tipo="palácio">...
   ```

   Em situações como esta o uso de **palácio** (sem obrigação de escolher qual dos pais) tenderia a simplicar as coisas[5].

[4] **Nota dos editores:** isso é precisamente o que a notação do HAREM quis dizer: que aquela ocorrência de Biblioteca pública é simultaneamente as duas coisas.

[5] **Nota dos editores:** essa é exactamente a filosofia do HAREM: não ver ambiguidades quando não existem. No caso em questão, seria **ambas** as coisas: `<OBRA | EDIFICIO>`. O HAREM nunca marca ambiguidade, porque assume que os humanos conseguem distinguir. O caracter ' '|' ' indica sempre vagueza.

4. por vezes o enquadramento das ferramentas concorrentes força estruturas classificativas diferentes das usadas e ligeiramente "antagónicas". Isto é apenas uma constatação que complica a participação e para a qual não há uma solução óbvia mas que ainda assim descrevemos:

   Considere-se o seguinte par de exemplos da colecção dourada:

   ```
   Visite o <OBRA TIPO="PRODUTO">DataGrama Zero</OBRA> a Revista
   Eletronica ( ...)
   A revista foi denominada <ABSTRACCAO TIPO="NOME">Medicina e
   Cultura</ABSTRACCAO> (...)
   ```

   Independentemente do contexto linguístico em que estas entidades possam estar a ser usadas, dum ponto de vista de representação de conhecimento pretende-se tirar partido de que esta duas revistas têm muito em comum (classes idênticas ou aparentadas) e será completamente inaceitável ignorar/esquecer que *Medicina e Cultura* é uma *revista*.

A granularidade e capacidade distintiva

Considere-se a questão ligada com os conceitos *Portugal*, *país*, *entidade geográfica*, etc:

O seguinte conjunto de relações binárias pode ser usado para descrever (algumas das) propriedades do conceito *Portugal*:

```
Portugal IOF  país
país     ISA  entidade geográfica
país     ISA  instituição administrativa
país     ISA  povo
...
```

Numa situação como a do IKF/RENA não dispomos de informação suficiente para resolver devidamente essa questão de escolher entre os vários pais possíveis e, assim, optámos por baixar a fasquia, crentes de que ter uma classificação que falhe 40% dos casos é pior do que dizer que é simplesmente um país.

Na visão IKF/RENA a nossa intenção corresponde a ir decorando a árvore de conhecimento com todos os atributos que conseguirmos obter (trata-se de uma finalidade específica nossa), ou seja pretendemos juntar a *Portugal* os atributos ligados a país nas suas várias acepções e usos (presidente da república, língua, rios, área, etc).

Esse tipo de junção e processamento de atributos, heranças, etc, cria restrições ao tipo de árvores classificativas a usar: a relação subclasse (nível1 – nível2 da estrutura HAREM) passa a ter maiores responsabilidades...

13.3.4 Sugestão para futuras edições

Em resumo, para futuras edições propunha:

- Etiquetagem mais prática:
  - uma única etiqueta Entidade `<ent ...>...</ent>`
  - um atributo *tipo* `<ent t="país">...</ent>`
  - com notação clara para alternativas `<ent t="t1|t2"> ...`
  - com notação clara para multiclassificação `<ent t="t1;t2"> ...`
  - para informação parcial = escolher um nó mais acima na árvore classificativa (caso extremo = topo = entidade)
  - um atributo de unificação para permitir ligar referências à mesma entidade
- Ontologia multi-nível de classes, com herança múltipla
- Identificadores de classe mais claros e únicos – a questão da clareza é crucial[6] para o contexto de extracção de informação onde o RENA se encaixa: dizer que *Palácio de Vila Flor* é um `LOCAL:ALARGADO` é inaceitável do ponto de vista de extracção de informação enciclopédica[7].

13.4 Conclusões e trabalho futuro

A participação no HAREM foi muito positiva, embora, por questões conjunturais, não tenha sido possível tirar partido de uma série de iniciativas.

A participação do RENA no HAREM seguiu uma abordagem que não visava maximizar o resultado final da avaliação, mas antes o tentar ajudar à evolução do RENA de acordo com os nossos objectivos imediatos (que por vezes não coincidiram com os do HAREM).

Apesar das evoluções conseguidas, o estado actual do RENA é de protótipo.

Ao nível do trabalho futuro, há genericamente o objectivo:

- melhorar as regras de inferência e unificação e resumo
- criar um processador estrutural
- melhorar o sistema de tratamento de nomes incluindo também dados estrangeiros
- documentar melhor a interface de biblioteca Perl.

[6] No geral, em teoria da classificação há a recomendação de que cada classificador deverá, sempre que possível, ter autonomamente uma leitura clara.

[7] Como dissemos, do nosso ponto de vista, palácio, monumento, etc, seria preferível. Classificações como `LOCAL`, localidade, edificação, são também claras; `LOCAL:ALARGADO` por si só é de leitura pouco clara e parece-me significar algo como *local que não se encaixa nas outras subcategorias*.

Capítulo 14

O SIEMÊS e a sua participação no HAREM e no Mini-HAREM

Luís Sarmento

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 14, p. 173–189, 2007.

O SIEMÊS foi desenvolvido por uma equipa de três elementos (Luís Sarmento, Luís Cabral e Ana Sofia Pinto) do Pólo do Porto da Linguateca, com o objectivo especifico de participar no HAREM (Seco et al., 2006). A ideia inicial da participação do Pólo do Porto no HAREM era aproveitar o conhecimento e a tecnologia de extracção de terminologia desenvolvida para o Corpógrafo (Sarmento et al., 2004) e melhorá-la para se conseguir a marcação e classificação de certos elementos que o HAREM contemplava, tais como *(T|t)eorema de Fermat*, *(C|c)onstante de Planck* ou *(S|s)índroma de Alzheimer*. Este género de estruturas, tradicionalmente mais próximas da terminologia, não têm sido tratadas devidamente pelos sistemas de REM mas, quer pelo facto de incluírem efectivamente um nome próprio quer pelo facto de serem muito frequentes em diversos géneros de texto, mereceram uma atenção especial por parte dos organizadores do HAREM. Apesar desta motivação inicial bem definida, a equipa do Pólo do Porto da Linguateca decidiu alargar o objectivo especifico e tentou desenvolver um sistema que fosse capaz de identificar e classificar todas as categorias previstas no HAREM. Esse sistema foi baptizado de SIEMÊS - Sistema de Identificação de Entidades Mencionadas com Estratégia Siamesa.

O SIEMÊS assenta na convicção de que o processo de classificação de entidades mencionadas poderá ser feito com maior robustez através da *combinação* de regras de análise do contexto com a consulta de almanaques, de onde se pode retirar informação muito relevante e que facilita a posterior análise. O SIEMÊS assume que, se for possível numa primeira fase, através da informação existente em almanaques, gerar um conjunto de hipóteses de classificação para um determinado candidato, torna-se possível numa segunda fase desambiguar semanticamente a *classe* e a *forma de menção* do referido candidato usando regras de análise do contexto relativamente simples. Esta dupla estratégia de classificação - que faz uso de um almanaque e de um banco de regras - foi a inspiração para o nome do sistema.

A filosofia base do SIEMÊS tem como principal objectivo garantir um desempenho *robusta* em cenários onde se pretenda classificar uma grande variedade de entidades. Procura-se assim amenizar as dificuldades provenientes da enorme combinatória de contextos que se encontra em tais cenários. No caso da tarefa definida no HAREM, a diversidade de cenários torna-se particularmente complexa dado o elevado número de classes a discriminar, o que apontaria para a criação de enormes bancos de regras capazes de lidar com todos os casos. Tais regras podem necessitar de recursos semânticos bastante desenvolvidos (tais como léxicos categorizados semanticamente) que não se encontram publicamente disponíveis.

Note-se que foi assumido desde início que a forma de utilização dos almanaques pelo SIEMÊS não se limitaria à simples consulta booleana de entradas, isto é de verificar se determinada entrada faz ou não parte do almanaque. O SIEMÊS procura explorar a informação nos almanaques de uma forma mais flexível, seguindo a ideia de que há palavras típicas de certas classes de entidades, cujos nomes acabam por apresentar alguma homoge-

neidade lexical que poderá ser explorada para fins de classificação. No SIEMÊS, o papel do almanaque é o de poder servir de base de comparação com um determinado candidato e gerar hipóteses de classificação em conformidade. As hipóteses de classificação mais verosímeis para o candidato em causa são as classes do almanaque onde se encontra exemplos mais "semelhantes" ao próprio candidato.

Foi com o objectivo de testar esta ideia que o SIEMÊS participou no HAREM, fazendo uso do almanaque REPENTINO (Sarmento et al., 2006) que foi desenvolvido paralelamente e em estreita relação. O REPENTINO armazena 450.000 exemplos de nomes de entidades distribuídos por 11 classes e 103 subclasses. Grande parte das instâncias presentes no REPENTINO foram compiladas usando métodos semi-automáticos a partir de grandes corpora, ou foram obtidas a partir de sítios web que continham listas de instâncias específicas. Os exemplos recolhidos através destas duas estratégias foram verificados e organizados manualmente.

Os resultados obtidos pelo SIEMÊS no HAREM foram suficientemente interessantes para continuar a investir nesta aproximação. Assim, no sentido de resolver vários problemas de engenharia de *software* da primeira versão SIEMÊS, decidiu-se, já no âmbito do plano de doutoramento do autor, re-implementar totalmente o sistema mantendo a filosofia de classificação, e expandindo-a ainda com novas capacidades. Assim, a actual versão do sistema, o SIEMÊS v2, possui uma arquitectura totalmente modular, o que permitiu realizar durante o Mini-HAREM uma avaliação por componentes do sistema. Esta avaliação ajudou a retirar indicações interessantes acerca da natureza do problema de REM e da eficiência das várias estratégias possíveis na sua resolução. Neste capítulo iremos por isso também apresentar alguns dos resultados dessa avaliação por componentes porque são ilustrativos da forma de funcionamento desta segunda versão do SIEMÊS, e também porque sugerem indicações valiosas para futuros desenvolvimentos.

14.1 A participação no HAREM

A arquitectura e a estratégia de classificação da primeira versão do SIEMÊS foi descrita em Sarmento (2006b), pelo que iremos neste capítulo focar mais os resultados obtidos na tarefa de classificação semântica do HAREM.

Os resultados obtidos no HAREM pelo SIEMÊS v1 foram interessantes (ver Tabela 14.1) tendo sido alcançado o segundo lugar global em medida F na tarefa de classificação, apesar de desempenhos relativamente pobres no que diz respeito às categorias numéricas (`TEMPO` e `VALOR`). Note-se contudo que, do ponto de vista absoluto, os resultados foram bastante modestos, com valores totais de precisão em torno dos 57,3% e valores de abrangência de 48,7%, resultando numa medida F de 0,537. Estes valores parecem bastante baixos quando comparados com os obtidos em provas como as MUC (Grishman e Sundheim, 1996) onde os sistemas possuem medidas F superiores a 0,9. Há contudo que referir que

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	2º	41,8	28,6	0,340
ACONTECIMENTO	1º	47,3	43,0	0,451
COISA	2º	30,0	13,3	0,185
VALOR	8º	53,3	37,4	0,434
TEMPO	4º	55,8	61,4	0,584
LOCAL	1º	64,1	69,8	0,668
PESSOA	4º	65,3	52,2	0,580
ORGANIZACAO	2º	57,6	41,2	0,480
OBRA	1º	29,8	12,0	0,171
TOTAL	2º	57,3	48,7	0,537

Tabela 14.1: Resultados da avaliação global da classificação semântica combinada do SIEMÊS no HAREM.

a dificuldade da tarefa HAREM é muito superior à da definida para as MUC tanto pelo facto de a classificação ser feita em dois níveis num total de 41 tipos, como pelo facto de a tarefa do HAREM passar por classificar a forma como a entidade é *mencionada* (ver Seco et al. (2006) e Santos et al. (2006)).

Nos resultados obtidos pelo SIEMÊS v1 no HAREM há alguns pontos interessantes. Em primeiro lugar, e apesar da estratégia simples de classificação principalmente baseada em informação de almanaque, o desempenho do SIEMÊS parece não ser inferior ao de sistemas que utilizam estratégias mais baseadas em análise do contexto. Este resultado pode parecer surpreendente até certo ponto, porque se o objectivo do HAREM era classificar a forma como a entidade é *mencionada* então o factor preponderante nessa classificação deveria ser naturalmente o *contexto*. Refira-se que o SIEMÊS recorre a uma quantidade muito reduzida de informação contextual, normalmente tem em conta apenas uma ou duas palavras de contexto para desambiguar entre algumas possibilidades geradas anteriormente em função de semelhança com o almanaque.

Em segundo lugar, o desempenho da primeira versão do SIEMÊS é elevado, do ponto de vista relativo, para classes que parecem exibir uma certa regularidade lexical. Por exemplo, no caso das categorias ACONTECIMENTO, ORGANIZACAO e ABSTRACCAO os bons resultados poderão advir do facto de as respectivas entidades serem unidade multipalavra com estrutura interna muito específica (por exemplo *Simpósio Nacional...*, *Universidade do...*, *Teorema de...*), com constituintes iniciais facilmente previsíveis, o que quase só por si discrimina a categoria intrínseca. A desambiguação da forma de menção da entidade pode, em muitos casos, ser feita com regras muito simples após a obtenção da informação acerca da respectiva categoria intrínseca; noutros casos, porém, torna-se difícil prever formas de menção diferentes da forma directa (por exemplo, para ABSTRACCAO).

Em terceiro lugar, e ao contrário de certos estudos (Mikheev et al., 1999), os resultados do SIEMÊS parecem apontar para a importância fundamental dos almanaques no reco-

nhecimento de certas classes de entidades, nomeadamente para LOCAL e OBRA. Na verdade, pode-se afirmar que quando se encontra um candidato para o qual existe uma entrada no REPENTINO correspondente a um LOCAL, é quase certo que essa entidade se refere de facto a um LOCAL. Haverá certamente casos ambíguos em que a mesma representação lexical é partilhada por várias categorias de entidades, frequentemente PESSOAs, mas, na maior parte dos casos, se não for possível identificar que a entidade corresponde a outra categoria (usando informação do contexto ou de co-referência), então pode assumir-se com bastante segurança que se trata de um LOCAL. No caso das OBRAs, a classe é de tal forma complexa (como se pode verificar da medida F, que não ultrapassou os 0,18) que a construção de regras de contexto parece ser muito difícil. Neste sentido, os almanaques acabam por ser fundamentais na classificação destas entidades, quer porque armazenam directamente o candidato em causa, quer porque permitem estabelecer semelhanças entre o candidato e outros elementos armazenados.

Quanto ao baixo desempenho do SIEMÊS nas categorias numéricas, podemos dizer que tal "falha" não é demasiado preocupante, já que a identificação e classificação deste género de entidades é feita normalmente usado gramáticas bastante extensas. No SIEMÊS estas gramáticas não foram alvo de grande cuidados, já que as limitações de arquitectura do sistema impediram a construção e manutenção de grande bancos de regras. Estas limitações de arquitectura foram, aliás, uma das grandes motivações para a construção de raiz da segunda versão do SIEMÊS onde tais problemas não subsistem.

14.2 A segunda versão do SIEMÊS

A nova versão do SIEMÊS (SIEMÊS v2) resulta de uma re-implementação total do sistema, já no âmbito do doutoramento do autor, tentando manter a filosofia geral da primeira versão mas com especial cuidado em garantir a sustentabilidade a médio e longo prazo do desenvolvimento do *software*. Deste ponto de vista, uma das grandes vantagens da segunda versão do SIEMÊS é a possibilidade de criar bancos de regras externos que são interpretados por um motor genérico, também desenvolvido para o efeito, separando totalmente o processo de criação das regras do processo de desenvolvimento do código. Tornou-se desta forma possível criar um elevado número de regras para lidar com contextos bem definidos, complementando a estratégia proveniente da versão anterior, que era quase exclusivamente assente em regras de semelhança sobre o almanaque.

Funcionalmente, a segunda versão do SIEMÊS pode ser decomposta em duas camadas principais:

1. Camada de identificação de candidatos, usando pistas formais, como a presença de maiúsculas ou de números. Esta camada recorre a um banco de regras para a identificação de candidatos alfabéticos e um outro onde é feita em simultâneo a identificação e classificação semântica de entidades numéricas: datas, quantidades, numerário,

etc. Relativamente a estas entidades, a identificação e a classificação são feitas num mesmo passo já que não há grandes problemas de ambiguidade. Nesta fase, a segunda versão do SIEMÊS quase não difere da primeira, tirando o facto de todas as regras estarem codificadas externamente.

2. Camada de classificação para as entidades alfabéticas. Esta camada é composta por uma cadeia de classificação com cinco componentes, capazes de gerar hipóteses de classificação dos candidatos usando estratégias diferentes. Após esta cadeia, aplica-se o componente final de desambiguação, que tenta escolher de entre as várias hipóteses geradas qual a correcta tendo em conta informação adicional acerca do contexto. Este componente de desambiguação tenta também identificar a forma de menção da entidade.

Sobre a primeira camada não há nada de particularmente relevante a destacar, para além do facto de no SIEMÊS v2 ter sido possível criar um banco com várias dezenas de regras que identificam e classificam vários tipos de entidade numéricas. Como nota, e em comparação com o SIEMÊS v1, o desempenho do SIEMÊS v2 na classificação de entidades da categoria `TEMPO` subiu de F=0,59 para F=0,71 e das entidades da categoria `VALOR` subiu mais de 30 pontos na medida F, de F=0,43 para F=0,77.

Como referido anteriormente, a camada de classificação possui uma cadeia de geração de hipóteses com cinco componentes, que são invocados sequencialmente e recorrem a estratégias diferentes para a geração de hipóteses. Os componentes, e as respectivas estratégias de geração de hipóteses são, pela ordem de invocação:

1. Bloco de regras "simples" sobre o contexto (que se supõem de elevada precisão)
2. Bloco de pesquisa directa no REPENTINO
3. Bloco de emparelhamento de prefixo sobre o REPENTINO (2 opções)
4. Bloco de semelhança sobre o REPENTINO (2 heurísticas)
5. Bloco posterior de recurso

Na actual versão do SIEMÊS (v2), estes blocos são chamados sequencialmente, embora nos pareça que em futuras versões deve ser explorada a possibilidade de invocar os blocos em parelelo de forma a poder combinar as contribuições de todos os componentes. A fusão dos resultados para uma decisão de classificação final poderá ser feita usando um mecanismo de votação especializado por categorias, já que, como iremos ver, o desempenho dos componentes varia em função destas. Nas secções seguintes iremos explorar com mais detalhe cada um destes componentes.

14.2.1 Bloco de regras "simples"

Este componente é composto por um conjunto de regras manualmente codificadas que tenta explorar pistas contextuais muito explícitas. A composição das regras é feita de uma forma compacta recorrendo ao conhecimento de certas classes semânticas de palavras, nomeadamente ergónimos ou cargos, tipos de povoação (*cidade*, *vila*, *aldeia*,...), tipos de organizações, e outros grupos de palavras que são altamente relevantes no contexto de REM. Toda esta informação é mantida numa base exterior ao SIEMÊS para desenvolvimento autónomo. Um exemplo de uma regra pertencente a este bloco é:

```
{{ -1:@cargo =>
   meta(-1,CLASSE=SER); meta(-1,SUBCLASSE=CARGO);
   meta(CLASSE=SER); meta(SUBCLASSE=HUM);
   sai();
}}
```

Relembre-se que estas regras são invocadas já após a *fase de identificação*, e são disparadas para cada candidato identificado, pela que a regra anterior tem a seguinte leitura: «se o candidato identificado (posição 0) for precedido por uma palavra da lista `@cargo` (posição -1), então marca o referido elemento precedente com as meta-etiquetas `CLASSE=SER` e `SUBCLASSE=CARGO` e marca o candidato com as meta-etiquetas `CLASSE=SER` e `SUBCLASSE=HUM`».

Um possível resultado desta regra seria algo como:

```
O <EM CLASSE=SER SUBCLASSE=CARGO>imperador</EM>
<EM CLASSE=SER SUBCLASSE=HUM>Hirohito</EM> chegou.
```

já que o termo *imperador* se encontra catalogado com a etiqueta `CARGO`. Este bloco tem 23 regras, destinadas quase exclusivamente a classificar instâncias da classe `PESSOA`.

14.2.2 Bloco de pesquisa directa no REPENTINO

Este bloco tem um funcionamento muito simples, consistindo numa pesquisa sobre o almanaque REPENTINO através de um módulo Perl que armazena toda a informação do almanaque. Para um dado candidato, é verificado o número de entradas no REPENTINO que possuem a mesma representação lexical e é guardada a informação acerca das respectivas classes e subclasses, que passam a ser consideradas hipóteses de classificação.

14.2.3 Bloco de emparelhamento de prefixo sobre o REPENTINO

Este bloco é uma generalização do anterior e consiste numa tentativa de encontrar no REPENTINO as instâncias que possuam o mesmo conjunto de palavras iniciais (prefixo) que

o candidato. Pretende-se explorar heuristicamente a informação que se encontra no prefixo de um candidato, que em certos casos possui grande potencial discriminativo. A pesquisa é iniciada considerando inicialmente um certo número de palavras do candidato (as duas primeiras ou as quatros primeiras) e são pesquisadas as instâncias no REPENTINO que se iniciam pelas mesmas palavras. As instâncias obtidas do REPENTINO são agrupadas por categorias e quando uma dessas categorias inclui mais de 40% das referidas instâncias é gerada uma hipótese de classificação que consiste nessa categoria e nas suas subcategorias mais representadas nos exemplos encontrados.

Se o limite mínimo de 40% não for alcançado, então reduz-se uma palavra à pesquisa de prefixos (isto é considera-se apenas uma ou três palavras) e tenta-se um novo emparelhamento com entradas do REPENTINO. Este procedimento é repetido até a tentativa de emparelhamento incluir apenas uma palavra ou se atingir o limite de cobertura de 40%. Pode não ser gerada nenhuma hipótese, continuando o processo de pesquisa de hipóteses nos outros blocos.

14.2.4 Bloco de semelhança sobre o REPENTINO

Neste bloco foram implementadas duas funções heurísticas que tentam estabelecer semelhanças entre um determinado candidato e o conteúdo do REPENTINO, permitindo assim obter informação acerca do grau de pertença do candidato relativamente às categorias definidas no REPENTINO. Quanto mais semelhante for o candidato relativamente às instâncias incluídas numa determinada categoria e subcategoria do REPENTINO, mais elevado é considerado o seu grau de pertença a essa categoria e subcategoria, sendo gerada uma hipótese de classificação em conformidade.

Para este cálculo foram definidas duas heurísticas, Difuso1 e Difuso2. A primeira heurística, Difuso1, tenta determinar para cada palavra do candidato qual a sua frequência relativa em cada uma das categorias/subcategorias do REPENTINO e estimar um grau de pertença do candidato com base numa média ponderada desses valores. Por exemplo, suponhamos que se pretende obter pela heurística Difuso1 o grau de pertença do candidato C_j, composto pela sequência de palavras $p_1 p_2 ... p_n$, relativamente às categorias/subcategorias do REPENTINO. Para cada palavra p_i pertencente ao candidato questiona-se o REPENTINO para obter informação acerca das subcategorias para as quais existem instâncias com a palavra p_i. É assim obtida uma lista com elementos da forma (Subcategoria S_1, nº entidades em S_1 contento palavra p_i) para cada palavra do candidato C_j. Vamos admitir que estes valores são obtidos usando a função $REP(S_i, p_i)$, que nos poderia levar a obter, por exemplo, os seguintes valores para p_i="silva":

- REP(Ser::Humano , "silva") = 1031;
- REP(Organização::Comercial , "silva") = 96;

- REP(Local::Endereço Alargado ,"silva") = 42;

Podemos então definir a função $P_{Difuso1}$ que fornece uma medida do grau de "pertença" do candidato C_j à subclasse S_i do REPENTINO, como:

$$P_{Difuso1}(C_j, S_i) = \frac{1}{tam(C_j)} \sum_{n=1}^{tam(C_j)} \frac{REP(S_i, p_n)}{REP(S_i, *)} \tag{14.1}$$

Sendo $tam(C_j)$ o número de palavras do candidato C_j, retirando preposições e outras palavras sem conteúdo. Após o cálculo de $P_{Difuso1}$ para todas as subcategorias onde qualquer uma das palavras de C_j ocorrem, podemos obter uma lista ponderada de hipóteses de classificação do candidato.

A segunda heurística, Difuso2, tenta explorar a *especificidade* das palavras existentes no candidato C_j. Cada palavra do candidato contribui para a geração das hipóteses de classificação finais tanto mais quanto menor for o número de subcategorias do REPENTINO onde existam instâncias (independentemente do seu número) que incluem a palavra em causa. A contribuição que uma palavra do candidato fornece é assim pesada por um factor inversamente proporcional ao número de subcategorias em que a palavra "ocorre", sendo assim promovida a contribuição de palavras que só ocorrem num número muito reduzido de subcategorias do REPENTINO. Desta forma, se um candidato possuir uma palavra para a qual só existe no REPENTINO uma subcategoria onde se encontram instâncias que incluem essa palavra, isso é interpretado por esta heurística como uma forte pista de que o candidato pertence a essa subcategoria.

Seja $NSUB(p_i)$ a função que retorna o número de subcategorias do REPENTINO nas quais existem instâncias contendo a palavra p_i. Para cada uma das subcategorias S_i pode ser calculado um grau de pertença do candidato C_j através da seguinte formula:

$$P_{Difuso2}(C_j, S_i) = \frac{1}{tam(C_j)} \sum_{n=1}^{tam(C_j)} ESP(p_n, S_i) \tag{14.2}$$

com:

$ESP(p_n, S_i) = \frac{1}{NSUB(p_n)}$

se pelo menos uma instância de S_i possui a palavra p_n, ou

$ESP(p_n, S_i) = 0$

se nenhuma instância de S_i possui a palavra p_n.

Tal como na heurística Difuso1, obtém-se uma lista ponderada de hipóteses de classificação do candidato C_j, que poderão posteriormente ser desambiguadas.

Note-se, contudo, que em qualquer dos casos as heurísticas recorrem apenas à informação das palavras simples para a obtenção das possibilidades de classificação. Faria sentido

que as heurísticas entrassem em consideração com *n*-gramas mais longos, permitindo que fossem tidas em consideração unidades lexicais composta mais discriminativa que palavras simples. É possível imaginar um esquema iterativo piramidal que parta da utilização da totalidade do candidato a marcar para obter um primeiro conjunto de hipóteses, e que em subsequentes iterações entre em consideração com os n-gramas constituintes de tamanho imediatamente inferior para refinar as hipóteses obtidas, até se atingir a utilização das palavras simples (como agora é feito). Este mecanismo piramidal seria semelhante ao de algoritmos como por exemplo o BLEU (Papineni et al., 2001), utilizado na avaliação de sistemas de tradução automática, e o resultado final consistiria numa combinação ponderada das hipóteses obtidas em cada nível da pirâmide. As hipóteses geradas a partir de n-gramas maiores seriam ponderadas com mais importância do que aquelas obtidas a partir dos n-gramas mais pequenos (no limite, palavras simples).

Contudo, a forma de ponderação a usar carece de um estudo que ainda não tivemos oportunidade de fazer. Além disso, a carga computacional envolvida em tal cálculo poderá afectar severamente o desempenho do SIEMÊS, pelo que questões de eficiência computacional do processo também deverão ser consideradas.

14.2.5 Bloco posterior de recurso

Este bloco contém um conjunto de regras muito simples a usar no fim da cadeia de classificação, como último recurso, e que pretendem explorar algumas pistas contextuais muito genéricas. Embora aparentemente pouco precisas, estas regras podem ser suficientes para resolver mais alguns casos que não foram tratados pelas estratégias anteriores. Um exemplo de uma regra é aquela que permite marcar um candidato com a etiqueta AMC (Arte, Media, Comunicação) do REPENTINO, a qual corresponde a um objecto média como por exemplo um título de um filme ou livro, verificando apenas se o mesmo se encontra entre aspas:

```
-1:"1:"=> meta(CLASSE=AMC); sai();
```

14.3 A participação no Mini-HAREM

A participação do SIEMÊS no Mini-HAREM tinha dois objectivos principais. Em primeiro lugar, pretendia-se reconfirmar a validade da aproximação já usada na primeira versão e verificar se certos problemas na identificação e classificação de expressões numéricas poderiam ou não ser facilmente corrigidos. De facto, para além dos mecanismos de semelhança já usados anteriormente, o SIEMÊS permite nesta segunda versão a construção e utilização de bancos de regras externos ao programa que podem por isso ser editados independentemente com grande facilidade. Desta forma, o SIEMÊS foi preparado com várias dezenas

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	1º	43,0	19,8	0,271
ACONTECIMENTO	5º	20,7	26,8	0,233
COISA	4º	40,0	10,2	0,162
VALOR	7º	84,5	70,3	0,767
TEMPO	8º	85,1	61,0	0,710
LOCAL	7º	61,3	56,7	0,589
PESSOA	3º	59,8	57,5	0,586
ORGANIZACAO	3º	40,2	47.0	0,433
OBRA	2º	15,3	33,5	0,210
TOTAL	2º	53,02	51,4	0,522

Tabela 14.2: Resultados da avaliação global da classificação semântica combinada do melhor ensaio do SIEMÊS v2 no Mini-HAREM.

de regras destinadas exclusivamente ao processamento de expressões numéricas tentando assim resolver um dos mais notórios problemas da versão anterior. Esta facilidade na construção e aplicação de regras foi também aplicada no desenvolvimento do componente de regras de grande precisão, já apresentado anteriormente, embora infelizmente não tenham sido desenvolvidas regras num número tão grande como o desejado, essencialmente por limitações de tempo e indisponibilidade de recursos léxico-semânticos.

Em segundo lugar, pretendia-se realizar uma avaliação do sistema por componentes, para perceber exactamente qual a contribuição de cada um deles na resolução global do problema de REM e, dada a riqueza dos resultados de avaliação fornecidos pela organização, se a eficiência das estratégias varia com as categorias em análise. Colocam-se questões muito interessantes, tais como saber qual a dificuldade relativa na classificação de entidades diferentes e que tipos de recursos / estratégias é que poderão ser mais eficientes na classificação de uma dada categoria em particular.

Infelizmente, à data do Mini-HAREM, a segunda versão do SIEMÊS ainda não estava completa, em especial o componente de desambiguação, pelo que apesar da melhoria de desempenho para entidades numéricas já comentado anteriormente, os resultados globais do melhor ensaio do SIEMÊS no Mini-HAREM (Precisão = 53,0%; Abrangência = 51,4% e medida F = 0,522) foram ligeiramente piores que os resultados da primeira versão do SIEMÊS obtidos no HAREM. A título comparativo, apresentamos na Tabela 14.2 os resultados por categoria do melhor ensaio do SIEMÊS v2, directamente comparável com os resultados do SIEMÊS v1 apresentados na Tabela 14.1.

14.3.1 A decomposição da avaliação

No Mini-HAREM foram submetidos 9 ensaios (ver Tabela 14.3). Dois dos ensaios, *sms-total1* e *sms-total2*, fizeram uso de todos os componentes disponíveis, podendo ser considera-

Ensaio	*Smpl*	*Exct*	*Prfx*	*Dfs1*	*Dfs2*	*Pstr*
sms-simples	X					
sms-exacto		X				
sms-prefixo2			X(2)			
sms-prefixo4			X(4)			
sms-difuso1				X		
sms-difuso2					X	
sms-posterior						X
sms-total1	X	X	X	X		X
sms-total2	X	X	X		X	X

Tabela 14.3: A configuração dos nove ensaios enviados para avaliação.

dos duas configurações completas, embora distintas, do SIEMÊS. Os restantes sete ensaios consistiram em manter activo apenas um dos cinco componentes de geração de hipóteses descritos na secção anterior. Para dois dos componentes foram ainda experimentadas duas opções de funcionamento o que resulta nos referidos sete ensaios. As correspondências na Tabela 3 são:

1. *Smpl*: bloco regras "simples"activado.
2. *Exct*: Bloco de pesquisa directa no REPENTINO activado.
3. *Prfx*: Bloco de emparelhamento de prefixo sobre o REPENTINO activado. Foram testadas as duas opções disponíveis, isto é começar por tentar emparelhar 2 palavras ou 4 palavras.
4. *Dfs1*: Bloco de semelhança sobre o REPENTINO activado, usando a heurística Difuso1.
5. *Dfs2*: Bloco de semelhança sobre o REPENTINO activado, usando a heurística Difuso2.
6. *Pstr*: Bloco posterior de recurso activado.

Em todos os ensaios, sempre que não fosse possível chegar a uma hipótese de classificação (com um nível mínimo de confiança) era removida a marcação de identificação para que fosse possível testar e comparar mais convenientemente o desempenho na etapa de classificação, e não na etapa de identificação nos quais os ensaios não divergem. Desta forma, os dados de avaliação relevantes para o nosso estudo são aqueles que constam do cenário relativo previsto pela organização do HAREM, em particular aqueles que se referem à classificação semântica combinada. Todas as submissões incluiam a análise às EM

Ensaio	Precisão (%)	Abrangência (%)	Medida F
sms-total2	53,0	51,4	0,522
sms-total1	52,6	51,0	0,518
sms-prefixo4	57,2	46,1	0,511
sms-prefixo2	55,2	46,9	0,507
sms-difuso2	45,9	42,3	0,440
sms-exacto	66,0	33,0	0,440
sms-posterior	58,1	25,3	0,353
sms-difuso1	35,5	32,3	0,338
sms-simples	68,8	15,0	0,246

Tabela 14.4: O resultado global no Cenário Absoluto dos 9 ensaios.

"numéricas" (data, numerário...) o que em rigor não deveria ter sido feito, pois esta classificação mascara um pouco os resultados globais dos ensaios. Contudo, quando a comparação é feita por categorias este factor torna-se irrelevante. Em todo o caso, consideramos que as comparações são sempre indicativas das vantagens ou desvantagens relativas de cada um dos componentes e opções.

14.3.2 Resultados globais

Para melhor ilustrar o impacto das várias opções no desempenho global do sistema encontram-se na Tabela 14.4 os resultados no cenário absoluto dos 9 ensaios. Estes resultados correspondem à avaliação mais crua do sistema, em que se considera o desempenho do sistema na tentativa de marcação de todas as entidades existentes na Colecção Dourada. Como seria de esperar as duas configurações completas do sistema, *sms-total1* e *sms-total2*, obtiveram os melhores resultados mas há que destacar os desempenho muito próximos de certos ensaios parciais, como é o caso dos correspondentes à activação do componente de emparelhamento do prefixo, *sms-prefixo4* e *sms-prefixo2*, e os bons valores de precisão obtidos pelo ensaio *sms-exacto*, que recorre ao emparelhamento exacto sobre o REPENTINO, e pelo ensaio *sms-simples* que recorre a um (ainda) pequeno conjunto de regras sobre o contexto.

Para se poder compreender melhor as diferenças em termos de precisão entre os ensaios, são apresentados na Tabela 14.5 os resultados da classificação no cenário relativo, isto é apenas considerando as entidades correctamente identificadas.

Estes valores colocam no topo os dois paradigmas quase opostos de REM: a utilização de regras manualmente preparadas e a utilização directa dos almanaques. Por outro lado reforça-se a convicção que a informação contida nas primeiras palavras da entidade é de facto muito importante, já que os níveis de precisão foram também relativamente elevados. É interessante ver que os ensaios *sms-exact*, *sms-prefixo4* e *sms-prefixo2*, que correspondem a níveis crescentes de generalização na forma como se utiliza a informação de almanaque

Ensaio	Precisão (%)
sms-simples	77,2
sms-exacto	72,1
sms-prefixo4	64,9
sms-prefixo2	62,3
sms-total2	61,1
sms-posterior	62,4
sms-total1	60,7
sms-difuso2	53,0
sms-difuso1	41,0

Tabela 14.5: Valores de precisão no Cenário Relativo para os 9 ensaios

Categoria	Ensaio	Precisão (%)
ABSTRACCAO	*sms-exacto*	85,3
ACONTECIMENTO	*sms-exacto*	80,0
COISA	*sms-difuso2*	95,0
LOCAL	*sms-exacto*	95,3
PESSOA	*sms-posterior*	89,4
ORGANIZACAO	*sms-exacto*	91,6
OBRA	*sms-exacto*	88,7

Tabela 14.6: Os melhores ensaios para a classificação semântica por categorias no cenário relativo.

apresentam um desempenho consistentemente decrescente. Curiosamente, os ensaios *sms-difuso2* e *sms-difuso1* que correspondem à forma mais genérica de utilização do almanaque obtiveram os piores resultados, embora o ensaio *sms-difuso1* tenha tido um desempenho significativamente inferior ao *sms-difuso2*. Esta diferença reflecte-se directamente, embora mais suavemente, nos desempenhos relativos dos ensaios *sms-total1* e *sms-total2*.

14.3.3 Os melhores componentes por categoria

No sentido de perceber quais os componentes que poderão ser mais adequados para lidar com as diferentes categorias prevista no HAREM / Mini-HAREM, apresentamos na Tabela 14.6 os resultados dos melhores ensaios em cada categoria, no que diz respeito à precisão, no cenário relativo.

O dado que mais se destaca no que diz respeito à precisão no cenário relativo é a supremacia em 5 das 7 categorias do ensaio *sms-exacto*, que faz uso da pesquisa directa e booleana sobre o REPENTINO. Em particular, à excepção da categoria COISA, categoria cuja definição é complexa, e da categoria PESSOA, que o ensaio *sms-posterior* lida com grande precisão (embora com reduzidíssima abrangência), o resultado nas restantes categorias é indicativo da importância do uso dos almanaques no processo de REM, apesar da modesta abrangência global (mas não a mais baixa - ver Tabela 14.4) obtida no ensaio, que rondou

os 33%.

14.3.4 Alguns comentários

Os valores de precisão obtidos em torno dos 85% não devem ser ignorados e devemos questionar-nos acerca da melhor forma de aproveitar tais desempenhos no futuro do SIEMÊS.

Uma possibilidade será usar o SIEMÊS numa versão exclusivamente baseada no componente de emparelhamento exacto com o REPENTINO para marcar uma grande quantidade de texto. Este texto poderá ser usado posteriormente como base para inferência de novas regras de contexto, usando mecanismos semelhantes ao SnowBall (Agichtein e Gravano, 2000), DIPRE (Brin, 1998) ou AutoSlog-TS (Riloff, 1996), ou a aquisição de novas entradas para o léxico semântico, tal como realizado em (Pasca, 2004). De facto, o bloco de regras (que se encontrava activo no ensaio *sms-simples*), apesar de ter atingido o melhor desempenho em termos de precisão, possui um nível de abrangência muito reduzido que poderia ser aumentado com a inclusão de novas regras ou com a expansão do léxico semântico no qual algumas das regras estão ancoradas.

Um segundo ponto que convém explorar tem a ver com o próprio almanaque REPENTINO, que foi construído paralelamente à primeira versão do SIEMÊS sem no entanto ter sido alvo de um planeamento suficientemente independente do sistema. Com tal planeamento poderiam ter sido obtidos resultados melhores usando menos exemplos do que as actuais 450 mil instâncias que o REPENTINO possui. De facto, entre estas existe um grande desequilíbrio na sua distribuição pelas 11 categorias e 103 subcategorias do almanaque. Por exemplo, cerca de dois terços das instâncias do REPENTINO são nomes de pessoas, que na verdade poderão ser em grande parte dispensadas.

Além disso, o REPENTINO possui vários problemas típicos de outros recursos lexicais, como a presença de certas instâncias muito raras que poderão causar ambiguidades desnecessárias. Por exemplo, o REPENTINO armazena várias instâncias com o lexema *Paris*, entre as quais se encontra a referência a uma povoação, a um filme e a um produto consumível. Esta informação pode ser problemática se não for acompanhada de mais informação acerca do contexto que ajude à sua própria desambiguação. Não sendo isto possível na actual versão do REPENTINO, nos casos onde a desproporção entre a representatividade das entidades em causa é tão grande deveria manter-se no almanaque apenas a entrada correspondente à instância mais frequente (neste caso como *Povoação*). O ponto importante aqui é perceber quanto é que o SIEMÊS poderá ajudar neste processo de enriquecimento do REPENTINO com informação de contexto / frequência, ou possivelmente num processo de emagrecimento, isto é, de remoção de instâncias redundantes ou problemáticas. Tudo isto obrigará a pensar o REPENTINO como um sistema dinâmico, o que ainda não foi convenientemente equacionado mas deverá ser alvo de trabalho futuro.

Categoria	Ensaio	Precisão (%)	Abrangência (%)	Medida F
ABSTRACCAO	*sms-total2*	43,0	19,8	0,271
ACONTECIMENTO	*sms-prefixo2*	36,91	25,42	0,301
COISA	*sms-prefixo2*	41,05	10,43	0,166
LOCAL	*sms-total2*	61,29	56,69	0,589
PESSOA	*sms-total2*	59,78	57,49	0,586
ORGANIZACAO	*sms-total2*	40,25	46,95	0,433
OBRA	*sms-total1*	15,85	36,46	0,221

Tabela 14.7: Os melhores ensaios por categorias no Cenário Absoluto

É também muito interessante poder observar quais os melhores ensaios por categorias tendo em conta o desempenho no cenário absoluto. Os resultados encontram-se na Tabela 14.7 e, como seria de esperar, os ensaios completos, *sms-total1* e *sms-total2*, pelo seu elevado nível de abrangência, conseguem em quase todos os casos obter o nível de desempenho mais elevado em termos de medida F. O ensaio *sms-total2* obteve um desempenho superior nas categorias ABSTRACCAO, LOCAL, PESSOA e ORGANIZACAO. Quanto à categoria OBRA, o desempenho absoluto do ensaio *sms-total1* foi superior ao *sms-total2*.

Destacam-se também na Tabela 14.7 os bons resultados do ensaio *sms-prefixo2* nas categorias COISA e ACONTECIMENTO. Estes resultados sugerem que para estas categorias a informação contida nas duas primeira palavras é suficiente para as classificar, e que eventualmente o problema da definição de *menção* não é tão complexo. Os valores de abrangência são no entanto muito baixos, 10,4% para a categoria COISA e 25,4% para ACONTECIMENTO, o que sugere que uma expansão do REPENTINO nestas categorias poderá aumentar a abrangência do sistema.

14.4 Conclusões

A participação do SIEMÊS no HAREM e Mini-HAREM permitem tirar algumas conclusões acerca do problema de REM e, na nossa opinião, fornecem valiosas indicações acerca das opções em causa na construção de um sistema REM.

Em primeiro lugar parece-nos que fica confirmado que a utilização de almanaques não pode, pelo menos por enquanto, ser evitada, se se pretender desenvolver um sistema de REM de largo espectro. É evidente que com a construção de recursos linguísticos mais sofisticados se poderão desenvolver regras de análise de contexto (como as do bloco de regras do SIEMÊS) e de análise interna de candidatos que permitirão obter desempenhos superiores aos obtidos por estratégias exclusivamente assentes em almanaques. No entanto, o processo de construção desses recursos é demorado pelo que, enquanto estes não existirem, a utilização dos almanaques é indispensável. Por outro lado, e vendo a construção de um sistema de REM como um processo a médio prazo, os desempenhos obtidos

pelo SIEMÊS por utilização directa do almanaque, dado os razoáveis níveis de precisão num largo espectro de categorias, poderão servir de base a processo de inferência automática das referidas regras ou dos recursos linguísticos necessários.

A análise por categorias dos resultados do SIEMÊS e dos componentes que melhor lidaram com cada uma das categorias em causa sugere que o problema de REM não é homogéneo, e é necessário compreender melhor as características de cada uma das categorias, em termos de atributos lexicais, de contextos possíveis e de formas de menção admissíveis. Pela análise de componentes do SIEMÊS, e tendo em conta os desempenhos obtidos pelas diferentes estratégias em cada categoria, fica a ideia de que as categorias previstas no HAREM / Mini-HAREM possuem características radicalmente diferentes quanto aos itens anteriormente enunciados. Parece-nos que um re-estudo das categorias previstas no HAREM à luz das pistas obtidas a partir da avaliação de componentes do SIEMÊS poderá ser útil para a melhor definição do problema de REM.

Quanto ao desenvolvimento do SIEMÊS há três linhas de desenvolvimento que nos parecem essenciais para futuras versões do sistema:

1. melhoria das heurísticas de semelhanças sobre o REPENTINO. Uma possibilidade passaria pelo treino de um classificador automático de texto sobre o conteúdo do REPENTINO, de forma a inferir automaticamente regras de classificação que substituam as heurísticas manualmente desenvolvidas.

2. melhoria das regras de classificação de elevada precisão e o seu alargamento para outras categorias. Isto poderá necessitar de recursos léxico-semânticos mais desenvolvidos, pelo que deverá ser investido algum esforço paralelo na sua criação. Em ambos os casos deverão ser consideradas alternativas (semi)-automáticas.

3. re-organização dos vários componentes de geração de hipóteses numa estrutura que permita aproveitar as suas diferentes valências, algo que não aconteceu convenientemente na actual configuração do SIEMÊS. Uma estrutura paralela de funcionamento que envolva votação dos diferentes componentes poderá ser uma opção melhor do que a actual estrutura em cadeia (*pipeline*).

Como nota final, é importante destacar a enorme importância que a participação nas provas do HAREM / Mini-HAREM teve para a compreensão geral do problema de REM e para a definição das linhas futuras de desenvolvimento do SIEMÊS, pelo que esperamos que seja possível a realização de mais edições de exercícios de avaliação conjunta num futuro próximo. Termino, por isso, com o meu agradecimento à Linguateca pela organização deste esforço de avaliação.

Capítulo 15

Em busca da máxima precisão sem almanaques: O Stencil/NooJ no HAREM

Cristina Mota e Max Silberztein

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 15, p. 191–208, 2007.

A nossa participação no HAREM resulta de uma colaboração que é anterior à avaliação conjunta, enquadrando-se no âmbito do doutoramento da primeira autora. São dois os seus objectivos: (i) estudar as EM, bem como os contextos em que ocorrem, de um ponto de vista diacrónico; (ii) verificar se o desempenho de sistemas de REM é influenciado por variações temporais dos textos. Para tal, a primeira autora está a usar o CETEMPúblico, que abrange 8 anos (de 1991 a 1998), divididos em semestres.

A fim de alcançar o primeiro objectivo, foi necessário ultrapassar o obstáculo do corpus não se encontrar anotado com EM. Sendo inviável proceder à anotação manual do corpus, dada a sua extensão (180 milhões de palavras), a primeira autora optou por utilizar um ambiente de desenvolvimento para PLN que a auxiliasse nessa tarefa, o NooJ, concebido e implementado pelo segundo autor (Silberztein, 2004). Assim, desenhou e construiu uma série de recursos linguísticos (dicionários e gramáticas) para REM, designados Stencil, que são utilizados pelo sistema para produzir um texto anotado com EM. Estes recursos foram construídos manualmente e organizados de modo a serem aplicados numa cadeia de processamento que envolve três fases: (i) extracção de EM com base em regras precisas; (ii) extracção de EM com base em regras combinatórias que usam o almanaque extraído na primeira fase; (iii) anotação do texto por consulta ao almanaque extraído na segunda fase. Tanto a primeira como a segunda fase envolvem revisão manual do almanaque construído nessa fase.

O NooJ, ao ser utilizado com esses recursos, pode ser visto como um reconhecedor de EM, apesar de não ter sido desenvolvido exclusivamente com esse fim em vista. Alguns exemplos de ferramentas criadas com base em sistemas genéricos de desenvolvimento para PLN são: o ELLE (Marcelino, 2005), o AnELL (Mota e Moura, 2003) e o ExtracNP (Friburger, 2002), baseados no INTEX (Silberztein, 1993), o Glossanet (Fairon, 1999), baseado no Unitex (Paumier, 2002), e o MUSE (Maynard et al., 2003b), baseado no GATE (Cunningham et al., 2002). O ELLE (que também participou no HAREM), o ExtracNP e o MUSE são ferramentas de reconhecimento de EM.

A constituição do Stencil e a forma como os recursos que o compõem são usados pelo NooJ na análise de um texto foram condicionadas pelos objectivos do estudo anteriormente referido, sobretudo nos dois aspectos seguintes:

1. Pretende-se optimizar a anotação resultante quanto à precisão, ainda assim garantindo abrangência suficiente. Por outras palavras, é preferível anotar menos entidades, embora com maior certeza quanto à sua correcção em termos da delimitação e classificação, do que anotar mais entidades em detrimento da precisão nas anotações. Esta opção justifica-se pois só desta forma poderão os resultados da análise temporal ser precisos e representativos da totalidade das EM presentes no corpus.

Categoria	Tipo
PESSOA	INDIVIDUAL GRUPOIND CARGO GRUPOCARGO
ORGANIZACAO	OUTRO (HAREM) / INSTITUICAO (Mini-HAREM)
LOCAL	CORREIO ADMINISTRATIVO GEOGRAFICO VIRTUAL
TEMPO	DATA HORA PERIODO
VALOR	QUANTIDADE MOEDA

Tabela 15.1: Categorias e tipos considerados pelo Stencil/NooJ.

2. Não é desejável usar almanaques[1] de nomes próprios, a não ser os criados pelo próprio sistema a partir do texto que estiver a processar, porque isso poderia enviesar o resultado da anotação. Esse enviesamento surgiria, caso os nomes próprios contidos nos almanaques não estivessem igualmente distribuídos pelos vários semestres do corpus de estudo. Esta questão pode ser um problema uma vez que a anotação deve ser feita independentemente por semestre.

A realização do HAREM mostrou-se então a oportunidade de desenvolver e avaliar um etiquetador que produziria a anotação de EM segundo directivas acordadas entre um grupo de investigadores interessados na área em questão.

No entanto, quando começámos a trabalhar no etiquetador Stencil/NooJ tínhamos em vista o reconhecimento de entidades mencionadas no estilo do que foi proposto pelas conferências MUC (Chinchor, 1998b; Grishman e Sundheim, 1995), ou seja, reconhecimento de nomes próprios, em contexto, é certo, mas não o reconhecimento da função das EM no texto, que foi o que acabou por acontecer no HAREM. Por considerarmos a tarefa demasiado complexa, optámos por não readaptar completamente o nosso etiquetador às directivas propostas pela organização da avaliação. Essa complexidade dificultaria não só o trabalho de anotação manual a que teremos de proceder para termos uma colecção dourada por cada semestre do CETEMPúblico, como tornaria mais difícil a um sistema de anotação alcançar uma precisão e uma abrangência acima dos 90% e 40%, respectivamente, que nos permita fazer o estudo diacrónico com algum grau de fiabilidade (ou seja, as entidades que estudaremos cobrirão praticamente metade das entidades existentes no CETEMPúblico e estarão incorrectas em menos de um décimo dos casos).

Tendo em conta os nossos interesses de anotação, optámos por participar na tarefa de classificação em cinco categorias (ver Tabela 15.1): PESSOA, ORGANIZACAO, LOCAL, TEMPO e VALOR. Além disso, participámos na tarefa de classificação morfológica.

[1] Adoptámos aqui o conceito de almanaque (do inglês *gazetteer*) tal como definido por Mikheev et al. (1999): listas de nomes próprios de pessoas, locais, organizações e outra entidadades mencionadas. Note-se, no entanto, que outros autores consideram como almanaques apenas as listas constituídas por nomes próprios de locais (Grishman e Sundheim, 1995) e outros ainda, alargam a sua constituição a indicadores que possam ser úteis na classificação das EM, como por exemplo, os nomes de profissão (Sarmento et al., 2006; Bontcheva et al., 2002), ou distinguem dois tipos de almanaques: almanaques de entidades e almanaques-gatilho (*trigger gazetteers*) (Toral e Muñoz, 2006).

No que resta deste capítulo, começaremos por apresentar sucintamente o NooJ. Em seguida, descreveremos os Stencil e a cadeia de operações que é executada até obter o texto anotado. Na secção seguinte centrar-nos-emos em aspectos relacionados com a participação na avaliação: (i) mostraremos em que tarefas e categorias se focou a nossa participação, ilustrando ainda algumas das opções tomadas; (ii) contrastaremos a participação no HAREM e no Mini-HAREM, e (iii) faremos uma análise dos resultados alcançados, chamando a atenção para alguns problemas e dificuldades na anotação. Finalmente, apresentaremos algumas ideias para trabalho futuro.

15.1 O que é o NooJ?

O NooJ é um ambiente de desenvolvimento para PLN. À semelhança do INTEX (Silberztein, 1993), este ambiente permite, por um lado, construir descrições formais (dicionários e gramáticas) de ampla cobertura de linguagens naturais e, por outro, aplicar essas mesmas descrições a textos de grandes dimensões com grande eficiência. Essa eficiência advém do facto de ambos os sistemas manipularem descrições formais representadas por modelos computacionais de estados finitos: autómatos e transdutores, redes de transição recursivas (ou seja, transdutores que integram outros transdutores) e redes de transição recursivas com variáveis (as quais permitem replicar, condicionar e deslocar o seu conteúdo nas saídas dos transdutores).

Ambos os sistemas têm em comum diversas funcionalidades, não só porque ambos têm por objectivo fazer processamento de textos escritos, mas também por se enquadrarem no âmbito da metodologia e princípios estabelecidos por Gross (1975). Contudo, a arquitectura dos sistemas e as opções tomadas aquando do seu desenvolvimento são bastante diferentes, e o NooJ apresenta muitas funcionalidades novas.

O NooJ, cujo desenvolvimento se iniciou em 2002, foi inicialmente concebido para ser um INTEX aperfeiçoado. A primeira versão do sistema INTEX surgiu em 1992, tendo evoluído substancialmente nos 10 anos que se seguiram, sobretudo para dar resposta às necessidades dos utilizadores. Porém, a tecnologia do INTEX tornou-se obsoleta. Desenvolvido em C/C++, trata-se de um sistema monolingue, capaz de lidar com apenas um ficheiro de cada vez, sem suporte para diferentes formatos de texto, e sem suporte para XML.

Assim, em 2002, o NooJ foi desenhado de raiz, usando novas e entusiasmantes tecnologias: programação por componentes em C# para a plataforma .NET e manipulação de XML. Além disso, o seu novo motor linguístico tem a capacidade de processamento multilingue, em cerca de 100 formatos diferentes de ficheiros, incluindo documentos XML.

As funcionalidades do NooJ (das quais se destaca: análise de morfologia flexional e derivacional, elaboração de gramáticas locais, análise transformacional, indexação, localização e extracção de padrões morfo-sintácticos) estão disponíveis através de:

- um programa autónomo (*noojapply.exe*), que pode ser invocado directamente a partir de outros programas mais sofisticados;
- uma biblioteca dinâmica de .NET (*noojengine.dll*), que é constituída por classes e métodos de objectos públicos, os quais podem ser usados por qualquer aplicação .NET, implementada em qualquer linguagem de programação;
- uma aplicação integrada de janelas (*nooj.exe*), que permite executar uma série de funcionalidades num ambiente de janelas, incluindo a edição de gramáticas.

No HAREM utilizámos o ambiente de janelas.

15.1.1 Características dos recursos

Uma das principais vantagens do NooJ em relação ao INTEX foi ter unificado a formalização de palavras simples, palavras compostas e tabelas de léxico-gramática. Deste modo, os dicionários do NooJ permitem formalizar indistintamente palavras simples e compostas, e podem ser vistos como tabelas de léxico-gramática em que cada entrada corresponde à descrição de uma unidade lexical seguida das suas propriedades morfológicas, sintácticas e semânticas.

Estes dicionários assemelham-se aos dicionários DELAS-DELAC do INTEX, e, como tal, cada entrada é constituída por um lema seguido das suas propriedades, que no NooJ incluem, entre outras: categoria gramatical (*cat*), no máximo um código de flexão (*codflx*) introduzido por `+FLX`, zero ou mais códigos de derivação (*codderiv*) introduzidos por `+DRV` que poderão ser seguidos por um código de flexão para a forma derivada resultante (*codflex_deriv*), o qual é introduzido por ":", seguido de zero ou mais propriedades de natureza diversa; podem ainda ser especificadas, entre o lema e a categoria, variantes ortográficas ou terminológicas, tal como ilustra a seguinte entrada genérica:

```
lema{,variante}*,cat[+FLX=codflex]{+DRV=codderiv[:<codflex_deriv>]}*{+Prop}*
```

Embora estes dicionários possam ser flexionados automaticamente para efeitos de verificação e correcção (à semelhança do que acontecia no INTEX), para análise de texto não é necessário fazê-lo. Ou seja, a análise morfológica das palavras de um texto é feita directamente a partir da entrada de base (não flexionada) e do seu código de flexão no momento da aplicação do dicionário ao texto. Esta característica permite, por exemplo, a substituição de uma forma verbal que esteja no presente pela correspondente forma participial (o que poderá ser útil para transformar uma frase na forma activa na sua forma passiva).

Relativamente às gramáticas, cada gramática do NooJ corresponde a uma hierarquia de grafos constituída pelo grafo principal e todos os seus sub-grafos. Ou seja, ao contrário do que acontecia no INTEX, os sub-grafos chamados pelo grafo principal não são autónomos. Dado que, como veremos em seguida, as informações produzidas pelas gramáticas

são adicionadas incrementalmente a uma estrutura de anotação, isso torna possível a sua aplicação aos textos em cascata. Estas características permitem uma maior flexibilidade na criação, manutenção e aplicação de gramáticas.

Acrescente-se ainda, no que respeita às tabelas de léxico-gramática, que a sua unificação com os dicionários, bem como a possibilidade de processamento de análise morfológica durante a execução, permitem a sua utilização sem recorrer a meta-grafos. Este factor representa uma vantagem, em termos de descrição, já que os meta-grafos do INTEX tinham tendência a ficar demasiado grandes, e consequentemente difíceis de ler e alterar.

15.1.2 Processamento linguístico de textos

O motor linguístico do NooJ é baseado numa estrutura de anotação. Uma anotação é um par (posição, informação) que determina que uma certa posição no texto tem certas propriedades. Quando o NooJ processa um texto, produz um conjunto de anotações que são guardadas na Estrutura de Anotação do Texto (*Text Annotation Structure*, TAS) e estão sincronizadas com o mesmo. Portanto, a aplicação de dicionários ou de gramáticas ao texto nunca é destrutiva. Além disso, as gramáticas podem ser aplicadas em cascata, uma vez que vão sendo incrementalemente incluídas informações no TAS que podem ser usadas pelos recursos de níveis seguintes[2].

A partir das informações adicionadas ao TAS é possível criar um novo texto anotado em formato XML com essas informações integradas. Inversamente, também é possível abrir um documento XML no NooJ e integrar as anotações que nele existirem na estrutura de anotação do texto.

O sistema permite ainda a criação de colecções de textos. Esta funcionalidade torna possível aplicar a mesma operação (ou série de operações) a todos os textos de forma independente. Ou seja, a operação é aplicada a cada um dos textos individualmente, em vez de à união dos textos.

15.2 O que é o Stencil?

Antes do HAREM ser organizado, construímos uma série de grafos simples que faziam a anotação de nomes de pessoas, organizações e lugares no sistema INTEX. Essa classificação não tinha tipos, não tinha atributos morfológicos, mas estabelecia co-referência entre os nomes completos de organizações e as respectivas siglas ou acrónimos. Toda a informação necessária para fazer a anotação encontrava-se integrada nos grafos, não fazendo portanto uso de informações adicionais que estivessem formalizadas em dicionários, e também não tinha almanaques de nomes próprios a auxiliá-los na anotação.

[2] Saliente-se que a aplicação de gramáticas em cascata era possível no INTEX usando, por exemplo, a ferramenta CasSys (Friburger, 2002). No entanto, esta aplicação era destrutiva, pois em cada aplicação era criado um novo texto anotado.

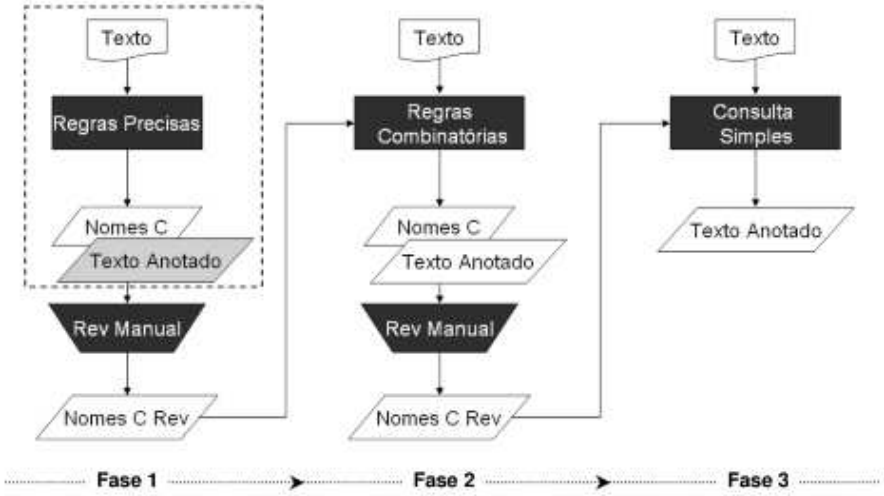


Figura 15.1: Arquitectura do etiquetador.

Uma vez que o NooJ apresentava várias vantagens em relação ao INTEX, tal como já referido na secção anterior, demos início à integração desses grafos no NooJ. Essa integração resultou praticamente numa reformulação dos grafos, pois tivemos de fazer várias modificações de acordo com as directivas do HAREM, nomeadamente: (i) prever novas categorias, (ii) fazer sub-categorização (iii) integrar classificação morfológica, e (iv) omitir a co-referência já que esta não foi contemplada na avaliação. Como o tempo era limitado, não nos aventurámos a fazer uma reestruturação completa dos grafos mais condizente com a filosofia do NooJ de construção de pequenas gramáticas para aplicação em cascata. A reformulação dos grafos também passou por uma simplificação do seu conteúdo, uma vez que muitas das informações que se encontravam explicitadas lexicalmente nos nós das gramáticas foram formalizadas em dicionários, e consequentemente essas informações lexicais passaram a ser categoriais. Por exemplo, em vez do nó conter os nomes de várias profissões (por exemplo, jornalista, linguista, pedreiro ou actor), passou a constar no nó apenas `<K+Profissão>`.

Este conjunto de recursos linguísticos, na forma de dicionários e gramáticas locais, que tem por fim fazer a anotação de EM, foi baptizado com o nome Stencil.

15.2.1 Organização dos recursos e forma de aplicação

Os recursos estão organizados de forma a serem aplicados em três fases distintas, como ilustrado na Figura 15.1.

Em cada uma das fases obtém-se não só um texto anotado, mas também uma lista

de nomes próprios classificados correspondentes às entidades que foram identificadas no texto. Uma vez que a última fase consiste apenas na anotação dos nomes que constarem na lista de nomes obtidos com o segundo passo, não é necessário extrair uma nova lista de nomes, pois seria idêntica à anterior. Dado que estamos interessados em fazer uma anotação optimizada quanto à precisão, as listas de nomes resultantes de cada um dos passos são revistas manualmente de modo a excluir potenciais fontes de erro nas fases seguintes. Por exemplo, se uma dada entidade for classificada com duas etiquetas distintas, em geral será eliminada da lista, pois quando a lista for reutilizada será criada uma falsa ambiguidade, que neste momento o Stencil "resolve" arbitrariamente; os nomes de pessoas ambíguos com nomes comuns também serão removidos, uma vez que a sua permanência não beneficia a análise ou poderá mesmo prejudicá-la (Baptista et al., 2006).

Através de experiências que fizemos com o CETEMPúblico, esta reutilização dos nomes encontrados no texto, sobretudo depois de revistos, permite o aumento da abrangência sem diminuir a precisão, mas apenas quando se trata da anotação de nomes próprios ao estilo das MUC. Isto porque, de uma forma geral, o nome de um local, por exemplo, não passa a ser o nome de uma organização dependendo do contexto, tal como acontece no HAREM. Este aumento de abrangência deve-se ao facto de as EM que foram encontradas pelas regras precisas poderem ocorrer noutros contextos que não foram previstos pelo primeiro conjunto de regras. Ao fazer a realimentação das EM irão ser encontradas essas ocorrências.

Dado que o nosso maior interesse era fazer a anotação do CETEMPúblico com vista à análise temporal das EM que nele ocorrem, não seria adequado o uso de almanaques de nomes próprios externos ao texto que está a ser analisado. Tal como justificado anteriormente, isso restringiria as EM encontradas, mesmo que em combinação com regras de reconhecimento com base em contexto. Embora possa parecer obscura essa opção, ela justifica-se porque, por um lado, não dispomos de recursos que estejam anotados em relação à época em que foram recolhidos e, por outro, queremos também estudar o aparecimento de novos nomes que não tenham sido previstos nos recursos.

15.2.2 Utilização de regras precisas

Na primeira fase, são aplicadas ao texto gramáticas locais que descrevem contextos muito restritivos que identificam e classificam EM com base em indícios internos e externos de acordo com a definição de McDonald (1996). Dado que não usámos almanaques, os indícios internos restringem apenas superficialmente a constituição interna do nome próprio dependendo da sua classificação. Por exemplo, o nome de pessoa é uma sequência de palavras em maiúsculas, eventualmente intercaladas por *de*, *do*, *das* e *dos*, não permitindo a ocorrência de *para*, como no caso das organizações; além disso, os indícios internos condicionam a primeira palavra do nome das organizações. Os indícios externos estabelecem

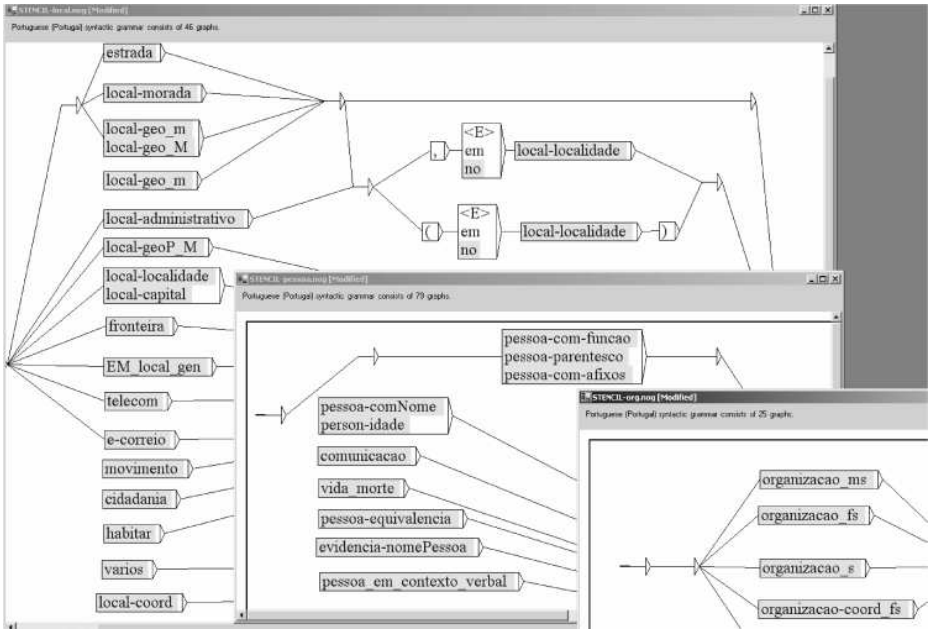


Figura 15.2: Primeiro nível das gramáticas aplicadas na primeira fase (apenas se mostra para ORGANIZACAO, PESSOA e LOCAL). O nome das sub-gramáticas encontra-se sombreado; alguns nós encontram-se desligados dos restantes por diminuirem a precisão.

contextos que com algum grau de certeza garantem a classificação da sequência em causa. Por exemplo, se uma sequência de palavras em maiúsculas que tem a constituição interna de nome de pessoa, for imediatamente precedida pelo nome de um cargo, então essa sequência será etiquetada como nome de pessoa.

As gramáticas utilizadas nesta fase estão organizadas de acordo com o tipo de entidade que reconhecem (ver Figura 15.2).

Nos casos em que era necessário fazer a classificação morfológica, os caminhos foram desdobrados de acordo com a flexão em género e número (i) do determinante que precede a sequência candidata a EM, ou (ii) do nome (no caso de ser um cargo, função, parentesco, etc.) que precede ou sucede a sequência, ou (iii) da primeira palavra que constitui a sequência, no caso dessa palavra ser um nome comum. Esse desdobramento permite atribuir a informação morfológica adequada à sequência que estiver a ser analisada como candidata a entidade mencionada. Este desdobramento deixou de ser necessário em versões do NooJ posteriores à realização do HAREM, pois passou a ser possível atribuir implicitamente atributos de elementos constituintes de uma sequência, a toda a sequência.

Adicionalmente, as gramáticas que classificam as entidades com a categoria PESSOA, tipos INDIVIDUAL e GRUPOIND, segmentam a sequência identificada como sendo nome de

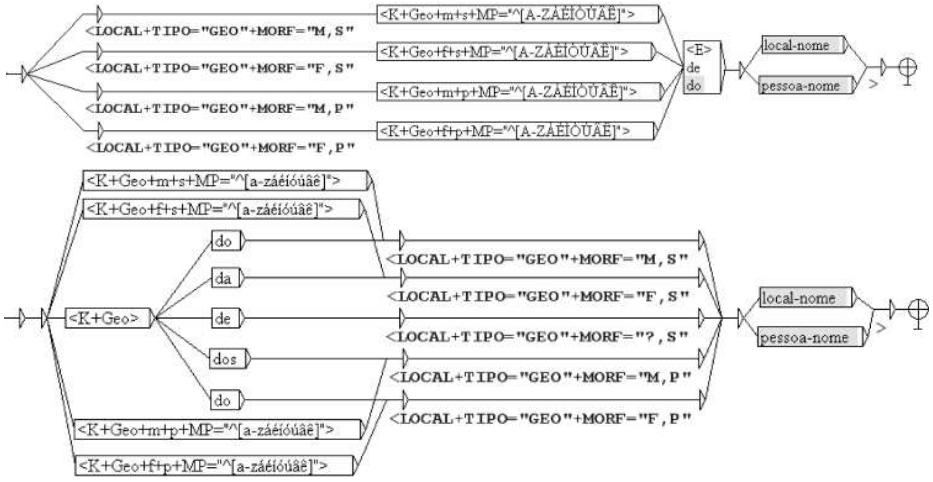


Figura 15.3: Detalhes da gramática de classificação de entidades de categoria LOCAL.

pessoa, associando a cada nome constituinte a etiqueta NOMEP. As regras combinatórias do passo seguinte usam os nomes classificados com essa etiqueta para descobrir novos nomes.

A funcionalidade do NooJ que permite combinar expressões regulares com sintaxe semelhante à do Perl com as palavras-chave do sistema permitiu simplificar e melhorar o processo de análise. Por exemplo, como se vê na Figura 15.3, dependendo de um marcador geográfico (K+Geo) começar por letra maiúscula (+MP= "[A-ZÁÉÍÓÚÂÊ]") ou minúscula (+MP="[a-záéíóúâê]") levará a que o mesmo seja ou não incluído dentro da anotação. A Figura 15.3 também ilustra o desdobramento das regras.

Estas gramáticas são aplicadas após a aplicação de um dicionário auxiliar que fornece as informações necessárias às gramáticas. Esse dicionário contém entradas nominais e adjectivais que se encontram sub-categorizadas de forma a poderem ser usadas na descrição tanto de indícios internos como externos. A constituição desse dicionário encontra-se descrita e exemplificada na Tabela 15.2.

De modo a flexionar estas formas, foram criados 51 paradigmas, dos quais 16 servem para flexionar compostos.

15.2.3 Utilização de regras combinatórias

A partir da anotação feita na primeira fase são geradas listas de nomes próprios classificados. Os que forem associados à etiqueta NOMEP são utilizados em regras combinatórias que identificam sequências de palavras em maiúsculas em que pelo menos um dos elementos tem essa classificação. Por exemplo, se a sequência *Jorge Sampaio* for identificada no primeiro passo como sendo PESSOA será integrada no almanaque do texto; além disso, tanto

Tipo	Formas canónicas	Formas flexionadas	Exemplo
Adjectivos patronímicos e gentílicos	530	2110	alentejano,A+FLX=Pato+Pátrio
Substantivos que designam profissões e funções	1581	6180	actor,K+FLX=Actor+Profissão
Substantivos que designam cargos	26	104	ministro,K+FLX=Cantor+Cargo
Parentescos	29	86	cunhado,K+FLX=Pato+Parentesco
Substantivos que introduzem instituições (mais 6 que introduzem departamentos)	81	162	escola,K+FLX=Mesa+Org+Cabeça
Substantivos que introduzem empresas	25	50	café,K+FLX=Carro+Emp+Cabeça
Substantivos geográficos, dos quais 8 são geopolíticos	39	78	comarca,K+FLX=Mesa+GeoP lago,K+FLX=Carro+Geo
TOTAL	2311	8770	

Tabela 15.2: Constituição do dicionário auxiliar.

Jorge como *Sampaio* serão igualmente adicionados a essa lista com a classificação `NOMEP`. Se neste passo, surge a sequência *Daniel Sampaio*, mesmo que esta não tenha sido identificada pelo passo anterior, então por conter *Sampaio* passará toda ela a ser identificada como `PESSOA` também. Por outro lado, mesmo que esses nomes ocorram isolados também serão classificados com essa categoria.

As restantes entidades que foram igualmente colocadas no almanaque do texto serão utilizadas directamente para identificar ocorrências dessas entidades em contextos que não foram previstos pelo primeiro passo.

Com excepção da abrangência dos nomes completos de organizações cuja classificação depende exclusivamente de indícios internos (e como tal, todas as ocorrências são encontradas no primeiro passo), a abrangência dos restantes tipos de nomes vai aumentar com a execução deste passo; a abrangência das organizações só aumenta ao nível das siglas e acrónimos que no primeiro passo apenas são identificadas quando estão no contexto do nome completo da organização.

15.2.4 Consulta simples dos dicionários de nomes próprios extraídos

Finalmente, na terceira fase, as listas de nomes classificados extraídos a partir da anotação feita no segundo passo, são aplicadas directamente ao texto sem recurso a novas regras de combinação nem de contexto. Ou seja, este passo consiste apenas numa consulta aos almanaques revistos (manualmente) de nomes próprios gerados a partir do próprio texto com as fases anteriores.

Esta fase tem sobretudo por objectivo aumentar a abrangência dos nomes de pessoa,

uma vez que com o passo anterior mais alguns novos nomes de PESSOA passaram a constar da lista de nomes próprios.

15.3 Participação no HAREM

O Stencil foi desenhado a pensar numa tarefa mais simples do que a que foi proposta pelo HAREM, ou seja, a classificação dos nomes das EM. Por esse motivo, como previamente referido, não fizemos algumas distinções estabelecidas nas directivas. Eis alguns exemplos em que não respeitámos as directivas:

- Independentemente de uma organização, como seja *Hotel Alfa*, estar a ser usada como locativo (*O congresso decorrerá no Hotel Alfa*) considerámo-la como ORGANIZACAO.
- Mesmo que um nome geográfico, como *Moçambique*, esteja na posição de um sujeito humano (*Moçambique fornecia muito café*) considerámo-lo como LOCAL.
- A uma data como *6 de Novembro* que em *No dia 6 de Novembro comemora-se...* devia ser considerada do tipo CICLICO, foi atribuído o tipo DATA.

Mesmo assim, adaptámos alguns aspectos de modo a que a participação não fosse completamente desadequada:

a) Alargámos a classificação às categorias TEMPO e VALOR;

b) Integrámos a atribuição de tipos;

c) Introduzimos a classificação morfológica;

d) Adaptámos algumas regras. Por exemplo, em alguns casos, os cargos, formas de tratamento e parentescos passaram a fazer parte das entidades classificadas como PESSOA, tipo INDIVIDUAL.

Dado que não se espera numa avaliação conjunta que exista intervenção humana durante o processo de anotação, a Colecção HAREM foi anotada apenas com base no primeiro passo descrito anteriormente. Poderíamos ter considerado automatizar o processo de revisão ou eliminá-lo, antes de fazer a realimentação. Porém, tendo em conta que no HAREM a classificação de uma entidade varia com a função que desempenha na frase, o processo de realimentação tal como está desenhado seria desastroso (já que esse processo assume exactamente que a função da entidade não varia). Naturalmente, que um processo de realimentação mais sofisticado poderia ajudar a resolver esta questão, como por exemplo o descrito por Mikheev et al. (1999), mas não tivemos tempo para o fazer. Além disso, as experiências que fizemos com a colecção dourada do HAREM, enquanto preparávamos o sistema para o Mini-HAREM, mostraram que o primeiro passo de extracção de EM não era

suficientemente preciso para fazer a reutilização, como se poderá confirmar pelos valores de precisão por categoria do resultado da experiência `stencil_1`, que foram ligeiramente superiores a 70% no caso da categoria `LOCAL` e entre 60% e 70% no caso das categorias `PESSOA` e `ORGANIZACAO` (ver secção 15.3.2).

15.3.1 HAREM vs. Mini-HAREM

Aquando do HAREM apenas a primeira fase do Stencil estava concluída. Existia apenas uma gramática principal organizada em sub-gramáticas de acordo com as entidades que classificava e foi construído o dicionário auxiliar. O NooJ não tinha sido sequer divulgado oficialmente, e muitas funcionalidades que existem agora, na altura ainda não estavam implementadas ou aperfeiçoadas[3]. Ao HAREM foram submetidos dois resultados, um oficial e outro não-oficial (ou seja, fora de prazo). Estes dois resultados distinguem-se pelo facto de ter sido corrigido um problema que nada tinha a ver com a análise das EM: na versão oficial, as anotações adicionadas ao TAS com base em contexto (por exemplo, indícios externos) não foram consideradas aquando da criação do texto anotado. Por exemplo, se no texto existisse a sequência *a irmã de Maria*, seria adicionada ao TAS a informação de que *Maria* tinha a categoria `PESSOA:INDIVIDUAL`; no entanto, essa informação não seria adicionada ao ficheiro anotado final.

No Mini-HAREM usámos a versão 1.21/b0322 do NooJ e as três fases do Stencil já estavam concluídas. Todavia, tal como anteriormente referido, a Colecção HAREM foi anotada apenas usando o primeiro passo (o qual corresponde à zona destacada com o rectângulo tracejado na Figura 15.1). Tendo em vista a aplicação em cascata, começámos a reestruturar a gramática que usámos no HAREM, dividindo-a em quatro gramáticas de acordo com as categorias: `PESSOA`, `ORGANIZACAO`, `LOCAL` e outra que reunia `TEMPO` e `VALOR`. Além disso, corrigimos alguns erros que as gramáticas tinham, restringimos os contextos descritos e introduzimos algumas regras novas. Com o objectivo de observar a diferença de desempenho com e sem almanaques submetemos, além do resultado anterior (que designaremos por `stencil_1`), mais três resultados:

- `stencil_pol`: obtido utilizando as gramáticas do passo 1 combinadas com a consulta simples de almanaques de nomes próprios extraídos do CETEMPúblico (extractos da secção de Política dos semestres 91a, 91b e 98b) usando o primeiro passo do Stencil com revisão. Este almanaque contém 14314 nomes de locais, 31764 nomes de pessoas, e 28510 nomes de organizações, num total de 75588 nomes próprios. Por lapso, os nomes de pessoa incluídos no dicionário não estavam a ser reconhecidos (por esse motivo, nos resultados seguintes mostra-se e comenta-se apenas o resultado corrigido, `stencil_polcor`);

[3] A primeira versão pública do NooJ (1.10) foi lançada em Março de 2005.

	Precisão	Abrangência	Medida F	Lugar
Identificação (cenário total)	78,25	58,83	0,6716	8º
Identificação (cenário selectivo)	64,09	63,17	0,6363	9º
Classificação combinada (cenário selectivo absoluto)	40,85	39,63	0,4023	9º

Tabela 15.3: Resumo das pontuações obtidas com o resultado não oficial no HAREM.

- `stencil_polcor`: obtido utilizando as gramáticas do passo 1 combinadas com o almanaque do passo anterior, com o reconhecimento de nomes de pessoas presentes no almanaque corrigido;
- `stencil_dic`: obtido utilizando as gramáticas do passo 1 e 2 em que o almanaque usado é o Npro (versão 5 sem nomes próprios ambíguos com nomes comuns) que contém 3544 nomes simples de pessoas classificados quanto a género e número, e quanto a serem nome de baptismo ou apelido (**?**).

15.3.2 Resultados

Relativamente à participação no HAREM, as pontuações obtidas ficaram muito aquém das expectativas, correspondendo a medida F do resultado oficial a cerca de metade do valor alcançado pelo resultado não oficial (por exemplo, no cenário total e absoluto a medida F foi 0,2073 e 0,4073, respectivamente). Essa diferença deveu-se ao facto de algumas anotações terem sido adicionadas ao TAS, sem terem sido integradas posteriormente no texto anotado oficial. Por esse motivo, não vamos sequer analisar esse resultado em mais detalhe, focando a análise de resultados no HAREM apenas nas classificações obtidas com o resultado não oficial, que acabou por não ser satisfatório devido a uma falha na gramática de reconhecimento. Por lapso, um dos caminhos da gramática que identifica as entidades do tipo `LOCAL` permaneceu demasiado genérico, o que levou a que boa parte das entidades do tipo `PESSOA` e `ORGANIZACAO`, bem como outras entidades que não pretendíamos identificar, fossem identificadas incorrectamente como `LOCAL` no resultado não oficial. Essa falha é sobretudo visível comparando a pontuação da identificação no cenário total com as pontuações da identificação no cenário selectivo e da classificação combinada no cenário selectivo (cf. Tabela 15.3). De notar que, corrigindo este erro, a medida F na classificação combinada seria inferior (23%). No entanto, observar-se-ia uma melhoria significativa em termos de precisão (66%). Como optámos por optimizar a precisão, essa correcção foi tida em conta no Mini-HAREM.

Saliente-se, no entanto, que o Stencil/NooJ obteve as melhores pontuações na identificação e classificação da categoria `TEMPO`, tendo alcançado a segunda melhor medida F e a melhor abrangência tanto na identificação como na classificação da categoria `VALOR` (ver Tabela 15.4, nos cenários total no caso da identificação, e total absoluto no caso da classificação combinada).

	Precisão	Abrangência	Medida F	Lugar
Identificação (cenário total) de TEMPO	85,74	76,65	0,8094	1º
Class. combinada (cenário total absoluto) de TEMPO	83,24	74,61	0,7869	1º
Identificação (cenário total) de VALOR	52,88	86,44	0,6562	2º
Class. combinada (cenário total absoluto) de VALOR	53.63	87,78	0,6659	2º

Tabela 15.4: Resumo das pontuações obtidas com o resultado não oficial no HAREM nas categorias TEMPO e VALOR.

Como se pode ver na Figura 15.4, o desempenho do Stencil/NooJ melhorou do HAREM (*stencil_no* – não oficial) para o Mini-HAREM (stencil_1, stencil_polcor e stencil_dic), em consequência de um aumento significativo da precisão.

Fazendo a análise por categoria (Figura 15.5), todas melhoraram excepto VALOR[4] que piorou em termos de medida F por ter havido uma diminuição da abrangência em troca de uma aumento de precisão. Também é possível observar que a categoria TEMPO melhorou ligeiramente a medida F como reflexo de um aumento da precisão, porém o sistema não conseguiu manter a melhor classificação nesta categoria, passando para terceiro lugar. Naturalmente, estas duas categorias não sofrem alterações nas experiências stencil_1, stencil_polcor e stencil_dic, uma vez que não dependem de almanaques. Na categoria LOCAL, em relação ao HAREM, houve uma descida da medida F, com as experiências stencil_1 e stencil_dic, como consequência da diminuição na abrangência compensada por um aumento significativo da precisão; com a experiência stencil_polcor, a medida F aumenta porque com base nos almanaques, que incluem nomes de locais, foi possível aumentar a abrangência sem prejudicar a precisão da experiência stencil_1. Com a categoria PESSOA, pelo contrário, a utilização de almanaques de nomes próprios de pessoas (quer simples, como na experiência stencil_dic, quer simples e compostos, como na experiência stencil_polcor) embora faça aumentar a abrangência, penaliza a precisão. No que respeita à categoria ORGANIZACAO, verifica-se um aumento mais significativo da medida F na experiência stencil_polcor, devido a um ligeiro aumento da abrangência, não tendo a precisão praticamente variado; esse aumento resulta sobretudo do reconhecimento de siglas que fazem parte do almanaque. O facto de não serem encontradas novas organizações para além das siglas deve-se ao facto das EM que estão no almanaque terem sido extraídas do CETEMPúblico com base em regras que dependem essencialmente dos mesmos indícios internos que estão a ser usados no reconhecimento de EM desse tipo na colecção HAREM. De acordo com as experiências de Wakao et al. (1996) esta categoria tem a beneficiar com o uso de indícios externos, nomeadamente porque muitas organizações são nomes de empresas, os quais não contêm em geral indícios internos bem definidos.

[4] No resultado stencil_1 a categoria VALOR, embora tenha sido adicionada ao TAS não foi exportada acidentalmente para o texto anotado final. Caso essas anotações tivessem sido exportadas, obter-se-ia para a categoria VALOR na classificação combinada uma precisão de 93,82%, uma abrangência de 37,18% e uma medida F de 53,26%. Estes valores são naturalmente semelhantes aos obtidos nas restantes experiências do Mini-HAREM.

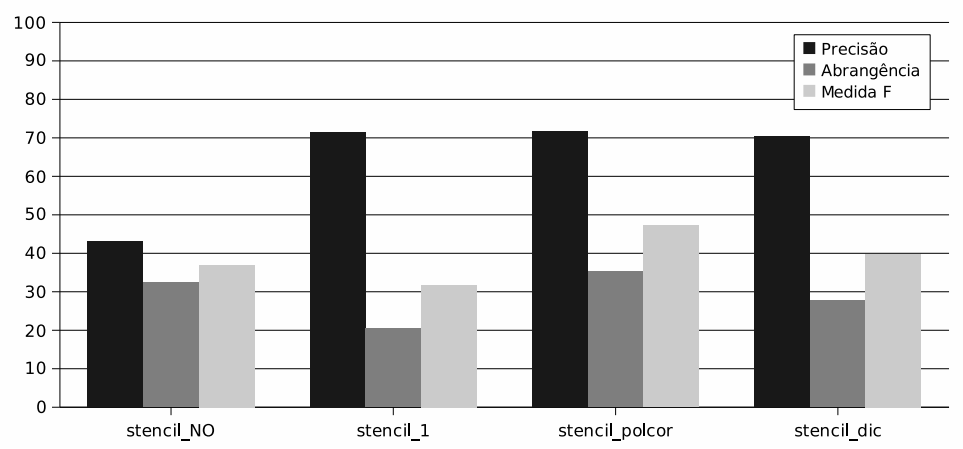


Figura 15.4: Classificação combinada no cenário total absoluto.

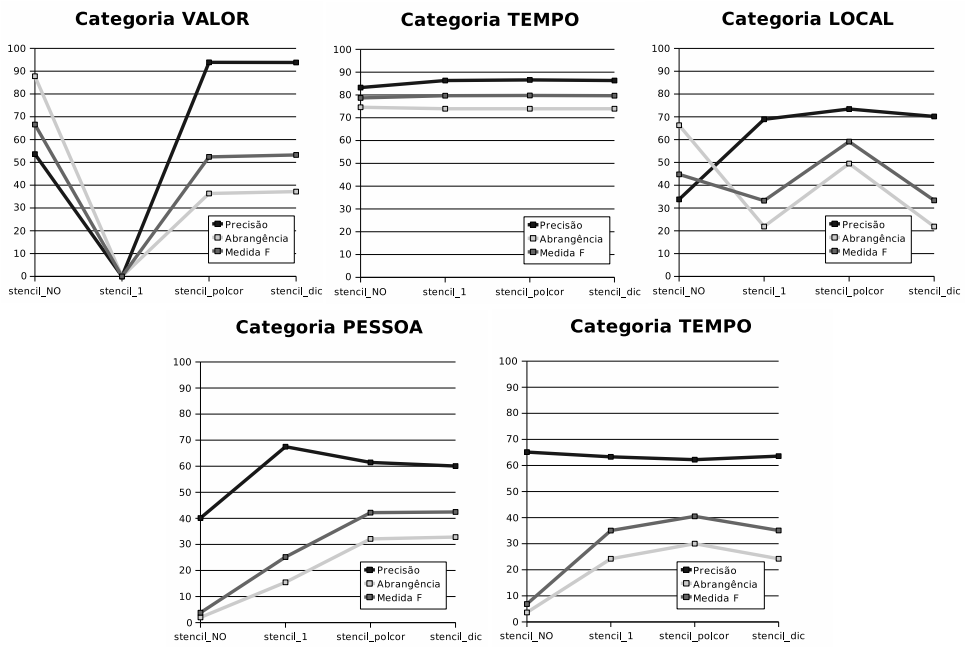


Figura 15.5: Classificação combinada por categoria no cenário total absoluto

Saliente-se ainda que os resultados obtidos para PESSOA, ORGANIZACAO e LOCAL, apesar de significativamente piores do que os de Mikheev et al. (1999), não são de espantar e sugerem a mesma conclusão: o reconhecimento da categoria LOCAL não consegue tirar partido tão facilmente do contexto e por isso o uso de almanaques ajuda, sobretudo, a melhorar o reconhecimento de entidades deste tipo.

Comparativamente com os outros sistemas participantes, apesar de não termos uma medida F tão boa devido à falta de abrangência, conseguimos mesmo assim estar entre os sistemas com melhor precisão.

No que diz respeito à classificação morfológica, acabámos por ser o único sistema a submeter resultados ao Mini-HAREM que a integrassem. Todavia, esses resultados não foram positivos. Para além da falta de abrangência (que não foi superior a 15% no melhor cenário total absoluto e mesmo no cenário total selectivo não ultrapassou os 20%), sobretudo nos resultados que foram obtidos com auxílio do almanaque do CETEMPúblico (stencil_polcor), a precisão foi baixa (61% no melhor caso no cenário total absoluto), sendo, no entanto, ligeiramente melhor em termos de número (no melhor caso, 35% de medida F no cenário total absoluto) do que em género (25% de medida F no cenário total absoluto, no melhor caso). Mesmo assim, tendo apenas em conta as entidades que são bem identificadas, os resultados são bem melhores (a medida F, passa de 25% no cenário total absoluto para 58% no cenário total relativo).

15.3.3 Problemas e dificuldades

Apesar de estarmos à espera de uma abrangência baixa, esta poderia ter sido mais alta se alguns pequenos lapsos na descrição das regras não tivessem ocorrido. Por exemplo, a regra que atribuía a categoria PESSOA a uma sequência de maiúsculas que ocorre após um cargo iniciado por letra minúscula tinha uma pequena falha que impediu a anotação das entidades neste contexto. Na experiência stencil_1, por exemplo, a correcção deste pequeno erro faria aumentar a abrangência de 15,46% para 16,89% e a precisão de 67,48% para 69,03% na classificação combinada da categoria PESSOA. Por outro lado, regras que em termos de precisão pudessem ser arriscadas por envolverem algum grau de ambiguidade não foram previstas. Por exemplo, se entre o nome de um cargo e uma sequência de maiúsculas existir a preposição *de* eventualmente contraída com um artigo definido, então é possível que essa sequência seja uma ORGANIZACAO (*o presidente da Sonae*); no entanto, também pode ser um LOCAL (*o presidente da China*); note-se, porém, que segundo as directivas do HAREM o segundo caso deve também ser anotado como ORGANIZACAO, mas terão tipos diferentes: EMPRESA no primeiro caso e ADMINISTRACAO no segundo.

O facto de termos dividido a gramática que tínhamos inicialmente em quatro gramáticas também trouxe algumas dificuldades. Por exemplo, com uma única gramática dada a sequência *o professor Ribeiro da Silva* que permite fazer a análise de *Ribeiro da Silva* como

PESSOA (por ocorrer a seguir a *professor*) bem como de LOCAL (por conter *ribeiro*), apenas a primeira anotação como PESSOA vai ser adicionada ao TAS por fazer parte de um caminho mais longo que tem precedência sobre análises mais curtas. Pelo contrário, usando as gramáticas separadas ambas as anotações são adicionadas ao TAS, o que leva a que no momento da geração do texto anotado o NooJ opte arbitrariamente por uma delas. Chamamos a atenção para o facto de neste momento já poderem ser geradas as duas anotações, o que, seja como for, não é a solução que pretendemos pois trata-se de uma falsa ambiguidade.

15.4 Comentários finais

Apesar de não termos seguido à risca as directivas da avaliação conjunta e termos acabado por concorrer com um sistema preparado para uma tarefa mais simples e com menos categorias, consideramos a participação positiva. Em particular, conseguimos uma precisão equiparável à do melhor sistema no Mini-HAREM (acima de 70%, enquanto o melhor sistema teve 73,55%), e por vezes ligeiramente melhor, apesar de ter tido uma medida F que variou entre 20% e 47%, quando o melhor sistema obteve quase 59%, no cenário total absoluto.

Contamos, numa futura edição do HAREM, caso se mantenham os objectivos de anotação da função das entidades, ser mais fiéis às directivas, mesmo que isso nos obrigue a manter dois sistemas diferentes: um para fins de anotação do CETEMPúblico com nomes próprios no âmbito da tese da primeira autora, e outro com o objectivo de competir conjuntamente na avaliação.

Mais do que a questão de quão bons foram os resultados na avaliação, interessa-nos saber quão melhores é que eles se tornarão no futuro. Para isso os programas avaliadores criados pela organização do HAREM (capítulo 19) são um instrumento fundamental para poder ir desenvolvendo e testando o sistema.

Agradecimentos

Os autores estão gratos ao grupo *Text Analysis and Language Engineering* do centro de investigação da IBM, T. J. Watson Research Center, por lhes terem dado a oportunidade de em 2001 trabalharem em conjunto em REM, o que serviu, em parte, de fonte inspiradora para o trabalho aqui apresentado. Os autores estão igualmente gratos ao Nuno Seco pelo apoio dado na utilização dos programas avaliadores, bem como ao Nuno Mamede, à Diana Santos, ao Nuno Cardoso, aos autores do CaGE, ao Luís Costa e ao Jorge Baptista pelas sugestões que nos deram para melhorar a versão final deste capítulo.

O trabalho da primeira autora foi financiado pela Fundação para a Ciência e a Tecnologia através da bolsa de doutoramento com a referência SFRH/BD/3237/2000.

Parte III

Capítulo 16

Directivas para a identificação e classificação semântica na colecção dourada do HAREM

Nuno Cardoso e Diana Santos

Este capítulo foi previamente publicado como Relatório Técnico DI/FCUL TR–06–18, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 16, p. 211–238, 2007.

Neste documento, apresentamos as directivas usadas na etiquetagem da colecção dourada da primeira edição do HAREM e do Mini-HAREM, e, consequentemente, qual o comportamento esperado pelos sistemas que nele participem.

Começamos por descrever o formato do que consideramos um texto anotado com entidades mencionadas (EM), e qual a definição operacional destas. Depois, para cada categoria, explicamos o significado atribuído e detalhamos a sua subcategorização.

No capítulo 17 será indicada a metodologia seguida para a anotação morfológica da colecção dourada.

16.1 Regras gerais de etiquetagem

Cada EM é rotulada por uma etiqueta de abertura e uma etiqueta de fecho, cujo formato é semelhante ao das etiquetas usadas em XML. A etiqueta de abertura contém a categoria atribuída, e possui atributos como o tipo ou a classificação morfológica. Na etiqueta de fecho, coloca-se a categoria usada na etiqueta de abertura. Um exemplo de uma EM etiquetada é:

```
os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Beatles</PESSOA>.
```

Os nomes das categorias e dos tipos não devem incluir caracteres com acentos e/ou cedilhas, e devem estar em maiúsculas. Ou seja, deverá ser usado `<ORGANIZACAO>` em vez de `<ORGANIZAÇÃO>`.

Os valores dos atributos `TIPO` e `MORF` devem ser rodeados por aspas.

Não deve haver nenhum espaço imediatamente a seguir à etiqueta de abertura e antes da etiqueta de fecho.

```
Certo:  O <PESSOA TIPO="INDIVIDUAL">João</PESSOA> é um professor.
Errado: O<PESSOA TIPO="INDIVIDUAL"> João</PESSOA> é um professor.
Errado: O <PESSOA TIPO="INDIVIDUAL">João </PESSOA>é um professor.
```

Se a EM contém espaços, esses devem manter-se inalterados.

```
Certo:  O <PESSOA TIPO="INDIVIDUAL">João Mendes</PESSOA> é um professor.
Errado: O <PESSOA TIPO="INDIVIDUAL">JoãoMendes</PESSOA> é um professor.
```

As aspas, parênteses, pelicas ou travessões não são para incluir na etiqueta, se englobarem a EM como um todo (ver caso 1). No entanto, são para incluir, caso apenas se apliquem a partes da EM (caso 2) ou façam parte integrante da mesma.

Caso 1

```
Certo:  A ''<OBRA TIPO="ARTE">Mona Lisa</OBRA>''
Errado: A <OBRA TIPO="ARTE">``Mona Lisa''</OBRA>
```

Caso 2

```
Certo: O <PESSOA TIPO="INDIVIDUAL">Mike ''Iron'' Tyson</PESSOA>
Certo: <PESSOA TIPO="INDIVIDUAL">John (Jack) Reagan</PESSOA>
Certo: Os resultados foram semelhantes aos produzidos por Diana Santos
       e colegas <OBRA TIPO="PUBLICACAO">(Santos et al, 2005)</OBRA>.
```

16.1.1 Recursividade das etiquetas

Não é permitido etiquetas dentro de etiquetas, como nos exemplos (errados) seguintes:

```
Errado: <PESSOA TIPO="GRUPO"><ORGANIZACAO TIPO="SUB">Bombeiros
        </ORGANIZACAO></PESSOA>
Errado: <ORGANIZACAO TIPO="INSTITUICAO">Departamento de <ABSTRACCAO
        TIPO="DISCIPLINA">Informática</ABSTRACCAO> do IST</ORGANIZACAO>
```

16.1.2 Vagueza na classificação semântica

No caso de haver dúvidas entre várias categorias ou tipos, deve utilizar-se o operador "|". Por exemplo, em *Ajudem os Bombeiros*, se se considerar que não existe razão para preferir uma das duas seguintes classificações para *Bombeiros*, nomeadamente entre `<PESSOA TIPO="GRUPO">` e `<ORGANIZACAO TIPO="INSTITUICAO">`, devem-se colocar ambas:

```
Certo: Ajudem os <PESSOA|ORGANIZACAO TIPO="GRUPO|INSTITUICAO">
       Bombeiros</PESSOA|ORGANIZACAO>!
```

Podem ser especificados mais do que uma categoria ou tipo, ou seja, `<A|B|C|...>`.

Caso a dúvida seja entre tipos, deve-se repetir a categoria. Por exemplo, em caso de dúvida sobre qual o tipo de organização (`EMPRESA` ou `INSTITUICAO`?) na frase *O ISR trata dessa papelada*, deve-se repetir a categoria `ORGANIZACAO` tantas vezes quantos os tipos indicados:

```
Certo: O <ORGANIZACAO|ORGANIZACAO TIPO="EMPRESA|INSTITUICAO">ISR
       </ORGANIZACAO|ORGANIZACAO> trata dessa papelada.
```

16.1.3 Vagueza na identificação

Se houver dúvidas (ou análises alternativas) de qual a identificação da(s) EM(s) que deverá ser considerada correcta, as várias alternativas são marcadas entre as etiquetas `<ALT>` e `</ALT>`, que delimitam e juntam as várias alternativas, que são separadas pelo caracter '|'. O exemplo abaixo mostra a etiquetagem a usar, quando não se consegue decidir por uma única identificação:

```
O <ALT><PESSOA TIPO="GRUPOMEMBRO">Governo de Cavaco Silva</PESSOA>
| Governo de <PESSOA TIPO="INDIVIDUAL">Cavaco Silva</PESSOA></ALT>
```

16.1.4 Critérios de identificação de uma EM

Uma EM deve conter pelo menos uma letra em maiúsculas, e/ou algarismos.

```
Certo:  <TEMPO TIPO="DATA">Agosto</TEMPO>
Errado: <TEMPO TIPO="DATA">ontem de manhã</TEMPO>
```

A única excepção a esta regra abrange os nomes dos meses, que devem ser considerados EM, ou parte de EM, mesmo se grafados com minúscula. Esta excepção deve-se ao facto de haver grafia maiúscula em Portugal e minúscula no Brasil nesse caso.

```
Certo: <TEMPO TIPO="DATA">agosto de 2001</TEMPO>
```

Existe também um conjunto de palavras relativas a certos domínios que também são excepções a esta regra, e que são as seguintes:

categoria `PESSOA` : *senhor, senhora, doutor, doutora, padre, cónego, deputado, chanceler, lorde, subprocurador-geral, presidente, rei, rainha, miss, major, comandante, capitão, brigadeiro, seu, tio, irmã, irmão, mana, mano, prima, primo, avô, avó, pai, mãe*

categoria `TEMPO` : *janeiro, fevereiro, março, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro, século, anos*

categoria `LOCAL` (tipo `ALARGADO`) : *número, n^o, sala,* abreviaturas de nomes de meses ligados por barra (para indicar o volume de uma revista, por exemplo *jan./dez.*)

categoria `ABSTRACCAO` (tipo `ESTADO`) : *doença, mal, sindroma, estado*

categoria `ABSTRACCAO` (tipo `NOME`) : Todos os casos descritos para a categoria `PESSOA`.

Se uma determinada EM, etiquetada como tal, aparecer depois sem maiúsculas no mesmo texto ou noutro, não deve ser outra vez etiquetada, ou seja, uma EM tem de conter obrigatoriamente pelo menos uma letra maiúscula e/ou algarismos.

No entanto, o inverso não é verdade, isto é, uma palavra com pelo menos uma letra maiúscula ou um número pode não ser uma EM. Um caso clássico são as palavras que iniciam as frases, mas também há que considerar o uso excessivo de maiúsculas em certos géneros de textos, como a *web*, onde casos como *Contactos*, *História*, *Página Inicial*, *Voltar*, *Menu*, *E-mail*, entre outros, não devem ser por regra identificados como EM.

Aplicando o mesmo raciocínio, as frases totalmente escritas em maiúsculas (como acontece em títulos de destaque) deverão ser analisadas cuidadosamente, e só deverão conter

etiquetas as EM claras. Por exemplo, se uma linha rezar *CLIQUE AQUI PARA VER A EDUCAÇÃO EM 1993*, *EDUCAÇÃO* não deve ser considerada uma EM, uma vez que, naquele contexto, a palavra não deveria conter nenhuma maiúscula. No entanto, o ano deve ser marcado como TEMPO, de tipo DATA ou PERIODO.

Outro exemplo: *ABALO EM LISBOA SEM VÍTIMAS*. Neste caso, consideramos correcto marcar LISBOA como EM, visto que assumimos que manteria a maiúscula se a frase não fosse exclusivamente grafada em maiúsculas. Note-se, de qualquer maneira, que estes casos caem um pouco fora do âmbito do HAREM, em que se utilizou um critério predominantemente gráfico, baseado nas convenções da língua escrita.

Palavras que foram incorrectamente grafadas apenas com minúsculas não são classificadas pelo HAREM como EM em caso nenhum.

16.1.5 Relação entre a classificação e a identificação

Embora a classificação deva ter em conta o significado da EM no texto, a identificação (ou seja a sua delimitação) deve restringir-se às regras das maiúsculas enunciadas acima. Ou seja, apenas a parte associada ao nome próprio deve ser identificada, embora classificada, se for caso disso, a entidade maior em que se enquadra. Vejam-se os seguintes exemplos:

```
Certo: a filha de <PESSOA TIPO="INDIVIDUAL">Giuteyte</PESSOA>
Certo: o tratado de <ACONTECIMENTO TIPO="EFEMERIDE">Tordesilhas
       </ACONTECIMENTO> dividiu o mundo
```

Embora apenas *Tordesilhas* tenha sido identificado, é o tratado de Tordesilhas que é classificado como um ACONTECIMENTO.

Isso também se aplica aos casos em que no texto um fragmento ou parte da EM é compreendida como relatando anaforicamente a uma entidade não expressa na sua totalidade. Por exemplo, na frase *A Revolução de 1930 foi sangrenta, e a de 1932 ainda mais*, deve marcar-se *1932* como <ACONTECIMENTO TIPO="EFEMERIDE"> e não como <TEMPO TIPO="DATA">.

Nos casos em que há enganos de ortografia ou grafia no texto, em particular quando uma palavra tem uma maiúscula a mais ou a menos e tal é notório, escolhemos corrigir mentalmente a grafia (maiúscula /minúscula) de forma a poder classificar correctamente. Além disso, estamos a pensar em marcar estes casos, na colecção dourada, com uma classificação META="ERRO".

```
Certo: O grupo terrorista <PESSOA TIPO="GRUPO" META="ERRO">Setembro
       negro</PESSOA>
```

Outras excepções, mais sistematicamente apresentadas, são as seguintes:

Para poder distinguir mais facilmente os casos de classes de objectos cujo nome inclui um nome próprio (geralmente de uma pessoa), adicionámos a seguinte regra de identifi-

cação para a categoria COISA: a preposição anterior também deve fazer parte da EM em *constante de Planck*, *bola de Berlim* ou *porcelana de Limoges*.

Por outro lado, consideramos que as EM de categoria VALOR e do tipo QUANTIDADE ou MOEDA devem incluir a unidade, independentemente de esta ser grafada em maiúscula ou minúscula.

Finalmente, no caso de doenças, formas de tratamento e certo tipo de acontecimentos consideramos aceitáveis um conjunto finito de nomes comuns precedendo a própria EM, cuja lista foi descrita anteriormente, na secção 16.1.4.

16.1.6 Escolha da EM máxima

Para evitar uma excessiva proliferação de EM com identificações alternativas, os sistemas e CD são construídos de forma a escolher a EM máxima, ou seja, aquela que contém, numa única interpretação possível, o maior número de palavras. Assim, e muito embora fosse possível ter tomado a decisão inversa e pedir, por exemplo, o máximo número de EM com uma interpretação possível separada, a escolha recaiu em preferir a EM maior.

Por exemplo:

```
Certo: O <PESSOA TIPO="CARGO">ministro dos Negócios Estrangeiros do
       Governo Sócrates</PESSOA>
Certo: <ORGANIZACAO TIPO="INSTITUICAO">Comissão de Trabalhadores da
       IBM Portugal</ORGANIZACAO>
Certo: <ACONTECIMENTO TIPO="EFEMERIDE">Jogos Olímpicos de Inverno de
       2006</ACONTECIMENTO>
```

As únicas excepções a esta regra são períodos descritos por duas datas, e intervalos de valores descritos por duas quantidades.

16.2 Categoria PESSOA

16.2.1 Tipo INDIVIDUAL

Títulos que precedem nomes

Os títulos (*dr.*, *eng.*, *arq.*, *Pe.*, *etc.*) usados no tratamento de uma pessoa devem ser incluídos na EM que delimita essa pessoa.

Formas de tratamento normalmente usadas para anteceder um nome, como presidente, ministro, etc. também devem ser incluídos, assim como graus de parentesco (*tia*, *irmão*, *avó*, etc) quando fazem parte da forma de tratamento. Outras relações profissionais como patrão, chefe, etc. não devem ser incluídos, nem profissões que não façam parte da forma de tratamento.

```
Certo:  O <PESSOA TIPO="INDIVIDUAL">Dr. Sampaio</PESSOA>.
Certo:  O <PESSOA TIPO="INDIVIDUAL">presidente Jorge Sampaio</PESSOA>.
Certo:  O <PESSOA TIPO="INDIVIDUAL">padre Melícias</PESSOA>.
Certo:  O <PESSOA TIPO="INDIVIDUAL">tio Zeca</PESSOA>.
Certo:  O acordeonista <PESSOA TIPO="INDIVIDUAL">Miguel Sá</PESSOA>.
Errado: O <PESSOA TIPO="INDIVIDUAL">acordeonista Miguel Sá</PESSOA>.
```

Cargos incluídos

Os cargos que não estejam separados por uma vírgula do nome devem ser incluídos no tipo `INDIVIDUAL`. Se houver vírgula, ficam de fora.

```
Certo: O <PESSOA TIPO="INDIVIDUAL">Presidente da República Jorge
       Sampaio</PESSOA>, disse...
Certo: O <PESSOA TIPO="CARGO">Presidente da República</PESSOA>,
       <PESSOA TIPO="INDIVIDUAL">Jorge Sampaio</PESSOA>, disse...
```

Caso o cargo seja descrito após o nome, aplica-se a mesma regra.

```
Certo: <PESSOA TIPO="INDIVIDUAL">Jorge Sampaio</PESSOA>, <PESSOA
       TIPO="CARGO">Presidente da República</PESSOA>, assinou...
```

Outros

Diminutivos, alcunhas, iniciais, nomes mitológicos e entidades religiosas são etiquetados nesta categoria.

```
Certo: <PESSOA TIPO="INDIVIDUAL">Zé</PESSOA>.
Certo: <PESSOA TIPO="INDIVIDUAL">‘‘Iron’’ Tyson</PESSOA>.
Certo: <PESSOA TIPO="INDIVIDUAL">John (Jack) Reagan</PESSOA>.
Certo: <PESSOA TIPO="INDIVIDUAL">JFK</PESSOA>.
Certo: <PESSOA TIPO="INDIVIDUAL">Deus</PESSOA>.
```

EM que não são cargos, mas que referem uma pessoa individual, são para ser etiquetados como tal.

```
Certo: <PESSOA TIPO="INDIVIDUAL">Vossa Excia</PESSOA>
```

16.2.2 Tipo GRUPOIND

Esta categoria representa grupo de indivíduos (do tipo `INDIVIDUAL`) que não têm um nome "estático" como grupo (ao contrário dos Beatles, por exemplo).

```
Certo: <PESSOA TIPO="GRUPOIND">Vossas Excias</PESSOA>.
Certo: O <PESSOA TIPO="GRUPOIND">Governo Clinton</PESSOA> foi a...
Certo: Foi em casa dos <PESSOA TIPO="GRUPOIND">Mirandas</PESSOA>.
Certo: O governo de <PESSOA TIPO="GRUPOIND">Cavaco Silva</PESSOA>
       esteve presente na cerimónia.
```

No caso de haver um grupo de pessoas discriminadas, deve-se etiquetar cada um dos nomes em separado. Na frase de exemplo *Os tenistas Carlos Guerra e António Gomes foram a Wimbledon*:

```
Certo:  Os tenistas <PESSOA TIPO="INDIVIDUAL">Carlos Guerra</PESSOA>
        e <PESSOA TIPO="INDIVIDUAL">António Gomes</PESSOA> foram a
        Wimbledon.
Errado: Os tenistas <PESSOA TIPO="GRUPO">Carlos Guerra e António
        Gomes</PESSOA> foram a Wimbledon.
```

16.2.3 Tipo CARGO

O tipo `CARGO` deve ser usado na referência de um posto que é ocupado por uma pessoa, mas que poderá no futuro ser ocupado por outros indivíduos. Ou seja, num dado contexto, `CARGO` pode representar uma pessoa em concreto, mas através da referência ao seu cargo. Note-se que noutros casos a mesma EM (que anotamos de qualquer maneira sempre da mesma forma, como `<PESSOA TIPO="CARGO">`) pode referir-se ao próprio cargo, que pode ser desempenhado por diferentes pessoas ao longo do tempo. Exemplos: *Papa*, *Ministro dos Negócios Estrangeiros*, *Rainha da Abissínia*.

Cargo associado a uma organização

Cargos que possuem na descrição uma organização, devem ter apenas uma etiqueta `<PESSOA TIPO="CARGO">` que abrange a organização.

```
Certo:  O <PESSOA TIPO="CARGO">Presidente da ONU</PESSOA> foi...
Errado: O <PESSOA TIPO="CARGO">Presidente</PESSOA> da
          <ORGANIZACAO>ONU</ORGANIZACAO> foi...
```

16.2.4 Tipo GRUPOCARGO

O tipo `GRUPOCARGO` é análogo ao `GRUPOIND`, designando EM que referem um conjunto de pessoas, através de um cargo.

```
Certo: os <PESSOA TIPO="GRUPOCARGO">Ministros dos Negócios
       Estrangeiros da União Europeia</PESSOA>
```

16.2.5 Tipo MEMBRO

Nos casos onde um indivíduo é mencionado pela organização que representa (e não um grupo), é marcado com o tipo MEMBRO.

```
Certo: Ele foi abordado por um <PESSOA TIPO="MEMBRO">GNR</PESSOA>
       à paisana.
Certo: O <PESSOA TIPO="MEMBRO">Mórmon</PESSOA> estava na sala ao lado.
```

No caso de entrevistas, quando o entrevistador é referenciado pelo nome da publicação, deve ser etiquetado como <PESSOA TIPO="MEMBRO">:

```
Certo: <PESSOA TIPO="MEMBRO">Jornal Nacional</PESSOA> - O que sente
       depois de ganhar o prémio?
Errado: <ORGANIZACAO TIPO="EMPRESA">Jornal Nacional</ORGANIZACAO>
        - O que sente depois de ganhar o prémio?
```

Os próprios nomes são ABSTRACCAO

Quando o texto foca o nome e não a referência do próprio nome, esse nome (independentemente de se referir a uma pessoa, animal, organização, etc.) é marcado como <ABSTRACCAO TIPO="NOME"> (detalhado na secção 16.9):

```
Certo:  Dei-lhe o nome de <ABSTRACCAO TIPO="NOME">João Sem Medo</ABSTRACCAO>.
Errado: Dei-lhe o nome de <PESSOA TIPO="INDIVIDUAL">João Sem Medo</PESSOA>.
Certo:  Uma organização suspeita denominada <ABSTRACCAO TIPO="NOME">Os
        Inimigos das Formigas</ABSTRACCAO> foi ilegalizada ontem no Cairo.
```

16.2.6 Tipo GRUPOMEMBRO

Este tipo, que apenas não se chama GRUPO para salientar a sua relação com o tipo MEMBRO, abrange EM que se referem a um conjunto de pessoas como membros de uma organização ou conceito semelhante (tal como equipa ou seita).

```
Certo:  Os <PESSOA TIPO="GRUPOMEMBRO">Mórmons</PESSOA> acreditam
        no profeta John Smith.
Certo:  Os <PESSOA TIPO="GRUPOMEMBRO">Genesis</PESSOA> deram um
        concerto ontem.
Certo:  O <PESSOA TIPO="GRUPOMEMBRO">BE</PESSOA> reuniu-se ontem.
Certo:  O <PESSOA TIPO="GRUPOMEMBRO">FC Porto</PESSOA> jogou muito
        bem e venceu o jogo.
Errado: O <ORGANIZACAO>FC Porto</ORGANIZACAO> jogou muito bem e venceu
        o jogo.
```

```
Certo:  O <ORGANIZACAO>FC Porto</ORGANIZACAO> tem um estádio...
Errado: O <PESSOA TIPO="GRUPOMEMBRO">FC Porto</PESSOA> tem um estádio.
```

16.3 Categoria ORGANIZACAO

16.3.1 Tipo ADMINISTRACAO

Este tipo pretende etiquetar as organizações relacionadas com a administração e governação de um território, tal como ministérios, municípios, câmaras, autarquias, secretarias de estado (Exemplos: *Secretaria de Estado da Cultura*, *Brasil*, *Prefeitura de São Paulo*, *Câmara Municipal de Leiria*). Inclui também as organizações que têm a ver com a governação a nível internacional ou supra-nacional (Exemplos: *ONU*, *UE*)

Países ou territórios como organização

EM referentes a países, territórios, regiões autónomas ou mesmo territórios ocupados ou ex-colónias, podem referir à organização, pelo que se deve usar as etiquetas `<LOCAL TIPO="ADMINISTRATIVO">` ou `<ORGANIZACAO TIPO="ADMINISTRACAO">`, dependendo do contexto.

```
Certo: <ORGANIZACAO TIPO="ADMINISTRACAO">Moçambique</ORGANIZACAO>
       votou a favor na ONU.
Certo: <LOCAL TIPO="ADMINISTRATIVO">Moçambique</ORGANIZACAO> faz
       fronteira com a Tanzânia.
```

Referências a ministérios

A referência à entidade organizativa deve ser explícita, para ser considerada uma EM de categoria `ORGANIZACAO`. No caso em que se refere a uma área de competência da organização, é uma `ABSTRACCAO` de tipo `DISCIPLINA` e não uma `ORGANIZACAO`.

```
Certo: O <ORGANIZACAO TIPO="ADMINISTRACAO">Ministério do Ambiente
       </ORGANIZACAO> gere a política ambiental.
Certo: O ministro do <ABSTRACCAO TIPO="DISCIPLINA">Ambiente
       </ABSTRACCAO> gere a política ambiental.
```

No seguinte caso, a vagueza da EM deve ser mantida (`ORGANIZACAO` ou `ABSTRACCAO`):

```
Certo: O <ORGANIZACAO|ABSTRACCAO TIPO="ADMINISTRACAO|DISCIPLINA">
       Ambiente</ORGANIZACAO|ABSTRACCAO> gere a política ambiental.
```

No caso da menção a Ministro com maiúscula, ambas as situações serão consideradas correctas.

```
Certo: O <PESSOA TIPO="CARGO">Ministro do Ambiente</PESSOA>
       gere a política ambiental.
Certo: O ministro do <ABSTRACCAO TIPO="DISCIPLINA">Ambiente</ABSTRACCAO>
       gere a política ambiental.
```

Esta excepção tem a haver com o facto de escrever Ministro com maiúscula, no contexto apresentado, está errado. As novas versões da colecção dourada estão etiquetadas de maneira a suportar erros como este, sem penalizar os sistemas.

16.3.2 Tipo EMPRESA

Este tipo abrange organizações com fins lucrativos, como empresas, sociedades, clubes, etc. (Exemplos: *Boavista FC*, *Círculo de Leitores*, *Livraria Barata*, (discoteca) *Kapital*) em contextos em que são mencionadas como tal.

```
Certo: O <ORGANIZACAO TIPO="EMPRESA">Boavista FC</ORGANIZACAO>
       contratou novos jogadores.
```

16.3.3 Tipo INSTITUICAO

Todas as organizações que não possuem fins lucrativos (não sendo, portanto, empresas), nem um papel directo na governação, são do tipo `INSTITUICAO`. Este tipo abrange instituições no sentido estrito, associações e outras organizações de espírito cooperativo, universidades, colectividades, escolas ou partidos políticos (Exemplos: *Associação de Amizade Portugal-Bulgária*, *Universidade Federal do Rio Grande do Sul*, *Liceu Maria Amália*, *PC do B (Partido Comunista do Brasil)*, *Museu do Ar*, *PSP*, *Amnistia Internacional*).

16.3.4 Tipo SUB

As EM de tipo `SUB` referem-se a determinados sectores de uma organização, mas sem autonomia para ser considerada ela própria uma organização, tais como departamentos, secções, assembleias gerais, comissões, comitês, secretarias, etc.

```
Certo: A sua queixa deve dirigir-se ao <ORGANIZACAO TIPO="SUB">
       Departamento dos Alunos de Mestrado do IST</ORGANIZACAO>
Certo: A <ORGANIZACAO TIPO="SUB">Assembleia Geral da Empresa
       PTO</ORGANIZACAO> tem poder para reprovar o orçamento
       proposto.
```

Nome da empresa incluído no SUB

No caso do nome da organização ser parte integrante do tipo `SUB`, este também faz parte da EM.

```
Certo:  O <ORGANIZACAO TIPO="SUB">Departamento de Marketing da
        General Motors</ORGANIZACAO>.
Errado: O <ORGANIZACAO TIPO="SUB">Departamento de Marketing da
         <ORGANIZACAO TIPO=EMPRESA">General Motors
         </ORGANIZACAO></ORGANIZACAO>.
Errado:  O <ORGANIZACAO TIPO="SUB">Departamento de Marketing
          </ORGANIZACAO> da <ORGANIZACAO TIPO=EMPRESA">General
          Motors</ORGANIZACAO>.
```

Sucursais e filiais

No caso de sucursais, filiais, empresas em regime de *franchising*, etc, ou seja, onde haja autonomia suficiente para as considerarmos uma organização autónoma, a EM deve ser classificada como uma `EMPRESA`, e não uma `SUB`.

```
Certo:  A <ORGANIZACAO TIPO=EMPRESA>VW Portugal</ORGANIZACAO> vai
        lançar uma iniciativa.
Errado: A <ORGANIZACAO TIPO=SUB>VW Portugal</ORGANIZACAO> vai
         lançar uma iniciativa.
Certo:  A <ORGANIZACAO TIPO=EMPRESA>GM</ORGANIZACAO> disse à
        <ORGANIZACAO TIPO=EMPRESA>GM Portugal</ORGANIZACAO> para
        recolher veículos.
```

Embora a organização designada pela segunda EM tenha uma dependência explícita em relação à designada pela primeira EM, a sua identificação sai do âmbito do tipo `SUB`, que pretende delimitar apenas EM que são sectores dentro de uma organização.

Como tal, resumindo:

```
Certo: <ORGANIZACAO TIPO="EMPRESA">GM Portugal</ORGANIZACAO>
Certo: <ORGANIZACAO TIPO="SUB">Departamento de Vendas da
       GM</ORGANIZACAO>
Certo: A <ORGANIZACAO TIPO="INSTITUICAO">Faculdade de Ciências
       da Universidade de Lisboa</ORGANIZACAO>
Certo: A <ORGANIZACAO TIPO="INSTITUICAO">Universidade de Lisboa
       </ORGANIZACAO> recomendou à <ORGANIZACAO TIPO="INSTITUICAO">
       Faculdade de Ciências</ORGANIZACAO>
Certo: A <ORGANIZACAO TIPO="SUB">Reprografia da Universidade de
       Lisboa</ORGANIZACAO> fecha às 16h.
```

Organizações dentro de cargos

Empresas incluídas na descrição dos cargos de pessoas não são para etiquetar.

```
Certo:  O <PESSOA TIPO="CARGO">CEO da Microsoft</PESSOA> foi a...
Errado: o <PESSOA TIPO="CARGO">CEO</PESSOA> da <ORGANIZACAO>
        Microsoft</ORGANIZACAO> foi a...
```

16.4 Categoria TEMPO

As EM de tipo `TEMPO` não devem conter palavras que não referem explicitamente a data ou a hora.

Textos como *final de 1999*, *próximo dia 22*, *entre 14 e 18*, *meados de Agosto*, ou *antes do dia 3*, só devem ter marcadas como EM, respectivamente, *1999*, *22*, *14*, *18*, *Agosto* e *3*.

A única excepção é para nomes de meses em português do Brasil, como já foi referido.

Note-se que, embora a idade de uma pessoa seja referida em anos (e, como tal, uma quantidade de tempo), deve ser marcada como `<VALOR TIPO="QUANTIDADE">` e não como `<TEMPO>`, uma vez que se refere a uma quantidade e não a uma localização temporal.

16.4.1 Tipo DATA

Referência a uma data

Inclui todas as referências a dias, mês e ano. Referências a mês e ano, ou só a ano, devem ser consideradas de tipo `DATA` se, no contexto, a referência indica uma localização temporal única. Esta pode ter diferentes granularidades (pode ser um dia ou vários meses).

```
Certo:  Camões morreu em <TEMPO TIPO="DATA">1580</TEMPO>.
Certo:  O EURO foi em <TEMPO TIPO="DATA">2004</TEMPO>.
Certo:  No dia <TEMPO TIPO="DATA">24 de Agosto de 1976</TEMPO>.
Certo:  Em <TEMPO TIPO="DATA">Agosto de 1976</TEMPO> foi a Final da Taça.
Errado: Em <TEMPO TIPO="DATA">Agosto de 1976</TEMPO> houve 54 suicídios.
Certo:  Em <TEMPO TIPO="PERIODO">Agosto de 1976</TEMPO> houve 54 suicídios.
Certo:  Em <TEMPO TIPO="DATA">1974</TEMPO> houve a Revolução.
Errado: Em <TEMPO TIPO="DATA">1974</TEMPO> vendeu-se 200.000 carros.
Certo:  Em <TEMPO TIPO="PERIODO">1974</TEMPO> vendeu-se 200.000 carros.
```

Referência a duas datas

Referências a períodos de tempo através da data de início e da data do final, devem ser etiquetadas com duas EM `<TEMPO TIPO="DATA">` separadas, e não com uma única etiqueta `<TEMPO TIPO="PERIODO">`.

```
Certo:  Entre <TEMPO TIPO="DATA">4</TEMPO> a <TEMPO TIPO="DATA">6 de
        Dezembro</TEMPO> há o Festival.
Errado: Entre <TEMPO TIPO="PERIODO">4 a 6 de Dezembro</TEMPO> há o
         Festival.
```

Neste caso, e apenas neste caso, aplicamos a regra de duas EM mínimas são melhores do que uma EM mais longa.

16.4.2 Tipo HORA

Referência a horas

Aplicam-se as mesmas regras descritas no tipo `DATA`, para as horas.

```
Certo: Às <TEMPO TIPO="HORA">2h00</TEMPO> vou ao dentista.
Certo: Entre as <TEMPO TIPO="HORA">2h00</TEMPO> e as <TEMPO
       TIPO="HORA">4h00</TEMPO> estou no dentista.
```

Referência a fusos horários

Horas com modificação referente a fusos horários devem abranger essa informação, uma vez que é parte essencial para interpretar o tempo da ocorrência.

```
Certo: O atentado ocorreu às <TEMPO TIPO="HORA">13h, hora de
       Lisboa</TEMPO>, e fez...
```

16.4.3 Tipo PERIODO

Engloba as EM que referem um intervalo de tempo contínuo e não repetido, com apenas um início e um fim (Exemplos: *Inverno*, *anos 80*, *século XIX*, *1984*, *pós-25 de Abril*, a *Idade do Bronze*). Note-se que a mesma EM pode referir um período único ou cíclico, ou uma data:

```
Certo: Vou três vezes a Londres no próximo <TEMPO TIPO="PERIODO">
       Inverno</TEMPO>.
Certo: O <TEMPO TIPO="CICLICO">Inverno</TEMPO> em Oslo costuma ser frio.
Certo: A Joana nasceu no <TEMPO TIPO="DATA">Inverno</TEMPO> passado.
```

Período referido como um acontecimento

É normal referir um determinado período de tempo através de um evento que decorreu durante esse período. Um exemplo é a *Segunda Guerra Mundial*, que pode ser referenciada como o evento ou como um período de tempo, sendo imprescindível a análise do contexto da EM para definir a semântica correcta, como se mostra nos seguintes exemplos:

```
Certo : Durante a <TEMPO TIPO="PERIODO">2ª Guerra Mundial</TEMPO>, surgiram
        os primeiros aviões a jacto.
Certo : A <ACONTECIMENTO TIPO="EFEMERIDE">2ª Guerra Mundial</ACONTECIMENTO>
        envolveu meio mundo.
```

A diferença é marcada pela expressão *Durante*, que desde logo indica que a EM é para ser interpretada como um PERIODO.

```
Certo: Durante a <TEMPO TIPO="PERIODO">Guerra Fria</TEMPO> não
       houve ataques nucleares.
```

Período implícito

Semelhante ao caso descrito acima, há outras referências a períodos que são implícitos a partir de diversas EM que, à primeira vista, parecem pertencer a outras categorias semânticas, como é ilustrado abaixo. Por exemplo, tome-se o caso de alguém que trabalhou na IBM e depois passou a trabalhar para a Sun. A sua menção à IBM na seguinte frase refere-se ao período no qual esteve lá empregado.

```
Certo: Depois da <TEMPO TIPO="PERIODO">IBM</TEMPO>, fui trabalhar
       para a <ORGANIZACAO TIPO="EMPRESA">Sun</ORGANIZACAO>.
```

Da mesma forma, no contexto de um pessoa que foi trabalhador no navio D. Luís, este deve ser anotado como TEMPO.

```
Certo: Depois do <TEMPO TIPO="PERIODO">D. Luís</TEMPO>, fiquei
       desempregado.
```

16.4.4 Tipo CICLICO

Compreende períodos recorrentes, quando empregues como tal (*Natal*, *1º de Janeiro*, *Páscoa*).

Há que ter atenção que uma dada EM da categoria TEMPO pode ter quase sempre duas interpretações:

```
No dia 6 de Novembro comemora-se...
No dia 6 de Novembro vai haver uma greve...
```

No primeiro caso, como acontece todos os anos, é <TEMPO TIPO="CICLICO">. No segundo caso, é <TEMPO TIPO="DATA">, porque se refere a um único dia.

16.5 Categoria ACONTECIMENTO

Esta categoria abrange acontecimentos que são únicos e, de uma maneira geral, irrepetíveis (EFEMERIDE), e outros cujo significado se reduz a designar um certo conjunto de actividades e de acções: ORGANIZADO (com sub-partes) e EVENTO (indivisível).

16.5.1 Tipo EFEMERIDE

Acontecimento ocorrido no passado e não repetível tal como o *25 de Abril*, o *11 de Setembro*, a *2ª Guerra Mundial*.

```
Certo: A <ACONTECIMENTO TIPO="EFEMERIDE">Revolução Francesa
       </ACONTECIMENTO> mudou a Europa.
Certo: O <ACONTECIMENTO TIPO="EFEMERIDE"> caso Whitaker
       </ACONTECIMENTO> abalou a Grã-Bretanha.
```

16.5.2 Tipo ORGANIZADO

Acontecimento multifacetado, que poderá durar vários dias, e geralmente conter vários EVENTO. Exemplos são a *Copa*, o *Euro 2004*, os *Jogos Olímpicos*, o *Festival de Jazz do Estoril*.

Acontecimentos periódicos

Quando o acontecimento em questão é um evento periódico, distinguido pelo ano do acontecimento ou pelo seu local, estes (data ou local) devem ser incluídos na etiqueta de acontecimento.

```
Certo:  <ACONTECIMENTO TIPO="ORGANIZADO">Jogos Olímpicos de
        2004</ACONTECIMENTO>
Certo:  <ACONTECIMENTO TIPO="ORGANIZADO">Jogos Olímpicos de
        Atenas</ACONTECIMENTO>
Errado: <ACONTECIMENTO TIPO="ORGANIZADO">Jogos
         Olímpicos</ACONTECIMENTO> de <TEMPO TIPO="DATA">2004</TEMPO>
Errado: <ACONTECIMENTO TIPO="ORGANIZADO">Jogos Olímpicos
         </ACONTECIMENTO> de <LOCAL TIPO="ADMINISTRATIVO">Atenas</LOCAL>.
```

16.5.3 Tipo EVENTO

Acontecimento pontual, organizado ou não, tal como *Benfica-Sporting*, *Chico Buarque no Coliseu*, *Buzinão na Ponte*.

Diferenças entre ORGANIZADO e EVENTO

Um bom exemplo da separação entre os tipos ORGANIZADO e EVENTO é o Euro'2004, que foi um acontecimento ORGANIZADO, que incluiu vários EVENTOs (jogos, festas, conferências, etc).

Quando se diz que um evento pode ser organizado ou não, dá-se o exemplo de um jogo de futebol (organizado) ou de uma manifestação popular espontânea (não organizada).

Note-se o caso apresentado, *Chico Buarque no Coliseu*, onde a combinação de uma PESSOA num determinado LOCAL produz um EVENTO, e como tal, deve ser etiquetado como tal, e não como duas EM distintas.

16.6 Categoria COISA

Esta categoria abrange coisas podem ser únicas e referenciadas como um item (OBJECTO), podem referir substâncias sem forma ou feitio determinado (SUBSTANCIA), podem representar uma categoria específica que descreve uma população de objectos (CLASSE), ou pode abranger EM cujo significado é um conjunto de objectos, discriminados a partir de uma propriedade comum, e que instancia uma classe de objectos (MEMBROCLASSE)

16.6.1 Tipo OBJECTO

Refere um objecto ou construção em particular, referido por um nome próprio. Inclui planetas, estrelas, cometas e sóis. Também pode conter objectos específicos.

```
Certo:  A fragata <COISA TIPO="OBJECTO">D. Luís</COISA> atracou ontem.
Certo:  Comprámos uma casa ao pé do chalé <COISA TIPO="OBJECTO">
        Sonho Perfeito</COISA>.
Certo:  Consegue-se ver <COISA TIPO="OBJECTO">Marte</COISA> hoje
        à noite.
```

16.6.2 Tipo SUBSTANCIA

Refere substâncias elementares que não se podem considerar objectos, por não serem contáveis (por exemplo, *Paracetamol*, H_2O).

```
Certo: O <COISA TIPO="SUBSTANCIA">DNA</COISA> é um poço de enigmas.
Certo: O médico disse que tenho falta de vitamina <COISA
       TIPO="SUBSTANCIA">D</COISA>.
```

16.6.3 Tipo CLASSE

Este tipo, que, convém salientar, junto com MEMBROCLASSE, é análogo à distinção feita na categoria PESSOA entre GRUPOMEMBRO e MEMBRO (CLASSE $\Longleftrightarrow$ GRUPOMEMBRO e MEMBROCLASSE $\Longleftrightarrow$ MEMBRO), representa classes de objectos que têm um nome e, como tal, dão origem a uma EM (Exemplos: *contador Geiger*, *flauta de Bisel*, *PC*, *SCSI*, *PDF*).

```
Certo: A FCCN exige relatórios em folhas <COISA TIPO="CLASSE">A4</COISA>.
Certo: Os móveis <COISA TIPO="CLASSE">Luís XV</COISA> são muito raros.
```

Muitas vezes, uma EM deste tipo refere o 'inventor' da classe (exemplo: *lâmpada de Edison*). O determinante deve ser incluído, para enfatizar essa semântica.

```
Certo:  pêndulo <COISA TIPO="CLASSE">de Foucault</COISA>.
Errado: pêndulo de <COISA TIPO="CLASSE">Foucault</COISA>.
```

"Consumíveis" tais como *pastéis de Belém*, *bolas de Berlim*, *Tiramisu de chocolate*, *vinho de Setúbal* (num contexto de tipo de vinho, e não como oriundo de um local) também são para ser etiquetados como `<COISA TIPO="CLASSE">` (ou tipo `SUBSTANCIA`, dependendo de serem contáveis ou não).

```
Certo: Receitas de <COISA TIPO="CLASSE">Bacalhau à Brás</COISA>.
Certo: Adoro bolas <COISA TIPO="CLASSE">de Berlim</COISA>.
Certo: Os pastéis <COISA TIPO="CLASSE">de Belém</COISA> têm muita fama.
```

16.6.4 Tipo MEMBROCLASSE

Este tipo abrange EM referentes a uma instanciação de classes, ou seja, objectos determinados que são referidos através da classe a que pertencem. Inclui produtos comercializados, e que são referidos por uma marca ou por uma empresa.

No exemplo *Eu gosto de comer Corn Flakes ao pequeno almoço*, estamos a referir-nos a uma série de produtos comerciais que representam uma classe de objectos (neste caso, cereais de pequeno almoço), mas quando dizemos *Os Corn Flakes de hoje estavam horríveis* referimo-nos a uma instância particular.

```
Certo: O meu <COISA TIPO="MEMBROCLASSE">Fiat Punto<COISA> foi à revisão.
Certo: O <COISA TIPO="MEMBROCLASSE">MS Word 2003</COISA> da Cristina
       rebentou hoje.
```

Note-se que, nos casos seguintes, estamos a referir-nos à `CLASSE` e não a um membro.

```
Certo : As consolas <COISA TIPO="CLASSE" MORF="F,P">Mega Drive</COISA>
        são compatíveis com ...
Certo : Os <COISA TIPO="CLASSE" MORF="M,P">Fiat Punto</COISA> têm
        bons travões.
```

16.7 Categoria LOCAL

A categoria `LOCAL` abrange todas as referências a sítios específicos.

16.7.1 Tipo CORREIO

O tipo CORREIO abrange todas as referências a locais com indicações completas, tais como moradas, números de salas, salas de cinema (Exemplos: *Sala 6*, *Caixa Postal 2400*, *Rua da Escola 15B*). Referências que não incluam endereços completos, ou cuja intenção não é facultar uma morada completa, devem ser do tipo ALARGADO e não CORREIO.

Abrangência do tipo CORREIO

Ao assinalar um <LOCAL TIPO="CORREIO">, deve-se incluir todos os locais inerentes à referência da localização exacta.

Certo: <LOCAL TIPO="CORREIO">Rua Augusta, nº 5 - Lisboa</LOCAL>.
Errado: <LOCAL TIPO="CORREIO">Rua Augusta, nº 5</LOCAL> - <LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>

16.7.2 Tipo ADMINISTRATIVO

Identifica localizações que foram criadas e/ou delimitadas pelo Homem. Inclui países, bairros, regiões geopolíticas, entre outros. Exemplos: *Rio de Janeiro*, *Alentejo*, *Bairro dos Anjos*, *Ásia Menor*, *Região Autónoma dos Açores*, *Jardim das Amoreiras*, *Médio Oriente*, *América Latina*, *África*, *Países de Leste*.

Não se deve incluir a referência ao tipo de local, caso haja, como são os exemplos distrito, concelho, aldeia, vila, cidade, bairro, região, etc (excepção feita se estas referências tiverem pelo menos uma letra maiúscula).

Certo: vou para a cidade de <LOCAL TIPO="ADMINISTRATIVO">Viseu</LOCAL>.
½*Errado: vou para a <LOCAL TIPO="ADMINISTRATIVO">cidade de Viseu</LOCAL>.*
Certo: vou para a <LOCAL TIPO="ADMINISTRATIVO">Cidade de Viseu</LOCAL>.

Locais dentro de organizações

Não há necessidade de colocar um <LOCAL TIPO="ADMINISTRATIVO"> dentro de ORGANIZACAO. Aliás, já tornámos explícito que não deve haver encaixe de EM dentro de EM.

Certo: <ORGANIZACAO>Câmara Municipal de Braga</ORGANIZACAO>.
Errado: <ORGANIZACAO>Câmara Municipal de <LOCAL>Braga</LOCAL> </ORGANIZACAO>.

'Locais' referidos como administração

Nomes de países, cidades, entre outros, designam locais... mas há casos em que a referência ao local é implícita ao seu Governo, ou seja, uma EM de categoria ORGANIZACAO e de tipo ADMINISTRACAO.

```
Certo:  <ORGANIZACAO TIPO="ADMINISTRACAO">Portugal</ORGANIZACAO>
        condenou a acção da <ORGANIZACAO TIPO="ADMINISTRACAO">
        Indonésia</ORGANIZACAO>.
Errado: <LOCAL TIPO="ADMINISTRATIVO">Portugal</LOCAL> condenou a
         acção da <LOCAL TIPO="ADMINISTRATIVO">Indonésia</LOCAL>.
```

16.7.3 Tipo GEOGRAFICO

Indica localizações de geografia física que apenas foram baptizadas (e não construídas) pelo Homem.

Não se deve incluir o tipo de acidente geográfico, ou seja, referências como rio, serra, mar, península, entre outras, exceptuando se estas contiverem pelo menos uma letra maiúscula.

```
Certo:  Vou ao estuário do <LOCAL TIPO="GEOGRAFICO">Douro</LOCAL>.
Certo:  Vou ao estuário do rio <LOCAL TIPO="GEOGRAFICO">Douro</LOCAL>.
Errado: Velejo no <LOCAL TIPO="GEOGRAFICO">rio Douro</LOCAL>.
Certo:  Velejo no <LOCAL TIPO="GEOGRAFICO">Rio Douro</LOCAL>.
```

16.7.4 Tipo VIRTUAL

O tipo VIRTUAL engloba locais como a Internet, e números de telefone ou de fax, etc., desde que contenham ou algarismos ou letras maiúsculas. URLs ou endereços de correio electrónico não são nunca considerados como EM. Também abrange locais de publicação, referidos pelos nomes dos meios de comunicação social.

Só se deve etiquetar os números de telefone, nunca o que os precede!

```
Certo:  Vê o meu sítio na <LOCAL TIPO="VIRTUAL">Internet</LOCAL>.
Certo:  Tel: <LOCAL TIPO="VIRTUAL">(096)347845 4563</LOCAL>.
Errado: <LOCAL TIPO="VIRTUAL">Tel: (096)347845 4563</LOCAL>.
```

Referência a local de publicação

Quando o local referido é um sítio abstracto que pode não corresponder a um local físico (como é exemplo um programa ou uma série de televisão ou de rádio), ou é mencionado na função de 'alojamento' de um item (como uma notícia de um jornal), a EM deve ser classificada como do tipo VIRTUAL:

```
Certo: Podes ler o meu artigo no <LOCAL TIPO="VIRTUAL">Jornal de
       Notícias</LOCAL>.
Certo: No <LOCAL TIPO="VIRTUAL">Diário de Notícias</LOCAL> de
       hoje, vinha referido...
```

Neste último caso, a interpretação certa é o local onde estão as notícias (que neste caso, pode ou não corresponder a um suporte de papel). No entanto, é o local que se pretende referenciar, e não o OBJECTO, MARCA ou EMPRESA.

16.7.5 Tipo ALARGADO

Deve conter referências a locais que não estão nas categorias acima, mas que referem um determinado sítio físico, como é o exemplo de pontos de encontro em edifícios, bares, hotéis, praças, centros de congressos, restaurantes, etc. (Exemplo: C*entro Comercial Amoreiras*).

'Organizações' referidas como LOCAL

Frequentemente, hotéis e centros de congressos são referenciados como sítios de ponto de encontro ou onde ocorrem eventos. Neste caso, nesse contexto essas EM são LOCAL de tipo ALARGADO.

```
Certo:  O Congresso decorrerá no <LOCAL TIPO="ALARGADO">Hotel Beta
        </LOCAL> e durará...
Errado: O Congresso decorrerá no <ORGANIZACAO TIPO="EMPRESA">Hotel Beta
          </ORGANIZACAO> e durará...
Certo:  O <ORGANIZACAO TIPO="EMPRESA">Hotel Beta</ORGANIZACAO> emprega
        500 funcionários...
Errado: O <LOCAL TIPO="ALARGADO">Hotel Beta</LOCAL> emprega 500
          funcionários...
```

Esta regra, aliás, aplica-se a todas as EM originalmente de outras categorias, sempre que no contexto remetam para um local concreto, como no seguinte exemplo de um ponto de encontro:

```
Certo : Encontramo-nos debaixo da <LOCAL TIPO="ALARGADO">Torre Eiffel
        </LOCAL>.
```

Diferença entre ALARGADO e CORREIO

No caso de se referir uma rua, avenida ou praça como um local onde ocorreu ou está localizada qualquer coisa, mas não como se de uma morada ou endereço se tratasse, é um <LOCAL TIPO="ALARGADO">.

Certo: Ex: O incêndio foi na <LOCAL TIPO="ALARGADO">Rua do Padrão</LOCAL>.
Errado: O incêndio foi na <LOCAL TIPO="CORREIO">Rua do Padrão</LOCAL>.
Certo: Eu deixei o carro na <LOCAL TIPO="ALARGADO">Praça da Alegria</LOCAL>.
Certo: Eu moro na <LOCAL TIPO="CORREIO">Praça da Alegria, nº 7</LOCAL>.
Errado: Eu moro na <LOCAL TIPO="ALARGADO">Praça da Alegria, nº 7</LOCAL>.

Diferença entre GEOGRAFICO e ADMINISTRATIVO

Fazemos uma diferença clara entre acidentes geográficos (naturais, objecto de estudo da geografia física) e localizações de geografia humana. Amazónia é um local GEOGRAFICO, Brasil é um local ADMINISTRATIVO. Nos casos em que existe uma coincidência exacta, como é o caso de por exemplo a Islândia, usa-se o tipo ADMINISTRATIVO.

16.8 Categoria OBRA

A categoria OBRA refere-se a qualquer coisa feita pelo Homem e que tenha um nome próprio (não comum).

16.8.1 Tipo REPRODUZIDA

Obras das quais há muitos exemplares, o nome representa o original a partir do qual se fazem as reproduções ("*Turn it on again*", "*Olhai os Lírios do Campo*", "*E Tudo o Vento Levou*", *"Sinfonia em si bemol", de Carlos Seixas*, *Bíblia*).

Certo: O álbum de música rock mais famoso é o
'<OBRA TIPO="REPRODUZIDA">Achtung Baby</OBRA>'.

16.8.2 Tipo ARTE

Obras ou objectos das quais há um exemplar único, tais como *Torre Eiffel*, *Guernica*, *Cristo-Rei*, *Capela Sistina*, *Igreja da Luz*, *Ponte da Arrábida*.

Certo: O <OBRA TIPO="ARTE">Mosteiro dos Jerónimos</OBRA> é o expoente
máximo do estilo manuelino.

'Arte' também como LOCAL

De reparar que, no caso anterior, onde a EM se refere a certos edifícios ou monumentos, o seu contexto pode ser a sua localização ou a obra em si. (Por exemplo, *Igreja da Luz* – LOCAL ou OBRA?).

```
Certo: A <OBRA TIPO="ARTE">Igreja da Luz</OBRA> tem um estilo único.
Certo: Encontramo-nos amanhã ao pé da <LOCAL TIPO="ALARGADO">Igreja
       da Luz</LOCAL>.
```

16.8.3 Tipo PUBLICACAO

Este tipo abrange obras escritas não referidas pelo nome, tais como citações de livros, artigos, decretos, directivas, entre outros. A etiqueta deve abranger todas as palavras relacionadas com a publicação, inclusivé nomes de editoras e/ou locais da publicação (Exemplos: *Maia et al. (2004)*, *Santos & Sarmento (2003:114)*, *Mota (op.cit.)*, *Decreto Lei 254/94*).

```
Certo: O <OBRA TIPO="PUBLICACAO">Decreto Lei nº 31/3 de 2005</OBRA>
       diz que isso é proibido.
Certo: Os resultados foram semelhantes aos produzidos por <OBRA
       TIPO="PUBLICACAO">(Santos et al, 2005)</OBRA>.
```

Citações a publicações no texto

O tipo `PUBLICACAO` engloba apenas produtos literários que são referidos por citações no texto. Quando se refere uma obra conhecida, é usada o tipo `REPRODUZIDA`.

```
Certo: <OBRA TIPO="REPRODUZIDA">Os Lusíadas</OBRA> descrevem
       a odisseia dos portugueses.
Certo: <OBRA TIPO="PUBLICACAO">Camões(1554)</OBRA> diz que...
```

Referências à obra ou estilo de um autor

Quando se refere a obra de um autor pelo nome do autorindexautor!nome, mencionando um estilo ou a totalidade do seu trabalho, deve-se marcar como `<ABSTRACCAO TIPO="OBRA">`, e não como `<PESSOA TIPO="INDIVIDUAL">` ou `<OBRA TIPO="PUBLICACAO">`.

```
Certo: Em <ABSTRACCAO TIPO="OBRA">Camões</ABSTRACCAO>, as musas são gregas.
```

16.9 Categoria ABSTRACCAO

Esta categoria exprime uma quantidade de ideias que são mencionadas por um nome próprio em português, que nos parecem também relevantes para um sistema de REM.

A categoria engloba áreas do conhecimento e práticas (`DISCIPLINA`), estados e funções (`ESTADO`), correntes de pensamento e facções (`ESCOLA`), planos e projectos (`PLANO`), marcas (`MARCA`), ideias abstractas (`IDEIA`) e os próprios nomes (`NOME`).

16.9.1 Tipo DISCIPLINA

Engloba disciplinas científicas, teorias, tecnologias e práticas, tais como *Inteligência Artificial*, *Neurofisiologia*, *Teoria da Relatividade*, *GSM*, *Tai-Chi*, *Futebol de 5*, *Java*.

Também inclui especialidades e áreas de governação, quando citadas como tal (pasta dos Negócios Estrangeiros, ministro/secretário de Estado do Interior).

```
Certo: O Dr. Silva foi demitido da pasta da <ABSTRACCAO
       TIPO="DISCIPLINA">Economia</ABSTRACCAO>.
Certo: Este programa foi escrito em <ABSTRACCAO
       TIPO="DISCIPLINA">Java</ABSTRACCAO>.
```

16.9.2 Tipo ESTADO

Engloba estados físicos, condições ou funções, tais como *doença de Alzheimer*, *AIDS*, *síndroma de Chang*, *Sistema Nervoso Central*. As EM de tipo `ESTADO` devem incluir os prefixos que os tornam estados no seu contexto (por exemplo, *mal de*, *estado de*, *doença de*, *síndroma de*), mesmo que apresentem minúscula.

```
Certo:  As vacas podem apanhar a <ABSTRACCAO TIPO="ESTADO">
        doença de Creutzfeldt-Jakob</ABSRACCAO>.
Errado: As vacas podem apanhar a doença de <ABSTRACCAO
         TIPO="ESTADO">Creutzfeldt-Jakob</ABSTRACCAO>.
```

16.9.3 Tipo ESCOLA

Compreende escolas, modas, facções, seitas, entre outros. Exemplos são *Barroco*, *Renascimento*, *Bushismo*, *Testemunhas de Jeová*, *Darwinismo*.

```
Certo : O <ABSTRACCAO TIPO="ESCOLA">Nazismo</ABSTRACCAO> surgiu na
        Alemanha.
```

16.9.4 Tipo MARCA

Compreende referências a marcas de produtos e raças de animais. Esta categoria pretende identificar as marcas como um conceito abstracto, como é o caso de a menção a uma marca sugerir credibilidade ou desconfiança.

```
Certo:  O meu cão é um <ABSTRACCAO TIPO="MARCA">Rotweiller</ABSTRACCAO>.
Certo:  A <ABSTRACCAO TIPO="MARCA">Vista Alegre</ABSTRACCAO> é
        reputadíssima no estrangeiro.
Certo:  Os <ABSTRACCAO TIPO="MARCA">Toyota</ABSTRACCAO> inspiram confiança.
```

Errado: O João vende <ABSTRACCAO TIPO="MARCA">Toyotas</ABSTRACCAO>.
Certo: O João vende <COISA TIPO="MEMBROCLASSE">Toyotas</COISA>.

Note-se que, no último caso, a EM *Toyota*, que foi classificada como <COISA TIPO="MEMBROCLASSE">, refere os produtos, e não a marca.

16.9.5 Tipo PLANO

Abrange medidas políticas, administrativas e/ou financeiras, assim como projectos ou acordos, que são designadas por um nome único (*Plano Marshall*, *Orçamento Larou*, *Rendimento Mínimo Garantido*).

```
Certo:  O <ABSTRACCAO TIPO="PLANO">Pacto de Varsóvia</ABSTRACCAO>
        proibiu o comércio da Polónia com o Ocidente.
```

O exemplo anterior considera *Pacto de Varsóvia* no seu contexto de acordo ou medida política. No entanto, o mesmo nome pode definir uma ORGANIZACAO ou, até, uma EFEMERIDE:

```
Certo: Os países do <ORGANIZACAO TIPO=ADMINISTRACAO>Pacto de
       Varsóvia</ORGANIZACAO> desenvolveram uma política comum.
Certo: O <ACONTECIMENTO TIPO="EFEMERIDE">Pacto de Varsóvia
       </ACONTECIMENTO> comemora 40 anos de idade.
```

16.9.6 Tipo IDEIA

As ideias ou ideais são muitas vezes EM que representam conceitos abstractos, mas que são normalmente referenciados por outros conceitos mais concretos, como é o exemplo de:

```
Certo: A honra da <ABSTRACCAO TIPO="IDEIA">França</ABSTRACCAO>
       estava em jogo.
```

Neste exemplo, o conceito abstracto é a honra, retirado a partir da referência *França*.

```
A candidatura para a <ABSTRACCAO|ORGANIZACAO TIPO="IDEIA|ADMINISTRACAO">
Presidência da República</ABSTRACCAO|ORGANIZACAO>
```

Neste caso, pode-se interpretar *Presidência da República* (note-se que não é um CARGO, CARGO seria *Presidente da República*) como uma ORGANIZACAO, mas também pode-se interpretar a EM como uma referência a um órgão de poder, um conceito mais abstracto do que a ORGANIZACAO.

16.9.7 Tipo NOME

Por vezes uma dada EM está a representar apenas o nome, e como tal deve ser identificada como um NOME.

```
Certo: Achei um cão. Vou dar-lhe o nome de <ABSTRACCAO TIPO="NOME">
       Bobi</ABSTRACCAO>.
Certo: O magnata criou uma empresa chamada <ABSTRACCAO TIPO="NOME">
       Cauca7</ABSTRACCAO>.
```

16.9.8 Tipo OBRA

Quando a referência a um autor pressupõe um estilo artístico ou o seu trabalho artístico, deve ser de categoria <ABSTRACCAO TIPO="OBRA">:

```
Certo: Em <ABSTRACCAO TIPO="OBRA">Camões</ABSTRACCAO>, as musas são gregas.
```

16.10 Categoria VALOR

Valores, como o nome indica, podem referir-se a quantidades absolutas ou relativas (QUANTIDADE), designar dinheiro (MOEDA) ou classificações desportivas, ordinais normais e outras (CLASSIFICACAO). Os itens numéricos a marcar ordem no texto não são considerados EM.

Quando há uma referência a um intervalo de valores, os seus limites devem ser etiquetados como duas EM distintas, e não como uma única EM, como ilustra o caso abaixo:

```
Certo:  Entre <VALOR TIPO="QUANTIDADE">7</VALOR> a
        <VALOR TIPO="QUANTIDADE">10 metros</VALOR>.
Errado: Entre <VALOR TIPO="QUANTIDADE">7 a 10 metros</VALOR>.
```

16.10.1 Tipo CLASSIFICACAO

Engloba valores que traduzem classificação, ordenação ou pontuação (Exemplos: *2-0*, *15'*, *3ª*). Enumerações de parágrafos, tópicos e outras secções não devem ser etiquetados.

Tempos como medida de classificação

No exemplo anterior, a EM *15'* só é uma classificação quando designa um tempo pelo qual se mede uma dada competição:

```
Certo:  1º lugar - Ferrari, com o tempo de <VALOR
        TIPO="CLASSIFICACAO">3' 57''</VALOR>.
Errado: O golo foi apontado aos <VALOR TIPO="CLASSIFICACAO">14'</VALOR>
```

por Deco.

Certo: O golo foi apontado aos <VALOR TIPO="QUANTIDADE">14'</VALOR> por Deco.

Números como pontuação

Muitas vezes, os números também pertencem a classificações, e como tal, devem ser etiquetados como tal.

Certo: Classificação: <VALOR TIPO="CLASSIFICACAO">1º</VALOR> FC Porto, <VALOR TIPO="CLASSIFICACAO">89</VALOR> pontos.

Números ordinais de eventos organizados

Expressões numéricas incluídas no nome de um evento ou de um cargo não são de categoria VALOR, como ilustram os seguintes exemplos:

Certo: Vai abrir a <ACONTECIMENTO>6ª Exposição Mundial de Cinema <ACONTECIMENTO>.

Errado: Vai abrir a <ACONTECIMENTO><VALOR TIPO="CLASSIFICACAO">6ª</VALOR> Exposição Mundial de Cinema<ACONTECIMENTO>.

Certo: <PESSOA TIPO="CARGO" MORF="M,S">33º Governador da Califórnia</PESSOA>.

Errado: <VALOR TIPO="CLASSIFICACAO">33º</VALOR> <PESSOA TIPO="CARGO" MORF="M,S">Governador da Califórnia</PESSOA>.

Graus escolares e académicos

Classificações referentes a anos escolares não devem ser etiquetados. Esta norma estende-se a graus académicos (*Mestrado*, *Licenciatura*, etc).

Certo: Reprovei na 4ª classe.

Errado: Reprovei na <VALOR TIPO="CLASSIFICACAO">4ª</VALOR> classe.

Certo: Tenho Mestrado em <ABSTRACCAO TIPO="DISCIPLINA">Pecuária </ABSTRACCAO>.

16.10.2 Tipo MOEDA

Abrange valores monetários (Exemplos: *300$00*, *$US 15*, *£39*, *Cr 500*, *50 contos*, *30 milhões de cruzeiros*). A etiqueta deve abranger a unidade monetária, mesmo que esta esteja em minúsculas.

Certo: O carro custou-me <VALOR TIPO="MOEDA">20000 euros</VALOR>.

Errado: O carro custou-me <VALOR TIPO="MOEDA">20000</VALOR> euros.

16.10.3 Tipo QUANTIDADE

Engloba percentagens, números soltos, e, caso uma quantidade tenha unidades, a própria unidade. Não engloba unidades monetárias, já abrangidas pelo tipo MOEDA. Exemplos: *15 m*, *30 kg*, *50 mm*, *1,4 kHz*, *27º C*, *23%*, *2.500*, *pH 2,5*.

Por unidades entendem-se as usadas para medir propriedades como distância, tempo, luz, área, volume, peso, massa, etc, e não objectos que sejam contados. Como tal, em *6 pessoas* ou *9 folhas de papel*, pessoas e folhas de papel não são para ser incluídas na etiqueta.

As unidades podem ser complexas, como em *23 metros quadrados*, *9m x 6m*, *3 Bar*, *4 quilogramas por metro cúbico*. Toda a especificação da quantidade é para ser incluída na EM.

Embora exista uma categoria TEMPO, esta não abrange evidentemente referências a quantidades de tempo, como ilustra o seguinte exemplo:

```
Certo:  Eu tenho <VALOR TIPO="QUANTIDADE">19 anos</VALOR>.
Errado: Eu tenho <VALOR TIPO="QUANTIDADE">19</VALOR> anos.
```

16.11 Categoria VARIADO

Esta etiqueta deve abranger outras referências que são relevantes e que cumpram as regras definidas acima para serem consideradas EM, mas que não são abrangidas nas outras categorias.

Exemplos (não exaustivos) são prémios, fenómenos naturais ou papéis de teatro (*prémio Valmor*, *tufão El Niño*, *voo 714*, *Rei Lear*).

Para simplificar a forma de processar a classificação semântica no HAREM, a categoria VARIADO deve ser obrigatoriamente expressa com o tipo OUTRO.

```
Certo: Eu recebi o <VARIADO TIPO="OUTRO">Prémio Camões</VARIADO> o
       ano passado.
```

Capítulo 17

Directivas para a identificação e classificação morfológica na colecção dourada do HAREM

Nuno Cardoso, Diana Santos e Rui Vilela

Este capítulo foi previamente publicado como Relatório Técnico DI/FCUL TR–06–19, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 17, p. 239–244, 2007.

Neste documento, apresentamos as directivas usadas na tarefa de classificação morfológica da colecção dourada do HAREM e, consequentemente, qual o comportamento esperado pelos sistemas que participem na tarefa. No capítulo 16 foi indicada a metodologia seguida na classificação semântica.

> O texto deste capítulo é exactamente idêntico ao do capítulo anterior, por isso foi aqui omitido. A razão dessa replicação deve-se ao facto de ter sido possível, no HAREM, de participar exclusivamente numa das duas tarefas de classificação. Como tal, e para que os participantes numa e não noutra tarefa não tivessem que ler as duas directivas, optou-se por repetir a secção respectiva às directivas de identificação, que neste capítulo, foi omitida.

17.1 Regras gerais da tarefa de classificação morfológica

Considerámos como passíveis de ser classificadas morfologicamente (isto é, EM que devem ter o atributo `MORF`):

- As categorias `PESSOA`, `ORGANIZACAO`, `COISA`, `ABSTRACCAO`, `ACONTECIMENTO`, `OBRA`, e `VARIADO` na sua totalidade.
- Na categoria `LOCAL`, os tipos `ADMINISTRATIVO` e `GEOGRAFICO`.
- Na categoria `TEMPO`, o tipo `CICLICO`.

As seguintes EM não têm atributo `MORF`:

- A categoria `VALOR` na sua totalidade.
- Na categoria `LOCAL`, os tipos `CORREIO`.
- Na categoria `TEMPO`, o tipo `HORA`.

E finalmente, nos seguintes casos as EM podem ou não ter o atributo `MORF`:

- Na categoria `LOCAL`, o tipo `VIRTUAL`.
- Na categoria `TEMPO`, os tipos `DATA` e `PERIODO`.

Uma série de exemplos de aplicação são apresentados posteriormente para clarificar em que situações ocorrem estas excepções.

17.1.1 Género (morfológico)

Consideramos que o género de uma EM pode ter três valores:

M: EM com género masculino.

F: EM com género feminino.

?: Para os casos em que o género é indefinido.

17.1.2 Número

Consideramos que o número de uma EM pode ter três valores:

S: EM no singular.

P: EM no plural.

?: Para os casos em que o número é indefinido.

17.1.3 Exemplos de não atribuição de `MORF` na categoria `LOCAL`

Em alguns casos particulares do tipo `VIRTUAL`, o atributo `MORF` foi omitido, devido ao facto de não ser possível avaliar morfologicamente números de telefone.

```
Certo: <LOCAL TIPO="VIRTUAL">(48) 281 9595</LOCAL>
```

Os casos que possuam a etiqueta `MORF` são, pelo contrário, geralmente casos em que a entidade é de outro tipo básico, mas é empregue no contexto na acepção de `LOCAL`.

```
Certo: Como capturar da <LOCAL TIPO="VIRTUAL" MORF="F,S">Internet</LOCAL>...
Certo: uma ordem do governo local publicada na "<LOCAL TIPO="VIRTUAL"
       MORF="F,S">Gazeta de Macau</LOCAL>" ordenava...
Certo: E só depois da publicação no '<LOCAL TIPO="VIRTUAL" MORF="M,S">
       Diário da República</LOCAL>' é que tomou-se conhecimento do traçado.
```

17.1.4 Exemplos de não atribuição de `MORF` na categoria `TEMPO`

Nos tipos `PERIODO` e `DATA` há casos distintos em que são aplicados o atributo `MORF`.

As datas especificadas em termos de anos ou de dias não possuem nunca a etiqueta `MORF`.

```
Certo: Este ano de <TEMPO TIPO="PERIODO">1982</TEMPO> deve...
Certo: <TEMPO TIPO="PERIODO">1914-1918</TEMPO>...
Certo: ia ser a <TEMPO TIPO="DATA">17 de Dezembro</TEMPO> porque saiu...
Certo: Em <TEMPO|TEMPO TIPO="DATA|PERIODO">91</TEMPO>, foram angariados...
```

As classificações que possuem atributo MORF são meses, séculos, e períodos históricos .

```
Certo: Cinema para o mês de <TEMPO TIPO="PERIODO" MORF="M,S">
       Maio</TEMPO>.
Certo: Mas já vem do <TEMPO TIPO="DATA" MORF="M,S">século XVI
       </TEMPO> o feriado.
Certo: os povoadores cristãos da <TEMPO|ACONTECIMENTO TIPO="PERIODO
       |EFEMERIDE" MORF="F,S">Reconquista</TEMPO|ACONTECIMENTO>.
Certo: Nesta <TEMPO TIPO="PERIODO" MORF="F,S">Primavera</TEMPO>,
       encontrei-me com os meus amigos.
Certo: está agora previsto para <TEMPO TIPO="DATA" MORF="M,S">Outubro
       </TEMPO> ou <TEMPO TIPO="DATA" MORF="M,S">Novembro</TEMPO>
```

17.2 Regras de atribuição de classificação morfológica

Considera-se o contexto e o texto adjacente para determinar o género e o número de uma dada EM, que à partida pode não ter género ou número definido.

Quando nem esse contexto nem o conhecimento lexical dos anotadores permite atribuir valores definidos, usa-se o valor '?', não especificado.

Exemplos:

```
Certo: O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">João</PESSOA>
       é um professor.
Certo: A <PESSOA TIPO="INDIVIDUAL" MORF="F,S">João</PESSOA>
       não veio.
Certo: O apelido <ABSTRACCAO TIPO="NOME" MORF="?,S">João
       </ABSTRACCAO> é muito raro.
```

Ou seja, o nome *João* tem diferentes interpretações da sua classificação morfológica, consoante o contexto em que se encontra inserido.

17.2.1 Exemplos na categoria LOCAL

Algumas localidades administrativas são precedidas por artigo, determinando assim o género e número da entidade que designam (*o Porto*, *a Madeira*, *o Brasil*, *a Guarda*, *o Minho*, *o Rio Grande do Sul*, *os Estados Unidos*). Contudo, muitas outras não levam artigo e torna-se mais difícil de atribuir uma classificação morfológica.

Pareceu-nos em alguns casos haver consenso, tal como para *Portugal* (M, S), *Lisboa* (F, S), *Bragança* (F, S), *Brasília* (F, S), *Nova Iorque* F, S) e *Colónia* (F, S), mas noutros casos apenas pudemos usar '?' no género, tal como em *Chaves*, *São Paulo* (estado ou cidade), *Castelo Branco*, *Braga* ou *Madrid*, excepto quando tal é especificado no contexto.

```
Certo: <LOCAL TIPO=ADMINISTRATIVO MORF="F,S">Leiria</LOCAL> é linda.
Certo: do concelho de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">
       Aregos</LOCAL>.
Certo: todo o noroeste(de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">
       Resende</LOCAL> ao...
Certo: em <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL>
       seria...
Certo: ...aqui em <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">São Paulo
       </LOCAL>.
Certo: ...em <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Nova Iorque
       </LOCAL> e saímos...
Certo: ...polícia de <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Colónia
       </LOCAL> foram suspensos...
```

17.2.2 Exemplos na categoria `ORGANIZACAO`

Geralmente o número e género de uma organização são definidos pelo número e género da primeira palavra do nome, *Charcutaria Brasil* (`F,S`), *Armazéns do Chiado* (`M,P`), *Banco X* (`M,S`) ou *Caixa Y* (`F,S`), enquanto empresas internacionais têm geralmente associado o género feminino: A *Coca-Cola*, *a Benetton*, *a IBM*, *a Microsoft*, *a Sun*, *a Lotus*, *a Ferrari*, etc.

```
Certo: junto do <ORGANIZACAO TIPO="EMPRESA" MORF="M,S">Banco Sotto
       Mayor</ORGANIZACAO>.
Certo: Uma acção da <ORGANIZACAO TIPO="EMPRESA" MORF="F,S">Cartier
       </ORGANIZACAO>.
Certo: A acção da <ORGANIZACAO TIPO="EMPRESA" MORF="F,S">Portugal
       Telecom</ORGANIZACAO> resultou...
Certo: Esta página tem o apoio da <ORGANIZACAO TIPO="EMPRESA"
       MORF="F,S">IP</ORGANIZACAO>.
```

17.2.3 Exemplos na categoria `PESSOA`

No caso de `GRUPOMEMBRO`, ou seja, grupos de pessoas, o número é geralmente plural, e o género depende do sexo dos membros. *As Doce*, *os ABBA*, *os Xutos e Pontapés*, *os Beatles*, *as Spice Girls*, *os GNR*...

```
Certo: os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Stones</PESSOA>
Certo: e antes dos <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">R.E.M.</PESSOA>
Certo: <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Peruanos</PESSOA>
       com diamantes falsos.
Certo: depois os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Mouros</PESSOA> que
```

```
       lhe deram o nome...
Certo: ...dez minutos o <PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Bastia
       </PESSOA>assegurou a presença na final...
```

17.2.4 Exemplos na categoria ACONTECIMENTO

No caso do tipo EVENTO, os acontecimentos desportivos que tenham duas equipas, o número é singular, e o género é masculino, visto que correspondem a um jogo.

```
Certo: seguintes jogos: <ACONTECIMENTO TIPO="EVENTO" MORF="M,S">
       Penafiel-Rio Ave</ACONTECIMENTO>
Certo: e o <ACONTECIMENTO TIPO="EVENTO" MORF="M,S">
       Nacional-Académica</ACONTECIMENTO>
```

17.2.5 Exemplos na categoria ABSTRACCAO

No caso do tipo DISCIPLINA, a maior parte das EM que se refiram a disciplinas na área da educação tem género feminino, o número pode variar consoante o primeiro átomo.

```
Certo: e <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">Filosofia</ABSTRACCAO>
       em todas as universidades.
Certo: <ABSTRACCAO TIPO= DISCIPLINA MORF= F,S >Ciência da Informação
       </ABSTRACCAO>.
Certo: futuros professores de <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">
       Educação Física</ABSTRACCAO>.
Certo: As <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,P">TI</ABSTRACCAO> são
       uma ferramenta...
```

Já em relação a desportos, o género é em geral masculino, embora haja alguns que, por serem originários de palavras portuguesas femininas, mantêm o género, tal como *Vela* ou *Luta livre*.

```
Certo: Página do time de <ABSTRACCAO TIPO="DISCIPLINA" MORF="M,S">
       Handebol</ABSTRACCAO>
```

Capítulo 18

Avaliação no HAREM: métodos e medidas

Diana Santos, Nuno Cardoso e Nuno Seco

Este capítulo foi previamente publicado como Relatório Técnico DI/FCUL TR–06–17, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 18, p. 245–282, 2007.

As directivas de avaliação descritas neste relatório técnico representam o conjunto de pontuações, regras, medidas e métricas usadas para medir e comparar as saídas dos sistemas de REM em relação às colecções douradas. O *software* de avaliação do HAREM, descrito ao detalhe no capítulo 19, implementa as directivas aqui expostas.

18.1 Terminologia

18.1.1 Pontuações

As *pontuações* são os valores atribuídos a cada EM marcada pelo sistema, após uma comparação com a respectiva marcação na CD. Cada tarefa possui as suas próprias pontuações, que são calculadas segundo um conjunto de *regras* que serão descritas e ilustradas com exemplos nas respectivas secções.

Um exemplo simples de pontuação: se um sistema identificar uma determinada EM tal como está na CD, obtém a pontuação `correcto` para a tarefa de identificação. A pontuação `correcto`, segundo as regras para a tarefa de identificação, corresponde a um valor igual a 1.

18.1.2 Medidas

As *medidas* representam formas de combinação das várias pontuações obtidas em cada tarefa. Assim, é possível representar diferentes componentes da avaliação, para cada saída.

As medidas são implementadas na avaliação das tarefas de classificação morfológica e de semântica, onde existe mais do que um parâmetro pontuável (no caso da morfologia, o género e o número; no caso da semântica, a categoria e o tipo).

Um exemplo de medidas: na tarefa de classificação semântica, uma EM é avaliada segundo a sua categoria e tipo. Assim, são geradas duas pontuações, uma relacionada com a categoria, e outra com o tipo. A combinação destas pontuações num único valor depende da *medida* usada.

18.1.3 Métricas

As *métricas* são formas de representar o desempenho dos sistemas em valores numéricos, de acordo com a marcação que fez para um dado grupo de EM.

Precisão: a precisão afere a "qualidade" da resposta do sistema, ao calcular a proporção de respostas correctas em relação a todas as respostas realizadas por este.

Abrangência: a abrangência afere a "quantidade" da resposta do sistema, ao calcular a proporção de respostas correctas em relação ao universo de possíveis respostas (no caso presente, as EM contidas na colecção dourada).

Medida F: A medida F combina as métricas de precisão e de abrangência para cada tarefa, de acordo com a seguinte fórmula:

$$\text{Medida F} = \frac{2 \times \text{precisão} \times \text{abrangência}}{\text{precisão} + \text{abrangência}}$$

Esta métrica é igual para todas as tarefas de avaliação.

Sobre-geração: a sobre-geração afere o excesso de resultados que um sistema produz, ou seja, calcula quantas vezes produz resultados espúrios.

Sub-geração: a sub-geração afere a quantidade de resultados que um sistema se esqueceu em analisar, ou seja, calcula quantas vezes produz resultados em falta, dada a solução conhecida (a CD).

Erro Combinado: o erro combinado reúne as métricas de sobre-geração e de sub-geração numa única métrica, de acordo com a seguinte fórmula:

$$\text{Erro combinado} = \frac{\sum \text{pontuações em falta} + \sum \text{pontuações espúrio} + \sum \text{factor de erro}}{\sum \text{Pontuação máx. sistema} \cup \text{Pontuação máx. CD}}$$

O factor de erro é calculado pela equação 18.2, apresentada mais à frente.

18.1.4 Cenários de avaliação

Os sistemas de REM são desenvolvidos para diferentes propósitos. Como tal, as directivas de avaliação prevêem a realização de avaliações segundo *cenários*, de forma a ajustar a avaliação às características de cada sistema de REM. O módulo de *software* responsável pela criação de cenários é o Véus, que se encontra detalhado no capítulo 19.

A avaliação do HAREM realizou-se segundo dois eixos de cenários:

Cenário absoluto–relativo: O cenário absoluto avalia o desempenho do sistema em relação à totalidade das EM na CD para a tarefa de REM completa, ou seja, a identificação e a classificação de EM. O cenário relativo, por seu lado, restringe a avaliação às EM pontuadas como correcto ou parcialmente correcto na tarefa de identificação. Este cenário permite avaliar o desempenho do sistema apenas na tarefa de classificação (semântica ou morfológica), independentemente do desempenho na tarefa de identificação.

Cenário total–selectivo O cenário total abrange todas as categorias de EM da CD, avaliando a tarefa de classificação (morfológica ou semântica) em relação à tarefa tal como foi proposta pelo HAREM. No cenário selectivo, o participante escolhe previamente um sub-conjunto de categorias e de tipos da categorização HAREM que o seu sistema

consegue processar. Assim, a tarefa da classificação (morfológica ou semântica) é avaliada segundo esse sub-conjunto de categorias e de tipos.

Tarefa de identificação

A tarefa de identificação é avaliada segundo o eixo de cenário total–selectivo:

Cenário de identificação total: considera para efeitos de pontuação todas as etiquetas na CD.

Cenário de identificação selectivo: considera apenas para efeitos de pontuação o leque de categorias semânticas que o sistema participante se propõe explicitamente identificar.

Tarefas de classificação

As tarefas de classificação (morfológica e semântica) são avaliadas segundo os dois eixos de cenários:

Total: considera todas as EM existentes na CD.

- **Absoluto:** considera todas as EM, incluindo as que não foram identificadas com pontuação correcta ou parcialmente correcta.
- **Relativo:** considera apenas as EM identificadas com pontuação correcta ou parcialmente correcta.

Selectivo: considera apenas as EM na CD de categorias/tipos que o participante se propôs classificar.

- **Absoluto:** considera todas as EM, incluindo as que não foram identificadas com pontuação correcta ou parcialmente correcta.
- **Relativo:** considera apenas as EM identificadas com pontuação correcta ou parcialmente correcta. correctas.

18.2 Tarefa de identificação

A avaliação da tarefa de identificação tem por objectivo medir a eficiência dos sistemas em delimitar correctamente os *termos* que compõem as EM na colecção, comparativamente com a CD).

Um termo é definido no HAREM como sendo qualquer sequência de letras (e somente letras) ou dígitos individuais. As preposições e conjunções são contabilizadas para efeitos de pontuação, mas não são considerados para efeitos de alinhamento. No capítulo 19, secção 19.2.3, está disponível uma lista das palavras que o AlinhEM, o módulo de *software* que realiza os alinhamentos, ignora, assim como as regras de atomização.

18.2.1 Pontuaçções

A avaliação do HAREM atribui a seguinte pontuação para a tarefa de identificação:

Correcto: quando o termo inicial e o termo final da EM são iguais na saída e na CD, e o número de termos da EM é o mesmo nas duas listas.

Parcialmente Correcto (por defeito): quando pelo menos um termo da saída do sistema corresponde a um termo de uma EM na CD, e o número total de termos da EM do sistema é menor do que o número de termos da respectiva EM da CD.

Parcialmente Correcto (por excesso): quando pelo menos um termo da saída do sistema corresponde a um termo de uma EM na CD, e o número total de termos da EM do sistema é maior do que o número de termos da respectiva EM da CD.

Em Falta: quando a saída do sistema falha em delimitar correctamente qualquer termo de uma EM da CD.

Espúrio: quando a saída do sistema delimita uma alegada EM que não consta na CD.

Às EM pontuadas como `correcto` é atribuído um valor igual a 1. As EM pontuadas como `parcialmente correcto` é atribuído o valor calculado pela equação 18.1:

$$p = 0,5\frac{n_c}{n_d} \tag{18.1}$$

Onde:

n_c representa o número de termos comuns entre a EM do sistema e a EM da CD, ou seja, a cardinalidade da intersecção dos termos.

n_d representa o número de termos distintos entre a EM do sistema e a EM da CD, ou seja, a cardinalidade da reunião dos termos.

O *factor de erro*, usado no cálculo da métrica *Erro Combinado*, é dado pela equação 18.2:

$$p = 1 - 0,5\frac{n_c}{n_d} \tag{18.2}$$

18.2.2 Métricas

Para a tarefa de identificação, as métricas são calculadas da seguinte forma:

Precisão

Na tarefa de identificação, a precisão calcula o teor de EM correctas e parcialmente correctas em todas as EM identificadas pelo sistema. Os valores para as EM pontuadas como parcialmente correctas são calculados pela equação 18.1.

$$\text{Precisão}_{\text{identificação}} = (\sum \text{EM correctas} + \sum \text{EM parcialmente correctas}) / (\sum \text{EM identificadas pelo sistema})$$

Abrangência

Na tarefa de identificação, a abrangência calcula o teor de EM contidas na CD que o sistema conseguiu identificar. Os valores para as EM pontuadas como parcialmente correctas são calculados pela equação 18.1.

$$\text{Abrangência}_{\text{identificação}} = (\sum \text{EM correctas} + \sum \text{EM parcialmente correctas}) / (\sum \text{EM na CD})$$

Sobre-geração

Na tarefa de identificação, a sobre-geração calcula o teor de EM que foram identificadas pelo sistema, mas que não existem na CD.

$$\text{Sobre-geração}_{\text{identificação}} = (\sum \text{EM espúrias} / \sum \text{EM identificadas pelo sistema})$$

Sub-geração

Na tarefa de identificação, a sub-geração calcula o teor de EM que existem na colecção dourada, mas que não foram identificadas pelo sistema.

$$\text{Sub-geração}_{\text{identificação}} = (\sum \text{EM em falta} / \sum \text{EM na CD})$$

18.2.3 Exemplo detalhado de atribuição de pontuação

Apresentemos uma frase hipotética da colecção dourada:

> Terminou ontem no **<LOCAL TIPO="ALARGADO">** **Laboratório Nacional de Engenharia Civil** **</LOCAL>**, em **<LOCAL TIPO="ADMINISTRATIVO">** **Lisboa** **</LOCAL>**, o **<ACONTECIMENTO TIPO="EVENTO">** **Encontro de Reflexão** **</ACONTECIMENTO>** sobre a concretização do **<ABSTRACCAO TIPO="PLANO">** **Plano Hidrológico** **</ABSTRACCAO>** espanhol.

Imaginemos a seguinte saída do sistema que pretendemos avaliar:

```
<PESSOA TIPO="INDIVIDUAL">Terminou</PESSOA> ontem no
<LOCAL TIPO="ALARGADO">Laboratório Nacional</LOCAL>
de <ABSTRACCAO TIPO="DISCIPLINA">Engenharia
Civil</ABSTRACCAO>, em <LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL>,
o Encontro de Reflexão sobre a concretização do <ABSTRACCAO
TIPO="PLANO">Plano Hidrológico espanhol</ABSTRACCAO>.
```

A Tabela 18.1 apresenta a pontuação pormenorizada, caso a caso, e na Tabela 18.2 os valores das métricas para a tarefa de identificação. A Tabela 18.3 apresenta 7 casos particulares com uma anotação (hipotética) feita por um sistema e na CD, e a Tabela 18.4 ilustra as regras de pontuação para esses casos, para a tarefa de identificação.

18.2.4 Identificações alternativas

No caso de considerarmos que há mais do que uma delimitação correcta na tarefa em questão, levando à identificação de uma ou mais EM alternativas, foi usada a etiqueta <ALT> para assinalar as várias opções na CD. Como tal, o avaliador do HAREM irá comparar a CD com a saída do sistema e optar pela melhor alternativa. A escolha é feita segundo o seguinte algoritmo:

1º – Melhor medida F para cada caso.

2º – Menor valor de erro combinado.

3º – Maior número de alinhamentos.

Para auxiliar na selecção da opção <ALT> nos casos mais difíceis, tais como alternativas sem EM, os programas de selecção de <ALT> para as tarefas do HAREM (*ALTinaID*, *ALTinaSEM* e *ALTinaMOR*, ver capítulo 19) introduzem no cálculo um alinhamento correcto em cada alternativa considerada. Tal introdução não prejudica a selecção, e evita que alternativas sem EM tenham uma medida F não definida (ou seja, zero no numerador e no denominador).

No Caítulo 19 explica-se em detalhe este processo do ALTinaID, e as Tabelas 18.5 a 18.9 ilustram como é feito esse cálculo, para o seguinte exemplo com três alternativas:

```
<ALT> <EM> Governo PSD de Cavaco Silva </EM>
<EM> Governo PSD </EM> de <EM> Cavaco Silva </EM>
Governo PSD de Cavaco Silva </ALT>
```

ALT1: Governo PSD de Cavaco Silva

Caso	Colecção dourada	Saída do sistema	Pontuação
1	-	Terminou	0 (Espúrio)
2	Laboratório Nacional de Engenharia Civil	Laboratório Nacional	$0,5 \times \frac{2}{5}$=0,2 (Parcialmente Correcto por Defeito)
3	Laboratório Nacional de Engenharia Civil	Engenharia Civil	$0,5 \times \frac{2}{5}$=0,2 (Parcialmente Correcto por Defeito)
4	Lisboa	Lisboa	1 (Correcto)
5	Encontro de Reflexão	-	0 (Em Falta)
6	Plano Hidrológico	Plano Hidrológico espanhol	$0,5 \times \frac{2}{3}$ = 0,333 (Parcialmente Correcto Por Excesso)

Tabela 18.1: Pontuação da tarefa de identificação, para o exemplo dado.

Métrica	Valor
Precisão	$\frac{1+0,2+0,2+0,333}{5} = 34,7\%$
Abrangência	$\frac{1+0,2+0,2+0,333}{4} = 43,3\%$
Medida F	$\frac{2 \times 0,347 \times 0,433}{0,347+0,433} = 0,385$
Sobre-geração	$\frac{1}{5} = 20\%$
Sub-geração	$\frac{1}{4} = 25\%$
Erro Combinado	$\frac{(1-0,2)+(1-0,2)+(1-0,333)+1+1}{6} = 71,1\%$

Tabela 18.2: Métricas da tarefa de identificação, para o exemplo dado.

Caso	Sistema participante	Colecção dourada
1	o novo presidente do CNPq, Evando Mirra	o novo presidente do CNPq, Evando Mirra
2	a partir de 1991	a partir de 1991
3	Graduou-se em Engenharia Mecanica e Elétrica	Graduou-se em Engenharia Mecanica e Elétrica
4	Rua 13 de Maio, 733 - Bela Vista-(11) 3262 3256	Rua 13 de Maio, 733 - Bela Vista - (11) 3262 3256
5	Senhores Comandantes das F-FDTL e da PNTL	Senhores Comandantes das F-FDTL e da PNTL
6	secretário-geral do Partido Revolucionário Institucional	secretário-geral do Partido Revolucionário Institucional
7	Estúdio da Oficina Cultural Oswald de Andrade	Estúdio da Oficina Cultural Oswald de Andrade
	São Paulo , 21 de novembro de 1994	São Paulo, 21 de novembro de 1994

Tabela 18.3: Lista de exemplos para ilustração da pontuação da tarefa de identificação.

Caso	Etiquetas	Pontuação	Termos	Total
1	**Saída:** presidente do CNPq, Evando **CD:** CNPq	$0,5 \times \frac{1}{4}$	n_c**:** CPNq n_d:presidente, do, CPNq, Evando	0,225
	Saída: presidente do CNPq, Evando **CD:** Evando Mirra	$0,5 \times \frac{1}{5}$	n_c: Evando n_d: presidente, do, CPNq, Evando, Mirra	
2	**Saída:** 991 **CD:** 1991	$0,5 \times \frac{3}{4}$	n_c: 9, 9, 1 n_d: 1, 9, 9, 1	0,375
3	**Saída:** Engenharia Mecânica **CD:** Engenharia Mecanica e Eléctrica	$0,5 \times \frac{2}{4}$	n_c: Engenharia, Mecânica n_d: Engenharia, Mecânica, e, Eléctrica	0,375
	Saída: Eléctrica **CD:** Engenharia Mecanica e Eléctrica	$0,5 \times \frac{1}{4}$	n_c: Eléctrica n_d:Engenharia, Mecânica, e, Elétrica	
4	**Saída:** Rua **CD:** Rua 13 de Maio, 733 - Bela Vista	$0,5 \times \frac{1}{10}$	n_c: Rua n_d:Rua, 1, 3, de, Maio, 7, 3, 3, Bela, Vista	1,35
	Saída: 13 de Maio **CD:** Rua 13 de Maio, 733 - Bela Vista	$0,5 \times \frac{4}{10}$	n_c: 1, 3, de, Maio n_d: Rua, 1, 3, de, Maio, 7, 3, 3, Bela, Vista	
	Saída: Bela Vista **CD:** Rua 13 de Maio, 733 - Bela Vista	$0,5 \times \frac{2}{10}$	n_c: Bela, Vista n_d: Rua,1, 3, de, Maio, 7, 3, 3, Bela, Vista	
	Saída: (11) 3262 3256 **CD:** (11) 3262 3256	1		
5	**Saída:** Senhores Comandantes das F- **CD:** Senhores Comandantes das F-FDTL e da PNTL	$0,5 \times \frac{4}{6}$	n_c: Senhores, Comandantes, das, F n_d: Senhores, Comandantes,das, F-, FDTL, PNTL	0,5
	Saída: FDTL **CD:** Senhores Comandantes das F-FDTL e da PNTL	$0,5 \times \frac{1}{6}$	n_c: FDTL n_d: Senhores, Comandantes, das, F-, FDTL, PNTL	
	Saída: PNTL **CD:** Senhores Comandantes das F-FDTL e da PNTL	$0,5 \times \frac{1}{6}$	n_c: PNTL n_d: Senhores, Comandantes, das, F-, FDTL, PNTL	
6	**Saída:** Partido Revolucionrio Institucional **CD:** secretário-geral do Partido Revolucionário Institucional	$0,5 \times \frac{3}{6}$	n_c: Partido, Revolucionário, Institucional n_d: secretário, geral, do, Partido, Revolucionário, Institucional	0,25
7	**Saída:** Oficina Cultural Oswald de Andrade **CD:** Estúdio da Oficina Cultural Oswald de Andrade	$0,5 \times \frac{5}{6}$	n_c: Oficina, Cultural, Oswald, de, Andrade n_d: Estúdio, Oficina, Cultural, Oswald, de, Andrade	1,11(1)
	Saída: São Paulo , 21 **CD:** São Paulo	$0,5 \times \frac{2}{4}$	n_c: São, Paulo n_d: São, Paulo, 2, 1	
	Saída: São Paulo, 21 **CD:** 21 de novembro de 1994	$0,5 \times \frac{2}{9}$	n_c: 2, 1 n_d: 2, 1, de, Novembro, de, 1, 9, 9, 4	
	Saída: novembro de 1994 **CD:** 21 de novembro de 1994	$0,5 \times \frac{6}{9}$	n_c: Novembro, de, 1, 9, 9, 4 n_d: 2, 1, de, Novembro, de, 1, 9, 9, 4	

Tabela 18.4: Pontuação na tarefa de identificação, para os exemplos da tabela 18.3.

ALT2: Governo PSD de Cavaco Silva

ALT3: Governo PSD de Cavaco Silva

O avaliador irá escolhar a alternativa que produz melhores resultados. A Tabela 18.5 apresenta vários exemplos de saídas de sistema (as células a negrito indicam a alternativa escolhida) e, para cada caso, a pontuação individual. Os valores da medida F e do erro combinado são calculados nas Tabelas 18.6 a 18.9, que se referem respectivamente à precisão, à abrangência, à medida F e ao erro combinado, escolhendo para cada caso qual das alternativas <ALT> será seleccionada. Como mencionado antes, as medidas nas Tabelas 18.6 a 18.9 são calculadas adicionando para cada alternativa um elemento correcto.

Caso	Sada do sistema	ALT1	ALT2	ALT3
1	<EM>Governo PSD de Cavaco Silva</EM>	**1 Correcto** **Medida F: 1** **Erro Combinado: 0%**	2 Parc. Correcto Medida F: 0,56 Erro Combinado: 53,3%	1 Espúrio Medida F: 0,67 Erro Combinado: 50,0%
2	Governo <EM>PSD de Cavaco Silva</EM>	1 Em Falta Medida F: 0,67 Erro Combinado: 50,0%	2 Em Falta Medida F: 0,5 Erro Combinado: 66,7%	**Sem pontuação** **Medida F: 1** **Erro Combinado: 0%**
3	Governo <EM>PSD de Cavaco Silva</EM>	**1 Parc.Cor. por Def.** **Medida F: 0,7** **Erro Combinado: 30%**	2 Parc.Cor. por Exc. Medida F: 0,54 Erro Combinado: 55%	1 Espúrio Medida F: 0,67 Erro Combinado: 50%
4	<EM>Governo</EM> <EM> PSD</EM> de Cavaco Silva	2 Parc. Correcto Medida F: 0,48 Erro Combinado: 60%	**2 Parc.Cor.+1 Em Falta** **Medida F: 0,5** Erro Combinado: 62,5%	2 Espúrio Medida F: 0,5 Erro Combinado: 66,7%
5	Governo <EM>PSD</EM> de Cavaco <EM>Silva<EM>	2 Parc. Correcto Medida F: 0,48 Erro Combinado: 60%	**2 Parc. Correcto** **Medida F: 0,5** **Erro Combinado: 50%**	2 Espúrio Medida F: 0,48 Erro Combinado: 66,7%
6	<EM>Governo PSD<EM> de Cavaco Silva	1 Parc. Correcto Medida F: 0,6 Erro Combinado: 40%	**1 Correcto, 1EmFalta** **Medida F: 0,8** **Erro Combinado: 33,3%**	1 Espúrio Medida F: 0,67 Erro Combinado: 50%
7	Governo PSD de Cavaco <EM>Silva</EM>	1 Parc. Correcto Medida F: 0,55 Erro Combinado: 45%	1 Parc. Cor., 1 Em Falta Medida F: 0,5 Erro Combinado: 58,3%	**1 Espúrio** **Medida F: 0,67** **Erro Combinado: 50%**
8	Governo <EM>PSD de Cavaco</EM> Silva	1 Parc. Correcto Medida F: 0,651 Erro Combinado: 45%	2 Parc. Correcto Medida F: 0,5 Erro Combinado: 58,3%	**1 Espúrio** **Medida F: 0,67** **Erro Combinado: 50%**

Tabela 18.5: Exemplos de selecção de alternativa na tarefa de identificação.

Caso	Precisão		
	ALT1	ALT2	ALT3
1	(1+1)/(1+1)=**100%**	(0,4+1)/(1+1)=70%	(0+1)/(1+1)=50%
2	(0+1)/(0+1)=**100%**	(0+1)/(0+1)=**100%**	(0+1)/(0+1)=**100%**
3	(0,4+1)/(1+1)=**70%**	(0,35+1)/(1+1)=67,5%	(0+1)/(1+1)=50%
4	(0,2+1)/(2+1)=40%	(0,5+1)/(2+1)=**50%**	(0+1)/(2+1)=33,3%
5	(0,2+1)/(2+1)=40%	(0,5+1)/(2+1)=**50%**	(0+1)/(2+1)=33,3%
6	(0,2+1)/(1+1)=60%	(1+1)/(1+1)=**100%**	(0+1)/(1+1)=50%
7	(0,1+1)/(1+1)=55%	(0,25+1)/(1+1)=**62,5%**	(0+1)/(1+1)=50%
8	(0,3+1)/(1+1)=**65%**	(0,25+1)/(1+1)=62,5%	(0+1)/(1+1)=50%

Tabela 18.6: Selecção de alternativa - cálculo de precisão.

Caso	Abrangência		
	ALT1	ALT2	ALT3
1	(1+1)/(1+1)=100%	(0,4+1)/(2+1)=**46,7%**	(0+1)/(0+1)=100%
2	(0+1)/(1+1)=50%	(0+1)/(2+1)=**33,3%**	(0+1)/(0+1)=100%
3	(0,4+1)/(1+1)=70%	(0,35+1)/(2+1)=**45%**	(0+1)/(0+1)=100%
4	(0,2+1)/(1+1)=60%	(0,5+1)/(2+1)=**50%**	(0+1)/(0+1)=100%
5	(0,2+1)/(1+1)=60%	(0,5+1)/(2+1)=**50%**	(0+1)/(0+1)=100%
6	(0,2+1)/(1+1)=**60%**	(1+1)/(2+1)=66,7%	(0+1)/(0+1)=100%
7	(0,1+1)/(1+1)=55%	(0,25+1)/(2+1)=**41,7%**	(0+1)/(0+1)=100%
8	(0,3+1)/(1+1)=65%	(0,25+1)/(2+1)=**41,7%**	(0+1)/(0+1)=100%

Tabela 18.7: Selecção de alternativa - cálculo de abrangência.

Caso	Medida F		
	ALT1	ALT2	ALT3
1	**2×1×1/(1+1)=1**	2×0,7×0,467/(0,7+0,467)=0,56	2×0,5×1/(0,5+1)=0,666
2	2×1×0,5/(1+0,5)=0,66	2×1×0,33/(1+0,33)=0,5	**2×1×1/(1+1)=1**
3	**2×0,7×0,7/(0,7+0,7)=0,7**	2×0,675×0,45/(0,675+0,45)=0,54	2×0,5×1/(0,5+1)=0,666
4	2×0,4×0,6/(0,4+0,6)=0,48	**2×0,33×1/(1+0,33)=0,5**	2×0,5×0,5/(0,5+0,5)=0,5
5	2×0,4×0,6/(0,4+0,6)=0,48	**2×0,5×0,5/(0,5+0,5)=0,5**	2×0,33×1/(1+0,33)=0,5
6	2×0,6×0,6/(0,6+0,6)=0,6	**2×1×0,666/(1+0,666)=0,8**	2×0,5×1/(1+0,5)=0,667
7	2×0,55×0,55/(0,55+0,55)=0,55	2×0,625×0,417/(0,625+0,417)=0,5	**2×0,5×1/(1+0,5)=0,667**
8	2×0,65×0,65/(0,65+0,65)=0,65	2×0,625×0,417/(0,625+0,417)=0,5	**2×0,5×1/(1+0,5)=0,667**

Tabela 18.8: Selecção de alternativa - cálculo de Medida F.

Caso	Erro Combinado		
	ALT1	ALT2	ALT3
1	**0/(0+1)=0%**	(2×(1-0,2))/(2+1)=53,3%	1/(1+1)=50%
2	1/(1+1)=50%	(2×1)/(2+1)=66,6%	**0/(0+1)=0**%
3	**0,6/(1+1)=30%**	((1-0,1)+(1-0,25))/(2+1)=55,0%	1/(1+1)=50%
4	(2×(1-0,1))/(2+1)=60%	**(2×(1-0,25)+1)/(3+1)=62,5%**	2/(2+1)=66,7%
5	(2×(1-0,1))/(2+1)=60%	**(2×(1-0,25))/(2+1)=50%**	2/(2+1)=66,7%
6	(1-0,2)/(1+1)=40%	1/(2+1)=33,3%	1/(1+1)=50%
7	(1-0,1)/(1+1)=45%	(1+(1-0,25))/(2+1)=58,3%	1/(1+1)=50%
8	(1-0,3)/(1+1)=35%	(2×(1-0,125))/(2+1)=58,3%	1/(1+1)=50%

Tabela 18.9: Selecção de alternativa - cálculo de Erro Combinado.

18.3 Tarefa de classificação semântica

A tarefa de classificação semântica avalia até que ponto os sistemas participantes conseguem classificar a EM numa hierarquia de categorias e de tipos definidos no HAREM, que foi especialmente criada o português e foi revista conjuntamente pelos participantes e pela organização.

18.3.1 Medidas

A classificação semântica é avaliada através de quatro medidas, que fornecem mais informação aos participantes sobre o desempenho dos seus sistemas:

Por categorias: pontua-se apenas a categoria da etiqueta.

Por tipos: pontua-se apenas as EM que tiveram categoria(s) pontuada(s) como `correcto`, e onde se avalia somente o atributo `TIPO` da etiqueta.

Combinada: avalia-se as categorias e os tipos da EM, através de uma pontuação que combina as duas através da equação 18.3.

Plana: avalia-se os pares categoria-tipo como folhas de uma classificação plana, considerando apenas como certos os casos que tenham a categoria e o tipo pontuados como `correcto`.

18.3.2 Pontuações

A pontuação na classificação semântica é feita para a categoria e para o tipo, em separado. São usados três valores possíveis:

Correcto: quando a categoria (ou tipo) da EM da saída é igual à categoria (ou tipo) da EM da CD.

Em Falta: quando a categoria (ou tipo) da EM da CD está ausente da categoria (ou tipo) da EM da saída.

Espúrio: quando a categoria (ou tipo) da EM da saída está ausente da categoria (ou tipo) da EM da CD.

Estas são as pontuações usadas para avaliar os alinhamentos, de uma forma genérica. No entanto, como as EM podem ter mais do que uma categoria e tipo (`<ABC... TIPO="XYZ...">`), estas pontuações não podem ser atribuídas assim de uma forma tão linear.

Como tal, vamos detalhar as regras de pontuação para cada medida em separado, ilustrada com exemplos.

Medida por categorias

A pontuação para a classificação semântica medida por categorias avalia as EM da seguinte maneira (ver exemplos na Tabela 18.10):

Caso	Saída Sistema	CD	Correcta	Em Falta	Espúria
1	<A>	<A>	A	-	-
2	<B>	<A>	-	A	B
3	<A>	<ABC>	A	-	-
4	<D>	<ABC>	-	A, B e C	D
5	<A>		-	-	A

Tabela 18.10: Pontuação na classificação semântica medida por categorias.

Correcta: Quando o sistema atribui à EM uma categoria, e se essa categoria for igual à da EM na CD, é pontuada como `correcto` (caso 1 da Tabela 18.10). Contudo, se a respectiva EM da CD possui um conjunto de categorias, basta a categoria da EM da saída corresponder a uma desse conjunto, que além de ser pontuado igualmente como `correcto`, o sistema não será prejudicado por faltarem as outras. Ou seja, o caso 3 da Tabela 18.10 resulta na mesma pontuação que o caso 1.

Em Falta: Se a categoria da EM de saída não corresponde à categoria da EM da CD, no caso de esta ter uma classificação única (caso 2 da Tabela 18.10), ou não corresponder a nenhuma das classificações múltiplas (caso 4 da Tabela 18.10), cada uma das categorias da EM da CD é pontuada como `Em Falta`. Contudo, se a categoria que o sistema classificou estiver incluída no conjunto presente na EM da CD, nada é considerado `Em Falta` (caso 3 da Tabela 18.10).

Espúria: no caso da EM do sistema atribuir uma categoria que não existe na EM da CD, essa categoria é pontuada como `espúria` (casos 2, 4 e 5 da Tabela 18.10). Esta marcação é atribuída quer em conjunção com `Em Falta`, quer se o sistema identificou algo como EM que não o seja.

Medida por tipos

Na classificação semântica medida por tipos, as EM são pontuadas de um modo semelhante à da classificação semântica por categorias, mas entrando em conta apenas com os casos em que as categorias foram correctamente identificadas, ou seja, é uma medida relativa por excelência. A Tabela 18.11 resume a pontuação atribuída nos diversos casos. O raciocínio é análogo ao caso anterior referente às categorias.

Caso	Saída Sistema	CD	Correcta	Em Falta	Espúria
1	<A>	<A TIPO="X">	-	X	-
2	<A TIPO="OUTRO">	<A TIPO="X">	-	X	-
3	<A TIPO="OUTRO">	<AAA TIPO="XYZ">	-	X,Y e Z	-
4	<A TIPO="X">	<A TIPO="X">	X	-	-
5	<A TIPO="X">	<A TIPO="Y">	-	Y	X
6	<A TIPO="X">	<ABC TIPO="XYZ">	X	-	-
7	<A TIPO="X">	<AAA TIPO="XYZ">	X	-	-
8	<A TIPO="X">	<AAA TIPO="WYZ">	-	W, Y e Z	X

Tabela 18.11: Pontuação na classificação semântica medida por tipos.

Medida combinada

A medida semântica combinada combina a pontuação da categoria e do tipo através de uma fórmula única, de modo a indicar o nível da classificação semântica como um todo:

$$\mathrm{P}_{CSC} = \begin{cases} 0 & \text{se a categoria não estiver correcta.} \\ 1 & \text{se a categoria estiver correcta mas o tipo não estiver correcto.} \\ 1 + \left(1 - \dfrac{n_c}{n_t}\right) - \dfrac{n_e}{n_t} & \text{se a categoria estiver correcta e pelo menos um tipo correcto.} \end{cases} \tag{18.3}$$

Onde n_c representa o número de tipos correctos, n_e o número de tipos espúrios, e n_t o número de tipos possível nessa categoria. Note-se que para calcular estes últimos valores, é preciso naturalmente conhecer quantos tipos diferentes cada categoria pode ter, o que está descrito na Tabela 18.12. Como o número de tipos de certas categorias foram alterados do HAREM para o Mini-HAREM, apresentamos os valores para cada evento:

Categoria	HAREM		Mini-HAREM	
	Número de tipos distintos	Valor máximo	Número de tipos distintos	Valor máximo
ABSTRACCAO	8	1,875	8	1,875
ACONTECIMENTO	3	1,666	3	1,667
COISA	3	1,666	4	1,75
LOCAL	5	1,8	5	1,8
OBRA	4	1,75	3	1,667
ORGANIZACAO	4	1,75	4	1,75
PESSOA	6	1,833	6	1,833
TEMPO	4	1,75	4	1,75
VALOR	3	1,667	3	1,667

Tabela 18.12: Quantidade de tipos distintos que uma categoria semântica pode ter, e valor máximo correspondente para o cálculo da medida combinada, para o HAREM e o Mini-HAREM.

Veja-se a Tabela 18.13 com alguns exemplos, em que assumimos que a categoria *A* tem quatro tipos distintos.

Caso	Saída do Sistema	CD	Medida combinada
1	<A TIPO="B">	<A TIPO="C">	1+(1-$\frac{0}{4}$) = 1
2	<A TIPO="B">	<A TIPO="B">	1+(1-$\frac{1}{4}$) = 1,75
3	<A TIPO="B">	<AZ TIPO="BY">	1+(1-$\frac{1}{4}$)-$\frac{1}{4}$ = 1,5
4	<A TIPO="B">	<AA TIPO="CD">	1+(1-$\frac{0}{4}$) = 1

Tabela 18.13: Exemplo para a classificação semântica na medida combinada, para uma categoria A com quatro tipos (n_t) = 4.

Medida plana

Caso	Saída Sistema	CD	Correcta	Em Falta	Espúria
1	<A TIPO="X">	<A TIPO="X">	(A,X)	-	-
2	<A TIPO="Y">	<A TIPO="X">	-	(A,X)	(A,Y)
3	<A TIPO="Y">	<AAA TIPO="XYZ">	(A,Y)	-	-
4	<A TIPO="W">	<AAA TIPO="XYZ">	(A,X|Y|Z)	(A,W)	
5	<B TIPO="Z">	<A TIPO="X">	-	(A,X)	(B,Z)

Tabela 18.14: Pontuação da classificação semântica, na medida plana.

A classificação semântica na medida plana tem como objecto de estudo o par (CATEGORIA, TIPO). Por exemplo, se as EM em análise fossem <LOCAL TIPO="GEOGRAFICO">Coimbra</LOCAL> e <PESSOA TIPO="INDIVIDUAL">Magalhães</PESSOA>, então os pares a serem avaliados seriam (LOCAL, GEOGRAFICO) e (PESSOA, INDIVIDUAL), respectivamente. Um par é pontuado como correcto quando a categoria e o tipo são o mesmo na entidade correspondente da CD. A Tabela 18.14 ilustra as regras da medida.

18.3.3 Métricas

Precisão

A precisão apresenta-se sobre dois cenários: absoluto (para todas as EM) e relativo (às EM correctamente identificadas).

Para a medida por categorias, a precisão é dada pela fórmula:

Absoluto: $\text{Precisão}_{\text{medida categorias}}$ = ($\sum$ EM correctamente identificadas e com categoria correcta + Y) / ($\sum$ EM classificadas pelo sistema)

Relativo: Precisão$_{\text{medida categorias}}$ = ($\sum$ EM correctamente identificadas e com categoria correcta + Y) / ($\sum$ EM parcial ou correctamente identificadas classificadas pelo sistema)

Em que Y corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria correctas. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

A classificação semântica na medida por tipos é, por definição, sempre relativa:

Relativo: Precisão$_{\text{medida tipos}}$ = ($\sum$ EM correctamente identificadas e com categoria e tipo correctos + Z) / ($\sum$ EM correctamente ou parcialmente identificadas)

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

Para a classificação semântica combinada, a precisão mede o grau de sucesso de acordo com a classificação máxima (calculada assumindo que todas as categorias e tipos propostos pelo sistema estiverem correctos):

Absoluto: Precisão$_{\text{medida CSC}}$ = (Valor de CSC obtida pelo sistema / Valor máximo da CSC para a saída do sistema)

Relativo: Precisão$_{\text{medida CSC}}$ = (Valor da CSC obtida pelo sistema / Valor máximo da CSC para a saída do sistema só considerando EM parcial ou correctamente identificadas)

Para a medida plana, a precisão é calculada da seguinte forma:

Absoluto: Precisão$_{\text{medida plana}}$ = ($\sum$ EM correctamente identificadas e com categoria e tipo correctos + Z) / ($\sum$ EM classificadas pelo sistema)

Relativo: Precisão$_{\text{medida plana}}$ = ($\sum$ EM correctamente identificadas e com categoria e tipo correctos + Z) / ($\sum$ EM parcial ou correctamente identificadas e classificadas pelo sistema)

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

Abrangência

A abrangência define-se de forma diferente para cada uma das quatro medidas, e de forma diferente para os cenários absoluto e relativo.

Para a medida por categorias, a abrangência é calculdade da seguinte forma:

Absoluto: $\text{Abrangência}_{\text{medida categorias}}$ = ($\sum$ EM correctamente identificadas e com categoria correcta + Y) / ($\sum$ EM classificadas na CD)

Relativo: $\text{Abrangência}_{\text{medida categorias}}$ = ($\sum$ EM correctamente identificadas e com categoria correcta + Y) / ($\sum$ EM partial ou correctamente identificadas e classificadas na CD)

Em que Y corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria correcta. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

A classificação semântica na medida por tipos é, por definição, sempre relativa:

Relativo: $\text{Abrangência}_{\text{medida tipos}}$ = ($\sum$ EM correctamente identificadas e com categoria e tipo correctos + Z) / ($\sum$ EM correctamente classificadas em categoria na CD)

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

Na avaliação da classificação semântica combinada, a abrangência mede o nível de cobertura de acordo com a classificação máxima (se tanto as categorias como os tipos enviados estiverem correctos). Mais uma vez, no cenário absoluto usam-se todas as EM na CD, e no relativo apenas o subconjunto parcial ou correctamente identificado.

Absoluto: $\text{Abrangência}_{\text{medida CSC}}$ = (Valor da medida semântica combinada obtida pelo sistema / Valor máximo da medida semântica combinada na CD)

Relativo: $\text{Abrangência}_{\text{medida CSC}}$ = (Valor da medida semântica combinada obtida pelo sistema / Valor máximo da medida semântica combinada na CD usando apenas as EM correctamente identificadas)

Para a medida plana, a abrangência calculada da seguinte forma:

Absoluto: $\text{Abrangência}_{\text{medida plana}}$ = (∑ EM correctamente identificadas e com categoria e tipo correctos + Z) / (∑ EM na CD)

Relativo: $\text{Abrangência}_{\text{medida plana}}$ = (∑ EM correctamente identificadas e com categoria e tipo correctos + Z) / (∑ EM parcial ou correctamente identificadas na CD)

Em que Z corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com categoria e tipo correctos. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

Sobre-geração

A sobre-geração na classificação semântica mede o número de EM com uma classificação semântica espúria, em comparação com a CD. A sobre-geração é calculada de forma diferente, de acordo com o cenário usado (absoluto ou relativo).

Para a medida por categorias, a sobre-geração é calculada da seguinte forma:

Absoluto: $\text{Sobre-geração}_{\text{medida categorias}}$ = (∑ EM com classificação semântica espúria na categoria / ∑ EM classificadas com categoria pelo sistema)

Relativo: $\text{Sobre-geração}_{\text{medida categorias}}$ = (∑ EM parcial ou correctamente identificadas com classificação semântica espúria na categoria) / (∑ EM parcial ou correctamente identificadas classificadas com categoria pelo sistema)

A classificação semântica na medida por tipos é, por definição, sempre relativa:

Relativo: $\text{Sobre-geração}_{\text{medida tipos}}$ = (∑ EM com classificação semântica espúria no tipo) / (∑ EM parcial ou correctamente identificadas classificadas com categoria e tipo pelo sistema)

A classificação semântica na medida plana é calculada da seguinte forma:

Absoluto: $\text{Sobre-geração}_{\text{medida plana}}$ = (∑ EM com classificação semântica espúria na categoria ou no tipo) / (∑ EM classificadas com categoria e tipo pelo sistema)

Relativo: $\text{Sobre-geração}_{\text{medida plana}}$ = (∑ EM correctamente identificadas com classificação semântica espúria na categoria ou no tipo + W) / (∑ EM parcial

ou correctamente identificadas classificadas com categoria e tipo pelo sistema)

Em que W corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com classificação semântica espúria na categoria ou no tipo. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

Sub-geração

A sub-geração na classificação semântica mede o número de EM com uma classificação semântica em falta, em comparação com a saída. A sub-geração é calculada de forma diferente, de acordo com o cenário usado (absoluto ou relativo).

Para a medida por categorias, a sub-geração é calculada da seguinte forma:

Absoluto: $\text{Sub-geração}_{\text{medida categorias}}$ = (∑ EM com classificação semântica em falta na categoria) / (∑ EM com categoria na CD)

Relativo: $\text{Sub-geração}_{\text{medida categorias}}$ = (∑ EM correctamente identificadas com classificação semântica em falta na categoria + R) / (∑ EM parcial ou correctamente identificadas com categoria na CD)

Em que R corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com classificação semântica em falta na categoria. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

A classificação semântica na medida por tipos é, por definição, sempre relativa:

Relativo: $\text{Sub-geração}_{\text{medida tipos}}$ = (∑ EM correctamente identificadas com classificação semântica em falta no tipo + S) / (∑ EM parcial ou correctamente identificadas com tipo na CD)

Em que S corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com classificação semântica em falta no tipo. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

A classificação semântica na medida plana é calculada da seguinte forma:

Absoluto: $\text{Sub-geração}_{\text{medida plana}}$ = (∑ EM com classificação semântica em falta na categoria ou no tipo) / (∑ EM com categoria na CD)

Relativo: Sub-geração$_{\text{medida plana}}$ = ($\sum$ EM correctamente identificadas com classificação semântica em falta na categoria ou no tipo + T) / ($\sum$ EM parcial ou correctamente identificadas com categoria e tipo na CD

Em que T corresponde ao somatório dos valores obtidos para as EM parcialmente identificadas e com classificação semântica em falta na categoria ou no tipo. Esses valores são calculados pela fórmula $\frac{n_c}{n_d}$.

18.3.4 Exemplo detalhado de atribuição de pontuação

Apresentamos um exemplo de texto, etiquetado por um sistema hipotético, e a respectiva CD. Para não sobrecarregar o presente documento, todas as EM da CD são identificadas correctamente ou parcialmente (portanto, os cenários relativo e absoluto produzem os mesmos resultados).

Exemplo da colecção dourada:

> Plano hidrológico de **<ORGANIZACAO | LOCAL TIPO="ADMINISTRACAO | ADMINISTRATIVO">** Espanha **</ORGANIZACAO** *mid* **LOCAL>** analisado em **<LOCAL TIPO="ADMINISTRATIVO">** Lisboa **</LOCAL>**. Terminou ontem no **<LOCAL TIPO="ALARGADO">** Laboratório Nacional de Engenharia Civil **</LOCAL>**, em **<LOCAL TIPO="ADMINISTRATIVO">** Lisboa **</LOCAL>**, o **<ACONTECIMENTO TIPO="EVENTO">** Encontro de Reflexão **</ACONTECIMENTO>** sobre a concretização do **<ABSTRACCAO TIPO="PLANO">** Plano Hidrológico **</ABSTRACCAO>** espanhol. Em análise esteve um documento que prevê a transferência de significativos volumes de água dos rios **<LOCAL TIPO="GEOGRAFICO">** Douro **</LOCAL>** e **<LOCAL TIPO="GEOGRAFICO">** Tejo **</LOCAL>** para a bacia hidrográfica do rio **<LOCAL TIPO="GEOGRAFICO">** Jucar **</LOCAL>**.

Exemplo de saída do sistema:

> **<LOCAL TIPO="GEOGRAFICO">** Plano hidrológico de Espanha **</LOCAL>** analisado em **<LOCAL TIPO="ADMINISTRATIVO">** Lisboa **</LOCAL>**. Terminou ontem no **<LOCAL TIPO="ALARGADO">** Laboratório Nacional **</LOCAL>** de **<ORGANIZACAO TIPO="SUB">** Engenharia Civil **</ORGANIZACAO>**, em **<LOCAL TIPO="ADMINISTRATIVO">** Lisboa **</LOCAL>**, o **<ABSTRACCAO TIPO="PLANO">** Encontro de Reflexão **</ABSTRACCAO>** sobre a

concretização do **<ABSTRACCAO TIPO="PLANO">** Plano Hidrológico **</ABSTRACCAO>** espanhol. **<ABSTRACCAO TIPO="DISCIPLINA">** Em análise **</ABSTRACCAO>** esteve um documento que prevê a transferência de significativos volumes de água dos rios **<LOCAL TIPO="GEOGRAFICO">** Douro **</LOCAL>** e **<LOCAL TIPO="GEOGRAFICO">** Tejo **</LOCAL>** para a bacia hidrográfica do rio **<ABSTRACCAO TIPO="PLANO">** Jucar **</ABSTRACCAO>**.

Nos alinhamentos parciais do exemplo, está associado um factor de correcção (calculado na tarefa de identificação) que influencia os cálculos das várias medidas. Especificamente, referimo-nos aos seguintes casos:

- **<LOCAL TIPO="GEOGRAFICO">** Plano hidrológico de Espanha **</LOCAL>**
- **<LOCAL TIPO="ALARGADO">** Laboratório Nacional **</LOCAL>**
- **<ORGANIZACAO TIPO="SUB">** Engenharia Civil **</ORGANIZACAO>**

Em qualquer destes casos a correspondência com a CD é parcial, logo a sua avaliação tem de ser condicionada por um factor de correcção que condiciona a contribuição desta entidade para a avaliação semântica global. O factor de correcção é dado pela fórmula $\frac{n_c}{n_d}$, onde n_c representa o número de átomos comuns entre as duas EM, e n_d representa o número de átomos distintos entre as duas EM.

Isto significa que a contribuição da EM **<LOCAL TIPO="GEOGRAFICO">**Plano hidrológico de Espanha**</LOCAL>** é de 0,25 ($n_c = 1$, e $n_d = 4$) e não 1 para o somatório total.

Medida por categorias

Na Tabela 18.15 apresentamos a pontuação para a classificação semântica segundo a medida por categorias, e na Tabela 18.16 os valores das métricas. Note-se que, no caso das identificações parciais, colocamos entre parênteses o correspondente factor de correcção $\frac{n_c}{n_d}$.

Medida por tipos

Na Tabela 18.17 apresentamos a pontuação para a classificação semântica segundo a medida por tipos, e na Tabela 18.18 os valores das métricas. De notar que os casos 4, 6, 8 e 11 da Tabela 18.17 não são classificados, porque não foram pontuados como correctos na Tabela 18.15.

Caso	Saída do Sistema	Correcta	Em Falta	Espúria
1	<LOCAL TIPO="GEOGRAFICO">Plano hidrológico de Espanha</LOCAL>	LOCAL (0.25)	-	-
2	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	LOCAL	-	-
3	<LOCAL TIPO="ALARGADO"> Laboratório Nacional</LOCAL>	LOCAL (0.4)	-	-
4	<ORGANIZACAO TIPO="SUB"> Engenharia Civil</ORGANIZACAO>	-	-	ORGANIZACAO
5	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	LOCAL	-	-
6	<ABSTRACCAO TIPO="PLANO"> Encontro de Reflexão</ABSTRACCAO>	-	ACONTECIMENTO	ABSTRACCAO
7	<ABSTRACCAO TIPO="PLANO">Plano Hidrológico</ABSTRACCAO>	ABSTRACCAO	-	-
8	<ABSTRACCAO TIPO="DISCIPLINA"> Em análise</ABSTRACCAO>	-	-	ABSTRACCAO
9	<LOCAL TIPO="GEOGRAFICO"> Douro</LOCAL>	LOCAL	-	-
10	<LOCAL TIPO="GEOGRAFICO"> Tejo</LOCAL>	LOCAL	-	-
11	<ABSTRACCAO TIPO="PLANO"> Jucar</ABSTRACCAO>	-	LOCAL	ABSTRACCAO
Total		5,65	2	4

Nota: No caso 4, como a EM anterior do sistema alinhou com a mesma EM da CD, e foi pontuada como correcta no alinhamento anterior, não podemos pontuar a categoria LOCAL como em falta.

Tabela 18.15: Pontuação da classificação semântica medida por categorias, para o exemplo dado.

Métrica	Valor
Precisão	$\frac{5,65}{11} = 51,36\%$
Abrangência	$\frac{5,65}{9} = 62,77\%$
Medida F	$\frac{2 \times 0,5136 \times 0,6277}{0,5136+0,6277} = 0,565$
Sobre-geração	$\frac{4}{11} = 36,36\%$
Sub-geração	$\frac{2}{9} = 22,2\%$

Tabela 18.16: Valores das métricas para a tarefa de classificação semântica, medida por categorias, para o exemplo dado.

Medida combinada

Na Tabela 18.19 apresentamos a pontuação para a classificação semântica segundo a medida combinada, e na Tabela 18.20 os valores das métricas. Salientamos que os casos 1 e 3

Caso	Saída do Sistema	Correcta	Em Falta	Espúria
1	<LOCAL TIPO="GEOGRAFICO"> Plano hidrológico de Espanha</LOCAL>	-	ADMINISTRATIVO	GEOGRAFICO
2	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	ADMINISTRATIVO	-	-
3	<LOCAL TIPO="ALARGADO"> Laboratório Nacional</LOCAL>	ALARGADO (0.4)	-	-
5	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	ADMINISTRATIVO	-	-
7	<ABSTRACCAO TIPO="PLANO"> Plano Hidrológico</ABSTRACCAO>	PLANO	-	-
9	<LOCAL TIPO="GEOGRAFICO"> Douro</LOCAL>	GEOGRAFICO	-	-
10	<LOCAL TIPO="GEOGRAFICO"> Tejo</LOCAL>	GEOGRAFICO	-	-
Total		5,4	1	1

Tabela 18.17: Pontuação da classificação semântica por tipos, para o exemplo dado.

Métrica	Valor
Precisão	$\frac{5,4}{7} = 77,14\%$
Abrangência	$\frac{5,4}{7} = 77,14\%$
Medida F	$\frac{2 \times 0,7714 \times 0,7714}{0,7714+0,7714} = 0,7714$
Sobre-geração	$\frac{1}{7} = 14,28\%$
Sub-geração	$\frac{1}{7} = 14,28\%$

Tabela 18.18: Valores das métricas para a classificação semântica, medida por tipos, para o exemplo dado.

da Tabela 18.19 são multiplicados pelo factor de correcção $\frac{n_c}{n_d}$, respectivamente, 0,25 e 0,4.

Medida plana

Na Tabela 18.21 apresentamos a pontuação para a classificação semântica segundo a medida plana, e na Tabela 18.22 os valores das métricas. Salientamos que os casos 1 e 3 da Tabela 18.21 são multiplicados pelo factor de correcção, 0,25 e 0,4, respectivamente.

Caso	Classificação
1	$1+0 \times \left(1-\frac{1}{5}\right) \times 0,25 = 0,25$
2	$1+1 \times \left(1-\frac{1}{5}\right) = 1,80$
3	$1+1 \times \left(1-\frac{1}{5}\right) \times 0,4 = 0,72$
4	0,0
5	$1+1 \times \left(1-\frac{1}{5}\right) = 1,80$
6	0,0
7	$1+1 \times \left(1-\frac{1}{8}\right) = 1,875$
8	0,0
9	$1+1 \times \left(1-\frac{1}{5}\right) = 1,80$
10	$1+1 \times \left(1-\frac{1}{5}\right) = 1,80$
11	0,0
Total	10,045

Tabela 18.19: Pontuação da classificação semântica segundo a medida combinada, para o exemplo dado.

Métrica	Valor
Precisão máxima do sistema	$\frac{10,045}{20,05} = 50,1\%$
Abrangência Máxima na CD	$\frac{10,045}{16,14} = 62,2\%$
Medida F	$\frac{2 \times 0,501 \times 0,6223}{0,501+0,6223} = 0,555$

Nota: o denominador do cálculo da precisão máxima do sistema corresponde ao somatório do cálculo da classificação semântica combinada assumindo que as classificações atribuídas pelo sistema estão totalmente correctas. Para melhor perceber este conceito imagine que as categorias da Tabela 18.15 e os tipos (agora com os restantes casos 4, 6, 8 e 11) da Tabela 18.17 estivessem a ser sempre considerados correctos.
Analogamente, o denominador do cálculo da abrangência máxima da CD utiliza a mesma fórmula para calcular o somatório das classificações combinadas para cada uma das entidades na CD.

Tabela 18.20: Valores das métricas para a tarefa de classificação semântica, segundo a medida combinada, para o exemplo dado.

Caso	Saída do Sistema	Correcta	Em Falta	Espúria
1	<LOCAL TIPO="GEOGRAFICO">Plano hidrológico de Espanha</LOCAL>	-	(LOCAL, ADMINISTRATIVO)	(LOCAL, GEOGRAFICO)
2	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	(LOCAL, ADMINISTRATIVO)	-	-
3	<LOCAL TIPO="ALARGADO"> Laboratório Nacional</LOCAL>	(LOCAL, ALARGADO)(0.4)	-	-
4	<ORGANIZACAO TIPO="SUB"> Engenharia Civil</ORGANIZACAO>	-	*	(ORGANIZACAO, SUB)
5	<LOCAL TIPO="ADMINISTRATIVO"> Lisboa</LOCAL>	(LOCAL, ADMINISTRATIVO)	-	-
6	<ABSTRACCAO TIPO="PLANO"> Encontro de Reflexão</ABSTRACCAO>	-	(ACONTECIMENTO, EVENTO)	(ABSTRACCAO, PLANO)
7	<ABSTRACCAO TIPO="PLANO"> Plano Hidrológico</ABSTRACCAO>	(ABSTRACCAO, PLANO)	-	-
8	<ABSTRACCAO TIPO="DISCIPLINA"> Em análise</ABSTRACCAO>	-	-	(ABSTRACCAO, DISCIPLINA)
9	<LOCAL TIPO="GEOGRAFICO"> Douro</LOCAL>	(LOCAL, GEOGRAFICO)	-	-
10	<LOCAL TIPO="GEOGRAFICO"> Tejo</LOCAL>	(LOCAL, GEOGRAFICO)	-	-
11	<ABSTRACCAO TIPO="PLANO"> Jucar</ABSTRACCAO>	-	(LOCAL GEOGRAFICO)	(ABSTRACCAO, PLANO)
Total	-	5,4	3	5

Tabela 18.21: Valores das métricas para a classificação semântica, segundo a medida plana, para o exemplo dado.

Métrica	Valor
Precisão	$\frac{5,4}{11} = 49,09\%$
Abrangência	$\frac{5,4}{9} = 60,00\%$
Medida F	$\frac{2\times 0,4909\times 0,6000}{0,4909+0,6000} = 0,5400$
Sobre-geração	$\frac{5}{11} = 45,45\%$
Sob-geração	$\frac{3}{9} = 33,33\%$

Tabela 18.22: Avaliação global da tarefa de classificação semântica segundo a medida plana.

18.4 Tarefa de classificação morfológica

A tarefa de classificação morfológica tem por objectivo avaliar a aptidão do sistema em definir qual o género e o número das EM identificadas, em comparação com as respectivas classificações morfológicas feitas manualmente na CD.

18.4.1 Medidas

A tarefa de classificação morfológica é avaliada segundo três medidas:

número: só é considerada a pontuação relativamente ao número.

género: só é considerada a pontuação relativamente ao género.

combinada: combina-se as pontuações para género e para o número.

Note-se, além disso, que a avaliação morfológica é apenas feita sobre as EM que também foram classificadas morfologicamente na CD. As classificações morfológicas feitas a EM que não estão classificadas na CD (como por exemplo as EM de categoria `TEMPO`) são simplesmente ignoradas no processamento subsequente.

18.4.2 Pontuações

As pontuações na tarefa de classificação morfológica podem variar de acordo com o cenário de avaliação usado. Em certos casos como é ilustrado no caso 10 da Tabela 18.23, podemos constatar que a pontuação no cenário absoluto é espúria, enquanto que no cenário relativo, a a EM é ignorada para efeitos de pontuação. Tal facto deve-se ao facto de, no cenário relativo, as EM que são espúrias na tarefa de identificação também são ignoradas na tarefa de classificação morfológica.

Quando uma EM é imperfeitamente reconhecida (ou seja, foi classificada na tarefa de identificação como parcialmente correcta), apenas contámos os casos em que essa identificação parcial concordava na primeira palavra da EM, multiplicando por um peso de 0,5 as EM que estão morfologicamente correctas.

A pontuação para cada uma das medidas segue as regras ilustradas na Tabela 18.23.

Nas tabelas seguintes, vamos mais uma vez considerar, para simplicidade de exposição, que os exemplos são relativos a EM que o participante queria classificar (cenário selectivo), ou então a todas as etiquetas da CD (cenário total), e que todas as identificações estavam correctas.

Se estivermos num cenário relativo (ou seja, só considerando as EM com valor de pontuação maior que 0 na tarefa de identificação) e os 10 exemplos da Tabela 18.23 como um exemplo de saída do sistema participante (note-se que os casos 9 e 10 serão ignorados e não contabilizados), a avaliação global produziria os resultados apresentados na Tabela 18.24.

Caso	Classificação		Medida		
	CD	Sistema	Género	Número	Combinada
1	M,S	M,S	Correcto	Correcto	Correcto
2	M,S	F,S	Incorrecto	Correcto	Incorrecto
3	M,S	M,P	Correcto	Incorrecto	Incorrecto
4	M,S	F,P	Incorrecto	Incorrecto	Incorrecto
5	M,S	?,S	Em Falta	Correcto	Em Falta
6	?,S	M,S	Sobre-especificado	Correcto	Incorrecto
7	?,S	?,S	Correcto	Correcto	Correcto
8	M,S	Não submetido	Em Falta	Em Falta	Em Falta
9	sem identificação	Não submetido	Ignorado	Ignorado	Ignorado
10	sem identificação	Submetido, sem ter ?	Ignorado (Cen. Relativo) Espúrio (Cen. Absoluto)	Ignorado (Cen. Relativo) Espúrio (Cen. Absoluto)	Ignorado (Cen. Relativo) Espúrio (Cen. Absoluto)

Tabela 18.23: Pontuação para a classificação morfológica, segundo as três medidas.

Cenário Absoluto

Métrica	Género	Número	Combinada
Precisão	$\frac{3}{8}$ = 37,5%	$\frac{5}{8}$ = 62,5%	$\frac{2}{8}$ = 25,0%
Abrangência	$\frac{3}{8}$ = 37,5%	$\frac{5}{8}$ = 62,5%	$\frac{2}{8}$ = 25,0%
Medida F	$\frac{2\times0,375\times0,375}{(0,375+0,375)}$=0,375	$\frac{2\times0,625\times0,625}{(0,625+0,625)}$=0,625	$\frac{2\times0,25\times0,25}{(0,25+0,25)}$=0,25
Sobre-especificação	$\frac{1}{8}$ = 12,5%	$\frac{0}{8}$ = 0%	-
Sub-geração	$\frac{2}{8}$ = 25,0%	$\frac{1}{8}$ = 12,5%	-

Cenário Relativo

Métrica	Género	Número	Combinada
Precisão	$\frac{3}{7}$ = 42,8%	$\frac{5}{7}$ = 71,4%	$\frac{2}{7}$ = 28,3%
Abrangência	$\frac{3}{8}$ = 37,5%	$\frac{5}{8}$ = 62,5%	$\frac{2}{8}$ = 25,0%
Medida F	$\frac{2\times0,428\times0,375}{(0,428+0,375)}$=0,40	$\frac{2\times0,714\times0,625}{(0,714+0,625)}$=0,666	$\frac{2\times0,283\times0,25}{(0,283+0,25)}$=0,266
Sobre-especificação	$\frac{1}{7}$ = 14,3%	$\frac{0}{7}$ = 0%	-
Sub-geração	$\frac{2}{8}$ = 25,0%	$\frac{1}{8}$ = 12,5%	-

Tabela 18.24: Valor das métricas para as três medidas da classificação morfológica, considerando os 10 casos da Tabela 18.23.

18.4.3 Métricas

Precisão

Na tarefa de classificação morfológica, a precisão mede o teor de classificações em género/número correctas de todas as produzidas pelo sistema (que tenham classificação morfológica na CD). Ou seja, excluindo sempre os casos em que a EM da CD não se encontra marcada morfologicamente.

Apresentamos a precisão para as três medidas (género, número e combinada), e para os dois cenários de avaliação: independente da identificação (absoluto), ou apenas para os casos em que a identificação obteve pontuação correcta ou parcialmente correcta (relativo).

Absoluto: $\text{Precisão}_{\text{género}}$ = (∑ EM identificadas correctamente e com género correcto + 0,5∑ EM identificadas parcialmente correctamente e com género correcto) / (∑ EM com classificações de género produzidas pelo sistema)

Relativo: $\text{Precisão}_{\text{género}}$ = (∑ EM identificadas correctamente e com género correcto + 0,5∑ EM identificadas parcialmente correctamente e com género correcto) / (∑ EM com classificações de género produzidas pelo sistema em EM identificadas correctamente ou parcialmente)

Absoluto: $\text{Precisão}_{\text{número}}$ = (∑ EM identificadas correctamente e com número correcto + 0,5∑ EM identificadas parcialmente correctamente e com número correcto) / (∑ EM com classificações de número produzidas pelo sistema)

Relativo: $\text{Precisão}_{\text{número}}$ = (∑ EM identificadas correctamente e com número correcto + 0,5∑ EM identificadas parcialmente correctamente e com número correcto) / (∑ EM com classificações de número produzidas pelo sistema em EM identificadas correctamente ou parcialmente)

Absoluto: $\text{Precisão}_{\text{combinada}}$ = (∑ EM identificadas correctamente e com género e número correcto + 0,5∑ EM identificadas parcialmente correctamente e com género e número correcto) / (∑ EM com classificações de número e género produzidas pelo sistema)

Relativo: $\text{Precisão}_{\text{combinada}}$ = (∑ EM identificadas correctamente e com género e número correcto + 0,5∑ EM identificadas parcialmente correctamente e com género e número correcto) / (∑ EM com classificações de número e género produzidas pelo sistema em EM identificadas correctamente ou parcialmente)

Abrangência

Na tarefa de classificação morfológica, a abrangência mede o teor de classificações em género/número que se encontram na CD e que o sistema conseguiu acertar. Tal como para a precisão, mede-se a abrangência no género morfológico, no número morfológico, e na combinação de ambos. No cenário relativo, restringe-se o denominador às EM da CD que foram parcial ou correctamente identificadas pelo sistema.

Absoluto: $\text{Abrangência}_{\text{género}}$ = (∑ EM correctamente identificadas com classificações de género correctas + 0,5∑ EM identificadas parcialmente correctamente com classificações de género correctas) / (∑ EM com classificações de género na CD)

Relativo: $\text{Abrangência}_{\text{género}}$ = (∑ EM correctamente identificadas com classificações de género correctas + 0,5∑ EM identificadas parcialmente correctamente com classificações de género correctas) / (∑ EM parcial ou correctamente identificadas com classificações de género na CD)

Absoluto: $\text{Abrangência}_{\text{número}}$ = (∑ EM correctamente identificadas com classificações de número correctas + 0,5∑ EM identificadas parcialmente correctamente com classificações de número correctas) / (∑ EM com classificações de número na CD)

Relativo: $\text{Abrangência}_{\text{número}}$ = (∑ EM correctamente identificadas com classificações de número correctas + 0,5∑ EM identificadas parcialmente correctamente com classificações de número correctas) / (∑ EM parcial ou correctamente identificadas com classificações de número na CD)

Absoluto: $\text{Abrangência}_{\text{combinada}}$ = (∑ EM correctamente identificadas com classificações de número e género correctas + 0,5∑ EM identificadas parcialmente correctamente com classificações de número e género correctas) / (∑ EM com classificação morfológica na CD)

Relativo: $\text{Abrangência}_{\text{combinada}}$ = (∑ EM correctamente identificadas com classificações de número e género correctas + 0,5∑ EM identificadas parcialmente correctamente com classificações de número e género correctas) / (∑ EM parcial ou correctamente identificadas com classificação morfológica na CD)

Note-se que os denominadores para as três medidas (género, número e combinada), embora formulados de maneira diferente, são exactamente iguais.

Sobre-geração

Relembramos que não se considera, para efeitos de avaliação, espúrios morfológicos (ou seja, só contam para avaliação os casos que também contêm classificação morfológica na CD). Assim, só no cenário absoluto é que há medida de sobre-geração, uma vez que num cenário relativo, não existem EM com morfologia identificadas como espúrias, sendo portanto o valor desta medida sempre 0.

Absoluto: $\text{Sobre-geração}_{\text{género}}$ = ($\sum$ EM com classificações em género espúrias) / ($\sum$ EM com classificações em género produzidas pelo sistema e que tenham também classificação morfológica na CD)

Absoluto: $\text{Sobre-geração}_{\text{número}}$ = ($\sum$ EM com classificações em número espúrias) / ($\sum$ EM com classificações de número produzidas pelo sistema e que tenham também classificação morfológica na CD)

Absoluto: $\text{Sobre-geração}_{\text{combinada}}$ = ($\sum$ EM com classificações em número ou género espúrias) / ($\sum$ EM com classificações de número ou género produzidas pelo sistema e que tenham também classificação morfológica na CD)

Sobre-especificação

Para a tarefa de classificação morfológica, consideramos também a medida de sobre-especificação, que mede a percentagem dos casos sobre-especificados em todos os casos analisados pelo sistema. Por sobre-especificado entendemos os casos em que na CD está "?"e o sistema escolheu um determinado valor concreto.

Absoluto: $\text{Sobre-especificação}_{\text{género}}$ = ($\sum$ EM com classificações de género sobre-especificadas em EM identificadas correctamente + $0,5\sum$ EM com classificações em género sobre-especificadas em EM identificadas parcialmente correctamente) / ($\sum$ EM com classificações de género produzidas pelo sistema)

Relativo: $\text{Sobre-especificação}_{\text{género}}$ = ($\sum$ EM com classificações de género sobre-especificadas em EM identificadas correctamente + $0,5\sum$ EM com classificações em género sobre-especificadas em EM identificadas parcialmente correctamente) / ($\sum$ EM com classificações de género produzidas pelo sistema em EM identificadas parcial ou correctamente)

Absoluto: Sobre-especificação$_{número}$ = (∑ EM com classificações de número sobre-especificadas em EM identificadas correctamente + 0,5∑ EM com classificações em número sobre-especificadas em EM identificadas parcialmente correctamente) / (∑ EM com classificações de número produzidas pelo sistema)

Relativo: Sobre-especificação$_{número}$ = (∑ EM com classificações de número sobre-especificadas em EM identificadas correctamente + 0,5∑ EM com classificações em número sobre-especificadas em EM identificadas parcialmente correctamente) / (∑ EM com classificações de número produzidas pelo sistema em EM identificadas parcial ou correctamente)

Absoluto: Sobre-especificação$_{combinada}$ = (∑ EM com classificações de número ou género sobre-especificadas em EM identificadas correctamente + 0,5∑ EM com classificações em número ou género sobre-especificadas em EM identificadas parcialmente correctamente) / (∑ EM com classificações morfológicas produzidas pelo sistema)

Relativo: Sobre-especificação$_{combinada}$ = (∑ EM com classificações de número ou género sobre-especificadas em EM identificadas correctamente + 0,5∑ EM com classificações em número ou género sobre-especificadas em EM identificadas parcialmente correctamente) / (∑ EM com classificações morfológicas produzidas pelo sistema em EM identificadas parcial ou correctamente)

Sub-geração

Na tarefa de classificação morfológica, a subgeração mede o número de classificações em falta comparadas com a informação morfológica na CD. Classificações em falta incluem tanto casos em que nenhuma classificação foi dada, como casos em que o sistema pôs ? para a classificação do género ou número enquanto na CD existe um valor mais específico. Como anteriormente, apresentamos separadamente as fórmulas para o cenário absoluto e relativo.

Absoluto: Sub-geração$_{género}$ = (∑ EM com classificações em género em falta /∑ classificações em género na CD)

Relativo: Sub-geração$_{género}$ = (∑ EM parcial ou correctamente identificadas com classificações em género em falta) / (∑ EM parcial ou correctamente identificadas com classificações em género na CD)

Absoluto: Sub-geração$_{\text{número}}$ = ($\sum$ EM com classificações em número em falta) / ($\sum$ classificações em número na CD)

Relativo: Sub-geração$_{\text{número}}$ = ($\sum$ EM parcial ou correctamente identificadas com classificações em número em falta) / ($\sum$ EM parcial ou correctamente identificadas com classificações em número na CD)

Absoluto: Sub-geração$_{\text{combinada}}$ = ($\sum$ EM com classificações em género ou número em falta /$\sum$ classificações morfológicas na CD)

Relativo: Sub-geração$_{\text{combinada}}$ = ($\sum$ EM parcial ou correctamente identificadas com classificações em género em falta /$\sum$ EM parcial ou correctamente identificadas com classificações morfológicas na CD)

18.5 Apresentação dos resultados

Os resultados da avaliação são depois apresentados sob duas formas:

Globais: centrados sobre os diversos aspectos da avaliação (por uma determinada categoria, um cenário ou um género textual, por exemplo). Aqui, o desempenho das várias saídas (devidamente anonimizadas) são reunidas em torno de tabelas e/ou gráficos, para permitir uma análise global sobre o comportamento dos sistemas para cada aspecto da avaliação.

Individuais: centrado sobre o desempenho de uma saída. As tabelas e/ou gráficos mostram a posição que a saída ocupou em relação às restantes saídas (devidamente anonimizadas). Estes relatórios possuem dados adicionais sobre o desempenho da saída que não são usados nos relatórios globais.

18.5.1 Resultados globais

Para os resultados globais, apresentam-se várias tabelas comparativas do desempenho dos sistemas. Cada tabela diz respeito a um conjunto dos seguintes parâmetros:

Tarefa: pode ser identificação, classificação morfológica ou classificação semântica.

Por critérios: pode ser global, ou discriminado por categorias, por género textual ou por variante.

Cenário: pode ser total (absoluto ou relativo, nas tarefas de classificação) ou selectivo (absoluto ou relativo, nas tarefas de classificação).

Medida: género. número ou combinada (classificação morfológica), ou por categorias, por tipos, combinada ou plana (na classificação semântica).

De reparar que, nos relatórios globais, os sistemas são devidamente anonimizados, tendo os nomes das saídas sido substituídos por pseudónimos.

As tabelas apresentam os valores para as métricas para cada medida / cenário usado. Um exemplo de tabela, para a tarefa de identificação, no global, em cenário total, é assim representada:

	Precisão (%)	Abrangência (%)	Medida F	Erro combinado	Sobre-geração	Sub-geração
riad	78,50	82,84	0,8061	0,2752	0,07913	0,07329
casablanca	77,15	84,35	0,8059	0,2721	0,09134	0,03575
ancara	76,85	83,56	0,8006	0,2781	0,08966	0,04035
sana	77,43	69,57	0,7329	0,3796	0,09524	0,2079
bahrein	59,45	64,39	0,6182	0,5056	0,2018	0,1607
asmara	56,95%	64,39%	0,6044	0,5230	0,2353	0,1607

Tabela 18.25: Exemplo de uma tabela no relatório global, que compara o desempenho de várias saídas para uma determinada tarefa.

Nos relatórios globais, a tabela é acompanhada também de gráficos. Os valores são apresentados em forma de gráfico de barras (ver Figura 18.1) e em forma de gráfico de pontos (ver Figura 18.2). Nos gráficos de barras, as saídas ficam no eixo das ordenadas, e nos gráficos de pontos, cada ponto representa uma saída.

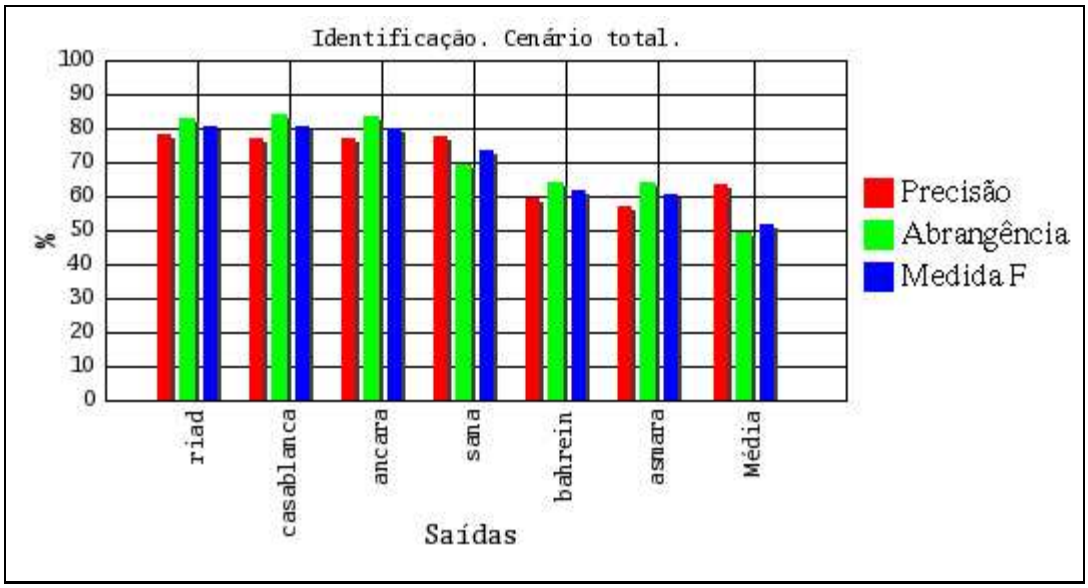


Figura 18.1: Exemplo de um gráfico de barras para o relatório global da tarefa de identificação (cenário total), apresentando os valores da precisão, abrangência e medida F.

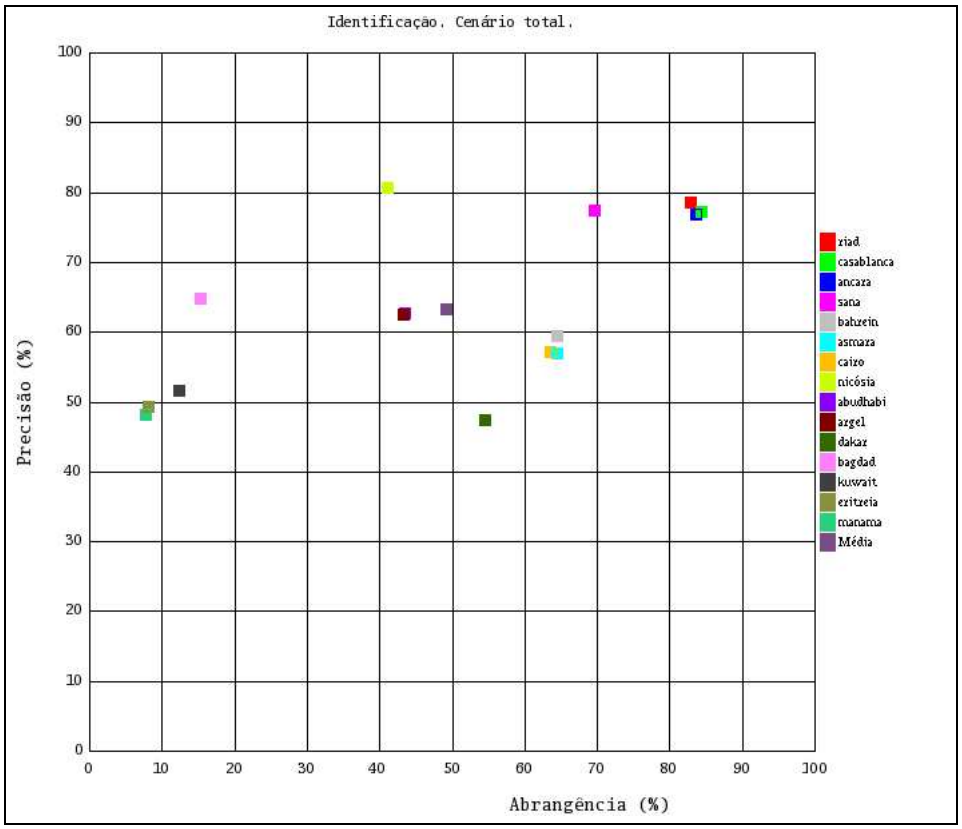


Figura 18.2: Exemplo de um gráfico de pontos para o relatório global da tarefa de identificação (cenário total).

18.5.2 Resultados individuais

Os resultados individuais de cada saída são gerados pelo módulo Alcaide (descrito em 19.2.15, na secção 19.2.15) com base nos relatórios globais, mas com os seguintes melhoramentos:

Resultados filtrados: Nas tabelas de resultados, só se mostra o desempenho das saídas do sistema. A tabela é complementada com informação adicional dos valores de avaliação detalhados. Nos respectivos gráficos de barras, mostra-se também o desempenho de todas as saídas, mas na legenda mostra-se o nome real das saídas do sistema, em vez dos respectivos pseudónimos. Note-se um exemplo de desempenho, para a saída do sistema RENA, na Tabela 18.26 e Figuras 18.3 e 18.4.

Agrupamento de cenários: Enquanto que nos relatórios globais, os resultados são discriminados por cada item (ou seja, há uma tabela para os desempenhos para cada ca-

Total na CD: 5002. Identificadas: 4494. Correctas: 3305 (66,07%).
Parcialmente Correctas: 836 (16,71%). Espúrias: 428 (8,56%). Em Falta: 1040 (20,79%).

Posição	Precisão (%)	Abrangência (%)	Medida F	Erro combinado	Sobre-geração	Sub-geração
4º	77,43	69,57	0,7329	0,3796	0,09524	0,2079

Tabela 18.26: Tabela do relatório individual para a saída RENA, para a tarefa de identificação.

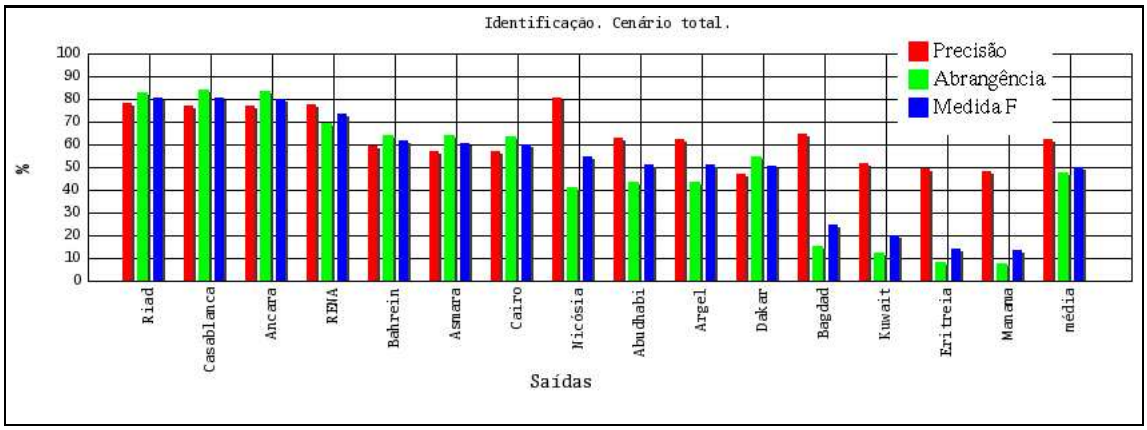


Figura 18.3: Exemplo de um gráfico de barras para o relatório individual da tarefa de identificação (cenário total) para a saída RENA, apresentando os valores da precisão, abrangência e medida F.

tegoria, género textual ou variante), nos relatórios individuais os desempenhos da saída são reunidos numa única tabela. O nome da saída é substituído pela sua posição relativa às outras saídas. Adicionalmente, os valores de avaliação detalhados são agrupados também em tabelas novas (ver Tabelas 18.27 e 18.28).

Gráficos de pontos individual: No caso de cenários (categoria, género textual ou variante), os gráficos de pontos apresentam o desempenho da saída para cada item, em vez de comparar para as restantes saídas como no relatório global (ver Figura 18.5).

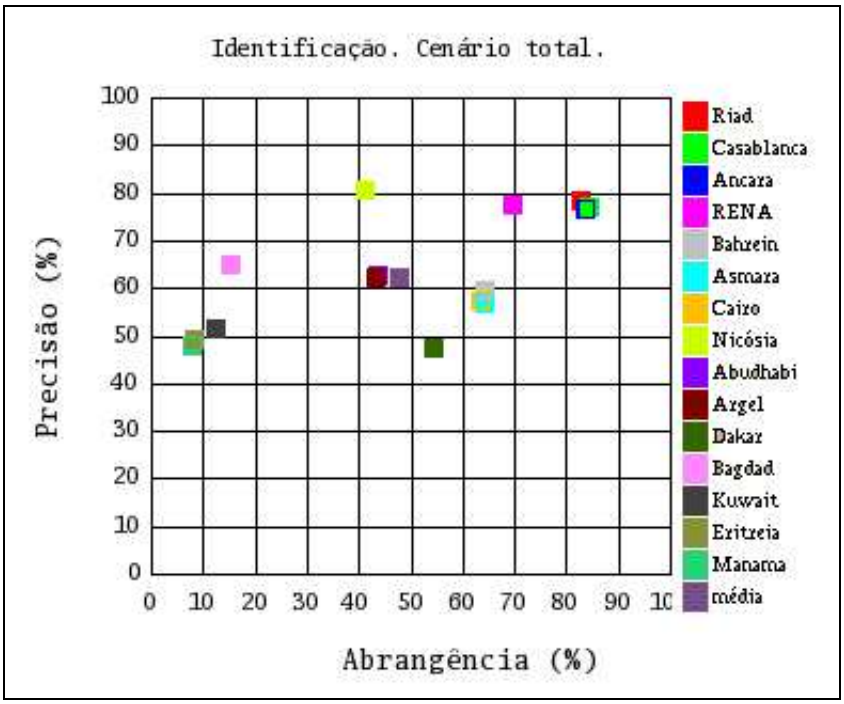


Figura 18.4: Exemplo de um gráfico de pontos para o relatório individual da tarefa de identificação (cenário total) para a saída RENA.

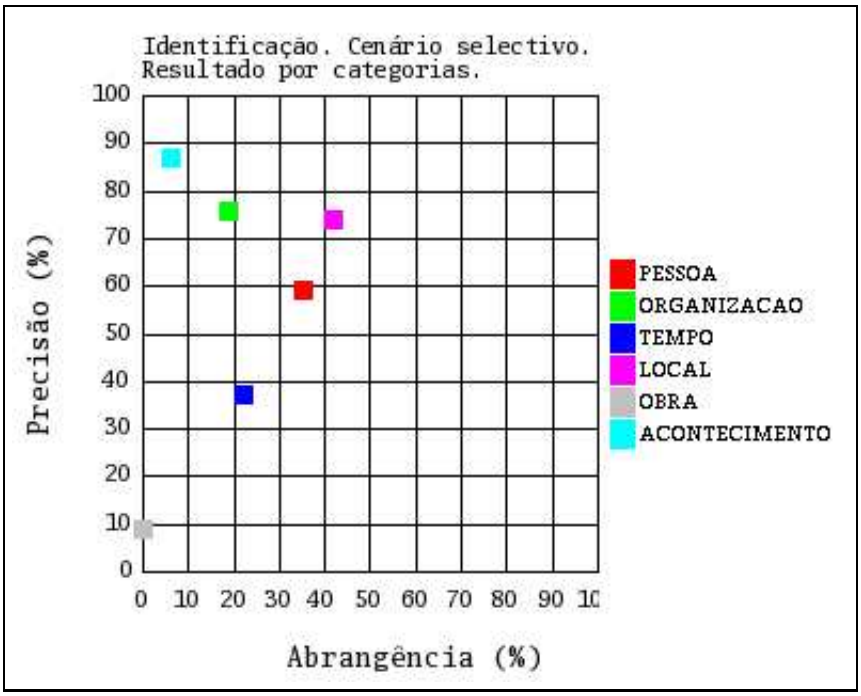


Figura 18.5: Exemplo de um gráfico de pontos para o relatório individual da tarefa de identificação (cenário total) para a saída RENA, discriminada por categorias.

Categoria	Total CD	Identificadas	Correctas		Parc. correctas		Espúrias		Em Falta	
			Total	%	Total	%	Total	%	Total	%
PESSOA	1024	619	339	33,11%	108	10,55%	178	17,38%	580	56,64%
ORGANIZACAO	955	242	176	18,43%	33	3,46%	36	3,77%	746	78,12%
TEMPO	434	264	96	22,12%	11	2,53%	161	37,10%	327	75,35%
LOCAL	1244	713	521	41,88%	47	3,78%	145	11,66%	678	54,50%
OBRA	215	4	0	0,00%	1	0,47%	3	1,40%	214	99,53%
ACONTECIMENTO	109	8	7	6,42%	0	0,00%	1	0,92%	102	93,58%
ABSTRACCAO	453	0	0	0,00%	0	0,00%	0	0,00%	453	100,00%
COISA	81	0	0	0,00%	0	0,00%	0	0,00%	81	100,00%
VALOR	479	0	0	0,00%	0	0,00%	0	0,00%	479	100,00%

Tabela 18.27: Exemplo de uma tabela com valores de avaliação detalhados do relatório individual. No caso presente, os valores referem-se aos desempenhos da saída RENA para a tarefa de identificação, discriminadas por categorias (cenário total).

Categoria	Posição	Precisão (%)	Abrangência (%)	Medida F	Erro combinado	Sobre-geração	Sub-geração
PESSOA	5º	59,23	35,80	0,4463	0,6958	0,2876	0,5664
ORGANIZACAO	7º	76,03	19,27	0,3074	0,8143	0,1488	0,7812
TEMPO	7º	37,44	22,77	0,2832	0,8339	0,6098	0,7535
LOCAL	7º	74,55	42,73	0,5432	0,6179	0,2034	0,5450
OBRA	5º	9,375	0,1744	0,003425	0,9983	0,7500	0,9953
ACONTECIMENTO	5º	87,50	6,422	0,1197	0,9364	0,1250	0,9358

Tabela 18.28: Exemplo de uma tabela de desempenho discriminado do relatório individual. No caso presente, os valores referem-se aos desempenhos da saída RENA para a tarefa de identificação, discriminadas por categorias (cenário total).

Capítulo 19

A arquitectura dos programas de avaliação do HAREM

Nuno Seco, Nuno Cardoso, Rui Vilela e Diana Santos

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 19, p. 283–306, 2007.

A plataforma de avaliação do HAREM consiste num conjunto de módulos utilizado nas avaliações conjuntas realizadas pela Linguateca para medir o desempenho dos sistemas de reconhecimento de entidades mencionadas (REM) participantes no HAREM. Estes programas foram concebidos de acordo com as directivas de avaliação do HAREM, aprovadas pela organização e pelos participantes, e que republicámos no capítulo 18.

A plataforma foi implementada segundo uma arquitectura modular, onde cada módulo executa uma tarefa simples e específica. O resultado final da avaliação é obtido através da sua execução numa determinada sequência.

A opção por uma arquitectura modular, desenvolvida por quatro autores em locais diferentes, foi motivada pelas seguintes considerações:

- a modularização facilita a depuração dos módulos, assim como a verificação de que o seu funcionamento cumpre as directivas de avaliação do HAREM;
- permite o desenvolvimento descentralizado e cooperativo dos programas, com os vários módulos a serem desenvolvidos por diferentes programadores;
- permite o desenvolvimento dos módulos na linguagem de programação em que o programador se sente mais confortável, visto que os módulos podem ser implementados em linguagens diferentes.

Este documento descreve detalhadamente cada um dos programas que compõem a plataforma de avaliação, já apresentada e motivada em Seco et al. (2006). Começamos por apresentar a arquitectura em termos globais, fornecendo depois a descrição pormenorizada de cada módulo.

19.1 Sinopse da arquitectura

A figura 19.1 apresenta o esquema da arquitectura da plataforma de avaliação do HAREM, indicando os módulos que a compõem, e a forma como interagem. A avaliação do HAREM pode ser dividida em quatro fases:

Fase 1: Extracção e alinhamento

A sintaxe das saídas dos sistemas é verificada e corrigida através de um **Validador**. O subconjunto de documentos da saída que também estão presentes na colecção dourada (CD, ver Santos e Cardoso (2006)), é extraído pelo **Extractor de CD**. As EM desse subconjunto são posteriormente alinhadas com as respectivas EM da CD pelo **AlinhEM**, gerando uma lista de *alinhamentos*. O **AvalIDa** processa os alinhamentos e produz os primeiros resultados para a tarefa de identificação.

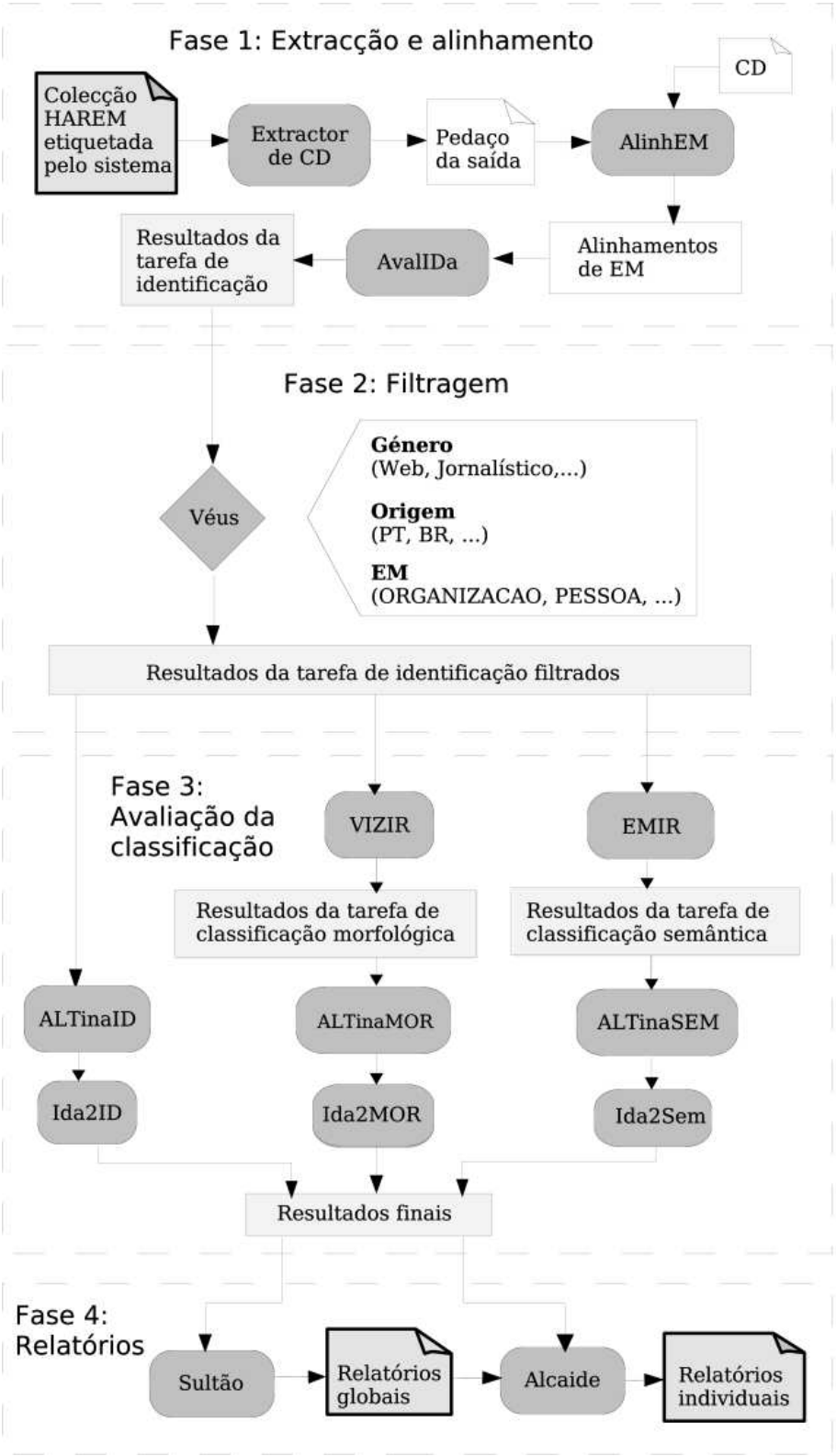


Figura 19.1: Esquema da plataforma de avaliação do HAREM.

Fase 2: Filtragem

A filtragem selectiva dos alinhamentos permite a avaliação parcial segundo diversos cenários específicos. O módulo **Véus** é responsável pela filtragem dos alinhamentos, a partir de uma lista de restrições, que pode incluir um conjunto de categorias e de tipos, um género textual, uma variante ou o resultado da avaliação na tarefa de identificação.

Fase 3: Avaliação da tarefa de classificação

A avaliação das tarefas de classificação morfológica e semântica é realizada em paralelo pelo **Vizir** e pelo **Emir**, respectivamente, a partir dos alinhamentos. O **ALTinaID**, o **ALTinaMOR** e o **ALTinaSEM** analisam as EM vagas em termos de delimitação na CD, e seleccionam as alternativas que conduzem à melhor pontuação para cada saída. Finalmente, o **Ida2ID**, o **Ida2MOR** e o **Ida2SEM** processam os alinhamentos finais e calculam os valores agregados das métricas para as três tarefas, respectivamente.

Fase 4: Geração de relatórios

Os resultados finais da avaliação são compilados em relatórios de desempenho que se desejam facilmente interpretáveis. O **Sultão** gera relatórios globais sobre os resultados de todas as saídas (devidamente anonimizadas), enquanto que o **Alcaide** gera relatórios individuais detalhados para cada saída.

19.2 Descrição pormenorizada de cada módulo

19.2.1 Validador

O módulo Validador verifica se o formato dos ficheiros de saída enviados durante a análise corresponde ao formato determinado pelas directivas do HAREM. Os documentos incluídos na saída deverão ter a seguinte estrutura, ilustrada abaixo através de uma DTD.

```
<!ELEMENT DOC ( DOCID, GENERO, ORIGEM, TEXTO ) >
<!ELEMENT DOCID ( #PCDATA ) >
<!ELEMENT GENERO ( #PCDATA ) >
<!ELEMENT ORIGEM ( #PCDATA ) >
<!ELEMENT TEXTO ( #PCDATA ) >
```

O formato adoptado pelo HAREM para estruturar os documentos na Colecção HAREM (CH) e nas respectivas CD é o formato SGML.

Veja-se o seguinte exemplo de um documento válido, ilustrando uma saída de um sistema REM que participasse nas tarefas de classificação semântica e morfológica.

```
<DOC>
<DOCID>HAREM-051-00043</DOCID>
<GENERO>Web</GENERO>
<ORIGEM>PT</ORIGEM>
<TEXTO>
<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Lions Clube de Faro</ORGANIZACAO>
DM-115CS
<LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL>
O <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Algarve</LOCAL> , a região mais a sul
do território continental de <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal
</LOCAL>, tem por capital a cidade de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">
Faro</LOCAL>.
</TEXTO>
</DOC>
```

O Validador tem em atenção as seguintes questões:

- Um `<DOC>` deve ser seguido, por esta ordem exacta, pelas etiquetas `<DOCID>`, `<GENERO>`, `<ORIGEM>` e `<TEXTO>`. Um `<DOC>` não pode conter outro `<DOC>`.
- Um `<DOCID>` deve possuir um único identificador `DOCID`. Este identificador é usado para identificar os documentos da CH, e é composto pela etiqueta `HAREM`, seguida de três caracteres alfanuméricos, e terminando por uma sequência de cinco algarismos. Estas três partes distintas são separadas por hífens. Um exemplo de um identificador `DOCID` válido é `HAREM-87J-07845`.
- A colecção não pode ter dois ou mais documentos com o mesmo `DOCID`.
- O texto marcado pelas etiquetas `<GENERO>` deve ser um dos géneros textuais especificados no ficheiro `harem.conf` (ver apêndice D.3).
- O texto marcado pelas etiquetas `<ORIGEM>` deve ser uma das variantes especificadas no ficheiro `harem.conf` (ver apêndice D.3).
- Dentro da etiqueta `<TEXTO>`, só são permitidas etiquetas válidas para a marcação de EM no texto.
- Se a saída não se referir à tarefa de classificação semântica, só pode conter etiquetas `<EM>`, que podem incluir o atributo opcional `MORF`.
- Se a saída se referir à tarefa de classificação semântica, não pode conter etiquetas `<EM>`. As etiquetas devem ter uma ou mais categorias separadas por um '|', e devem possuir obrigatoriamente o atributo `TIPO` com um ou mais tipos separados por um

'|', em número idêntico. Esses tipos devem corresponder às categorias, pela mesma ordem. A etiqueta pode incluir também o atributo opcional `MORF`.

- Para o atributo MORF, o formato aceite é "*x*,*y*", onde *x* pode tomar os valores M, F ou ?, e *y* os valores S, P ou ?.
- As etiquetas e os atributos devem conter apenas caracteres alfabéticos maiúsculos, além dos caracteres '|' (barra vertical), para especificar mais de uma categoria, e ',' (vírgula), que separa os valores para o género e para o número, dentro do atributo `MORF`, como explicado acima.
- Todos os atributos dos parâmetros TIPO e MORF devem estar delimitados por aspas.
- Não são aceites etiquetas de abertura quando ainda existe uma etiqueta à espera de ser fechada. Por outras palavras, não são aceites EM marcadas dentro de outras EM.

19.2.2 Extractor

O módulo Extractor extrai o subconjunto dos documentos contidos na CD, da saída do sistema dos participantes. No processo, o Extractor ordena os documentos numericamente pelo seu identificador, o `DOCID`, e escreve-os sem alterar o seu conteúdo.

19.2.3 AlinhEM

O módulo AlinhEM tem como objectivo produzir uma lista de *alinhamentos* das EM da saída do sistema com as EM da CD. Alinhamentos são linhas de texto que descrevem a correspondência existente entre as EM de dois documentos (no caso da avaliação do HAREM, entre a saída do sistema e a CD).

A tarefa do AlinhEM é muito importante, uma vez que os módulos seguintes baseiam-se nos alinhamentos gerados por este. Um requisito do AlinhEM é que as colecções de textos a alinhar possuam os mesmos documentos, podendo diferir apenas nas etiquetas de EM colocadas nos textos.

Formato de saída

O AlinhEM processa e escreve cada documento no seguinte formato:

```
HAREM_ID ORIGEM GÉNERO
<VERIFICACAO_MANUAL>Informação para o juiz humano</VERIFICACAO_MANUAL>
Alinhamento 1
Alinhamento 2
(...)
Alinhamento n
```

O AlinhEM escreve uma primeira linha com os seus metadados, uma linha (opcional) para depuração manual, seguida de uma lista de alinhamentos. O documento termina com uma ou mais linhas em branco. Os alinhamentos podem ser de cinco tipos:

um para um: uma EM da CD alinha exactamente com uma EM na saída.

um para muitos: uma EM da CD alinha com mais do que uma EM na saída.

muitos para um: mais do que uma EM da CD alinham com uma EM na saída.

nenhum para um: uma EM é identificada na saída mas não há uma EM correspondente na CD.

um para nenhum: uma EM da CD não foi marcada como tal na saída.

Para cada tipo de alinhamento, o AlinhEM representa cada uma destas situações num formato específico, para facilitar o processamento dos módulos seguintes. Todos os formatos exibem primeiro a correspondência na CD, seguido de um separador '--->' e a(s) correspondências na saída, entre parênteses rectos. Existem cinco formatos diferentes de alinhamentos, um para cada tipo:

1. No caso de um alinhamento do tipo **um para um**, a lista de entidades da saída contém uma EM:

   ```
   <EM>17:00<EM> ---> [<EM>17:00</EM>]
   ```

2. No caso de um alinhamento do tipo **um para muitos**, onde múltiplas EM da saída alinham com uma EM da CD, o alinhamento apresenta as várias EM da saída separadas por vírgulas, como é ilustrado a seguir:

   ```
   <EM>17:00<EM> ---> [<EM>17</EM>, <EM>00</EM>]
   ```

3. No caso de um alinhamento do tipo **muitos para um**, cada EM da CD alinhada é representada numa linha distinta:

   ```
   <EM>17<EM> ---> [<EM>17:00</EM>]
   <EM>00<EM> ---> [<EM>17:00</EM>]
   ```

4. No caso de um alinhamento do tipo **nenhum para um**, ou seja, EM espúrias na saída, esta é marcada com a etiqueta `<ESPURIO>`:

   ```
   <ESPURIO>Ontem</ESPURIO> ---> [<EM>Ontem</EM>]
   ```

5. No caso de um alinhamento do tipo **um para nenhum**, ou seja, EM que não foram identificadas na saída, a EM da CD aponta para uma lista com o termo `null`.

   ```
   <EM>Departamento de Informática</EM> ---> [null]
   ```

Etiquetas <ALT>

Nas situações em que as etiquetas `<ALT>` foram usadas na CD, o AlinhEM faz o alinhamento para cada alternativa, e marca cada uma das alternativas com uma etiqueta `<ALT`*n*`>`, com *n* a ser o número incremental da alternativa. De seguida pode-se ver exemplos de alternativas escritas pelo AlinhEM. A selecção da melhor alternativa é posteriormente realizada pelos módulos AltinaID, AltinaMOR e AltinaSEM.

Segue-se um exemplo de alternativas para um alinhamento do tipo **um para um**, com uma EM vaga na CD, para o caso em que na CD esteja `<ALT> <EM>98 anos</EM> e meio | <EM>98 anos e meio</EM> </ALT>` e a saída do sistema tenha sido `<EM> 98 anos </EM>`:

```
<ALT>
<ALT1>
<VALOR TIPO="QUANTIDADE">98 anos e meio</VALOR> ---> [<VALOR TIPO="QUANTIDADE">98 anos</VALOR>]
</ALT1>
<ALT2>
<VALOR TIPO="QUANTIDADE">98 anos</VALOR> ---> [<VALOR TIPO="QUANTIDADE">98 anos</VALOR>]
<ALT2>
</ALT>
```

O próximo é um exemplo de alternativas para um alinhamento do tipo **um para um** ou do tipo **muitos para um**, uma ou mais EM vagas na CD, para o caso em que na CD esteja `<ALT> <EM> Aves-Campomaiorense </EM> | <EM> Aves </EM> - <EM> Campomaiorense </EM> </ALT>` e a saída do sistema tenha sido `<EM> Aves-Campomaiorense </EM>`:

```
<ALT>
<ALT1>
```

```
<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO>]
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Aves</PESSOA> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO>]
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Campomaiorense</PESSOA> --->
[<ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Aves-Campomaiorense</ACONTECIMENTO>]
</ALT2>
</ALT>
```

Apresentamos agora um exemplo de alternativas para um alinhamento do tipo **nenhum para nenhum** ou do tipo **um para nenhum**, uma ou nenhuma EM na CD, para o caso em que na CD esteja `<ALT> Monárquico | <EM> Monárquico </EM> <ALT>` e a saída do sistema tenha sido `Monárquico`:

```
<ALT>
<ALT1>
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA> ---> [null]
</ALT2>
</ALT>
```

Finalmente, eis um exemplo de alternativas para um alinhamento do tipo **nenhum para um** ou do tipo **um para um**, uma ou nenhuma EM na CD, para o caso em que na CD esteja `<ALT> Monárquico | <EM> Monárquico </EM> <ALT>` e a saída do sistema tenha sido `<EM> Monárquico </EM>`:

```
<ALT>
<ALT1>
<ESPURIO>Monárquico</ESPURIO> --->
[<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA>]
</ALT1>
<ALT2>
<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA> --->
[<PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Monárquico</PESSOA>]
</ALT2>
</ALT>
```

Etiquetas <OMITIDO>

A etiqueta `<OMITIDO>` foi introduzida na versão 2.1 da CD de 2005, em plena avaliação do HAREM, por se ter achado necessário ignorar certos excertos de texto sem qualquer interesse do ponto de vista linguístico, sem interferir com a avaliação do HAREM. Assim, as

etiquetas <OMITIDO> identificam esses excertos de texto, alertando os módulos de avaliação para ignorarem o conteúdo. Apresentamos abaixo um exemplo contido num documento oriundo da Web, e que, do ponto de vista da tarefa de REM em português, é inadequado para avaliar o desempenho dos sistemas.

```
<OMITIDO>
Sorry, your browser doesn't support <OBRA TIPO="PRODUTO">Java</OBRA>.
</OMITIDO>
```

Numeração distinta de átomos

O AlinhEM, ao ser executado com a opção `-etiquetas sim`, regista todos os átomos presentes nos alinhamentos de cada documento, e depois numera-os sequencialmente por ordem de aparição. Desta forma, impede-se que haja emparelhamentos de EM com átomos em comum, mas que estão localizados em partes diferentes do documento.

Para ilustrar tais situações, considere-se o seguinte extracto de texto, marcado como uma CD (só para a categoria ORGANIZACAO):

```
<DOC>
<DOCID>HAREM-051-00043</DOCID>
<GENERO>Web</GENERO>
<ORIGEM>PT</ORIGEM>
<TEXTO>
<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S"><1>Lions</1> <1>Clube</1> de <1>Faro</1>
</ORGANIZACAO>

É no Hotel Eva, situado na lateral da marina, que se reune o <ORGANIZACAO TIPO="INSTITUICAO"
MORF="M,S"><2>Clube</2> <2>Lions</2></ORGANIZACAO>, nas primeiras quartas-feiras de cada mês.
</TEXTO>
</DOC>
```

E a respectiva (e hipotética) saída de um sistema participante:

```
<DOC>
<DOCID>HAREM-051-00043</DOCID>
<GENERO>Web</GENERO>
<ORIGEM>PT</ORIGEM>
<TEXTO>
<1>Lions</1> <1>Clube</1> de <1>Faro</1>

É no Hotel Eva, situado na lateral da marina, que se reune o <ORGANIZACAO TIPO="INSTITUICAO"
MORF="M,S"><2>Clube</2> <2>Lions</2></ORGANIZACAO>, nas primeiras quartas-feiras de cada mês.
</TEXTO>
</DOC>
```

O documento da CD tem duas EM, e ambas incluem o átomo *Lions*. Contudo, a saída do sistema apresenta apenas uma EM, com o átomo *Lions*. Se os textos não fossem marcados com etiquetas numéricas, o AlinhEM não tinha informação suficiente para saber qual das EM da CD é que vai alinhar com a EM da saída.

Nos processos de atomização e de etiquetação numérica, o AlinhEM pode ignorar ocorrências de um dado conjunto de átomos. Esta opção permite não só ignorar termos muito frequentes, como também permite ultrapassar situações em que os textos originais das saídas são alterados, especialmente nas suas EM.

O AlinhEM possui uma lista interna de termos a ignorar nas avaliações conjuntas do HAREM, que apresentamos abaixo. Esta lista pode ser complementada com outra lista, segundo a opção `-ignorar`, descrita no apêndice D.2.2.

```
a, A, à, À, ao, AO, Ao, as, AS, As, com, COM, Com, como, COMO, Como,
da, DA, Da, das, DAS, Das, de, DE, De, do, DO, Do, dos, DOS, Dos,
e, E,  é, É, em, EM, Em,for, FOR, For, mais, MAIS, Mais, na, NA, Na,
não, NÃO, Não, no, NO, No, nos, NOS, Nos, o, O, os, OS, Os, ou, OU, Ou,
para, PARA, Para, pela, PELA, Pela, pelo, PELO, Pelo, por, POR, por,
que, QUE, Que, se, SE, Se, um, UM, Um, uma, UMA, Uma.
```

O processo de atomização do AlinhEM não se preocupa em garantir que cada átomo gerado corresponda a algo que faça parte do léxico, uma vez que a preocupação principal é o alinhamento correcto das EM. O AlinhEM pode mesmo partir palavras e números em locais que os atomizadores para a língua portuguesa não o fariam. O AlinhEM utiliza as seguintes regras de atomização:

1. Todos os caracteres não alfa-numéricos são considerados delimitadores de átomos.

   ```
   alguem@algures.com ->  <1>alguem</1> @ <1>algures</1> . <1>com</1>
   ```

2. Todos os números são atomizados ao nível do dígito.

   ```
   1979 ->  <1>1</1> <1>9</1> <1>7</1> <2>9</2>
   1.975 -> <1>1</1> . <1><9/1> <1><7/1> <1><5/1>
   ```

3. A transição de um caracter numérico para um alfabético (ou vice-versa) delimita átomos.

   ```
   NBR6028 -> <1>NBR</1> <1>6</1>  <1>0</1> <1>2</1> <1>8</1>
   ```

Etiquetas <VERIFICACAO_MANUAL>

A etiqueta `<VERIFICACAO_MANUAL>` é gerada quando o AlinhEM é executado com a opção `-etiquetas sim`, e no final da etiquetagem numérica aos átomos do mesmo documento

na CD e na saída, os números finais não coincidem. Isto normalmente sugere que o texto original da saída foi alterado, o que pode impedir o alinhamento correcto das EM. Quando tal acontece, os alinhamentos com as etiquetas numéricas discordantes são envolvidas em etiquetas <VERIFICACAO_MANUAL>, para que sejam inspeccionados manualmente de forma a que a origem do problema seja identificada. Estas etiquetas são ignoradas pelos módulos seguintes.

19.2.4 AvalIDa

O módulo AvalIDa avalia e pontua os alinhamentos produzidos pelo AlinhEM, segundo as directivas de avaliação para a tarefa de identificação. Para tal, o AvalIDa acrescenta no final de cada alinhamento a respectiva pontuação dentro de parênteses rectos, com um caracter de dois pontos como separador, como é exemplificado abaixo:

```
<EM>17:00<EM> ---> [<EM>17:00</EM>]:[Correcto]
```

No caso de um alinhamento do tipo **um para muitos**, as várias pontuações são separadas por vírgulas, como é mostrado no exemplo abaixo. Este caso é sintomático de pontuações parcialmente correctas, que é complementado com a informação do valor do *factor de correcção* e do *factor de erro* (ver a secção 18.2.1):

```
<EM>17:00<EM> ---> [<EM>17</EM>, <EM>00</EM>] :[Parcialmente_Correcto_
por_Defeito(0.25; 0.75), Parcialmente_Correcto_por_Defeito(0.25; 0.75)]
```

Existem, no entanto, casos que requerem um processamento mais cuidado, como o caso exemplificado abaixo:

```
<EM>Gabinete do Instituto</EM> ---> [<EM>Gabinete do Instituto da Juventude
em Lisboa</EM>]:[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
<EM>Juventude em Lisboa</EM> --->  [<EM>Gabinete do Instituto da Juventude
em Lisboa</EM>]:[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
```

Este exemplo apresenta uma EM da saída (<EM>Gabinete do Instituto da Juventude em Lisboa</EM>) alinhada com duas EM da CD (<EM>Gabinete do Instituto</EM> e <EM>Juventude em Lisboa</EM>). Como o alinhamento é representado em duas linhas, os módulos seguintes (como por exemplo, o Ida2ID) precisam de saber se as duas linhas se referem a um único alinhamento (uma situação **muitos para um**) ou a dois alinhamentos (duas situações **um para um**), evitando cair no erro de contar mais de uma vez a mesma EM. O AvalIDa distingue entre as duas situações usando as etiquetas numéricas produzidas pelo AlinhEM.

19.2.5 Véus

O módulo Véus permite seleccionar criteriosamente grupos de documentos com determinadas características, tais como o seu género textual (Web, Jornalístico, etc) ou a sua variante (PT, BR, etc), ou filtrar os alinhamentos segundo as classificações semânticas das etiquetas das EM, permitindo a avaliação do desempenho do sistema segundo um determinado leque de categorias/tipos.

É dessa forma que o HAREM permite avaliar os sistemas segundo um cenário *selectivo*, ou seja, comparando a saída sobre a CD segundo o universo das EM de categoria/tipo que o sistema se propõe tentar identificar/classificar, e não segundo o universo total das EM. Além disso, o Véus ainda permite parametrizar as avaliações em três estilos: Além do do HAREM, descrito no presente capítulo e volume, também permite uma avaliação "relaxada" em que apenas o primeiro valor de um alinhamento com EM parcialmente correctas é contabilizado, e uma avaliação estilo "muc" em que nenhum caso parcialmente correcto é contabilizado (são todos considerados errados, veja-se Douthat (1998)).

Filtragem por género textual ou por variante

Quando o Véus é executado apenas com um filtro por género textual ou variante, apenas os cabeçalhos dos documentos são analisados, para decidir se o documento é ignorado ou se é copiado para a saída.

Nesse caso, a primeira linha escrita pelo Véus contém a informação sobre todas as categorias e tipos utilizadas na avaliação (ou seja, a repetição das categorias e tipos especificados no ficheiro `harem.conf`). A linha é ilustrada abaixo (o exemplo está abreviado para facilitar a leitura):

```
#PESSOA=["MEMBRO", "GRUPOIND", "CARGO", "GRUPOCARGO", "INDIVIDUAL",
 "GRUPOMEMBRO"]; LOCAL=["GEOGRAFICO", "ALARGADO", "ADMINISTRATIVO",
 "VIRTUAL", "CORREIO"]; (...)
```

Filtragem por categorias e tipos semânticos

Quando o Véus é executado com um filtro por categorias e/ou tipos, a primeira linha da saída do Véus reproduz todas as categorias e tipos aceites, para que não se perca a informação sobre o tipo de filtro aplicado e que originou o resultado do Véus.

Se, por exemplo, o Véus fosse executado com um filtro para obter apenas alinhamentos contendo a categoria `ORGANIZACAO` com todos os seus quatro tipos, e `LOCAL` nos seus tipos `GEOGRAFICO`, `ADMINISTRATIVO`, `CORREIO` e `ALARGADO` (ou seja, todos excepto o `VIRTUAL`), como é ilustrado na Figura 19.2, a primeira linha da saída do Véus seria:

```
#ORGANIZACAO=["INSTITUICAO", "ADMINISTRACAO", "SUB", "EMPRESA"];
LOCAL=["GEOGRAFICO", "ADMINISTRATIVO", "CORREIO", "ALARGADO"]
```

O símbolo '#' no início de cada ficheiro gerado pelo Véus indica aos módulos de avaliação seguintes qual o *cenário* de avaliação especificado, para efeitos de avaliação semântica e morfológica.

De seguida, o Véus filtra todos os alinhamentos previamente identificados, extraindo o subconjunto de alinhamentos que interessa considerar. Note-se que a filtragem por categorias só faz sentido quando o sistema em causa efectuou a respectiva classificação semântica, ou seja, quando a etiqueta genérica `<EM>` não é usada.

19.2.6 ALTinaID

O módulo ALTinaID analisa as alternativas na tarefa de identificação, marcadas com `<ALT>`, e selecciona a alternativa segundo os critérios descritos no capítulo 18. A alternativa escolhida é a única escrita como resultado do programa. As etiquetas `<ALT>` e `<ALT`*n*`>` também são eliminadas.

19.2.7 Ida2ID

O módulo Ida2ID calcula os valores das métricas de avaliação para a tarefa de identificação, fornecendo dados para aferir o desempenho do sistema REM participante.

O funcionamento do Ida2ID pode ser dividido em dois passos: em primeiro lugar, o Ida2ID percorre todos os alinhamentos do ficheiro fornecido, realizando várias contagens. No segundo passo, usa os valores finais dos contadores para chegar aos valores das métricas de avaliação.

À primeira vista, a tarefa do Ida2ID parece simples. Contudo, o formato usado para representar os alinhamentos pode induzir o Ida2ID à contagem errada de EM. Estes casos potencialmente problemáticos normalmente verificam as seguintes condições:

1. O alinhamento em consideração foi pontuado como `parcialmente_correcto`;
2. O alinhamento imediatamente anterior ao que está a ser considerado também foi pontuado como `parcialmente_correcto`;
3. A EM da saída identificada no alinhamento imediatamente anterior é idêntica à EM da saída identificada no alinhamento que está a ser considerado.

Quando estas três condições se verificam, o Ida2ID precisa de decidir se está na presença de uma EM nova, ou se está na presença da mesma ocorrência da EM anterior. Para decidir, o Ida2ID averigua se existe alguma sobreposição das EM da CD, com o auxílio das etiquetas numéricas.

Considere-se o seguinte exemplo (hipotético) de um alinhamento do tipo **muitos para um**:

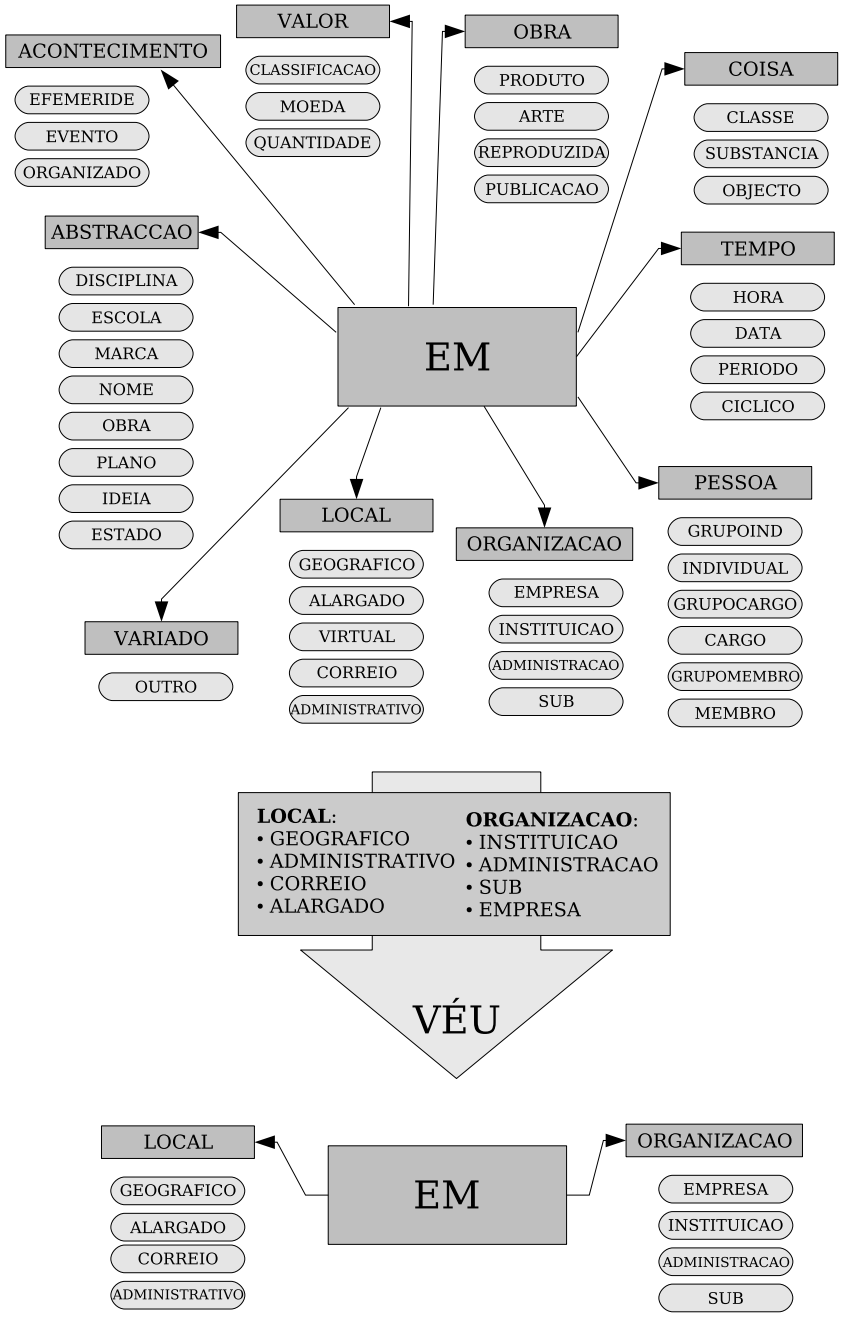


Figura 19.2: Esquema de um exemplo do processamento efectuado pelo Véus.

```
<EM><3>Gabinete</3> do <2>Instituto</2></EM> ---> [<EM><3>Gabinete
</3> do <2>Instituto<2> da <1>Juventude</1> em <5>Lisboa</5></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
<EM><1>Juventude</1> em <5>Lisboa</5></EM> ---> [<EM><3>Gabinete
</3> do <2>Instituto<2> da <1>Juventude</1> em <5>Lisboa</5></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
```

Com a ajuda das etiquetas numéricas, o Ida2ID consegue determinar que as duas linhas referem-se à mesma EM da saída, uma vez que essa EM, tal como está representada, refere-se à EM que contém a 3ª ocorrência do átomo 'Gabinete', ou a 5ª ocorrência do átomo 'Lisboa'. Como tal, o contador das EM de saídas do Ida2ID faz uma correcção e conta apenas uma EM na saída.

Agora, considere-se também o seguinte exemplo de dois alinhamentos do tipo **um para um**:

```
<EM><3>Gabinete</3> do <2>Instituto</2></EM> ---> [<EM><3>Gabinete
</3> do <2>Instituto<2> da <1>Juventude</1> em <5>Lisboa</5></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
<EM><2>Juventude</2> em <6>Lisboa</6></EM> ---> [<EM><4>Gabinete
</4> do <3>Instituto<3> da <2>Juventude</2> em <6>Lisboa</6></EM>]:
[Parcialmente_Correcto_por_Excesso(0.21; 0.79)]
```

Neste exemplo, há duas EM da saída alinhadas respectivamente com outras duas EM na CD. As etiquetas numéricas mostram que na saída há duas ocorrências de uma EM e, como tal, o Ida2ID conta duas EM na saída.

Um exemplo de um relatório (fictício) produzido pelo Ida2ID para um sistema, contendo as várias contagens e avaliações a levar em conta é apresentado em seguida:

```
Total na CD: 4995
Total Identificadas: 2558
Total Correctos: 1927
Total Ocorrências Parcialmente Correctos: 601
Soma Parcialmente Correctos: 128.57140579578655
Soma Parcialmente Incorrectos: 472.42859420421337
Espúrios: 73
Em Falta: 2545
Precisão: 0.8035853814682512
Abrangência: 0.41152580696612345
```

```
Medida F: 0.5443059461924498
Sobre-geração: 0.028537920250195466
Sub-geração: 0.5095095095095095
Erro Combinado: 0.600549668520057
```

É de notar que estes cálculos só podem ser efectuados após a escolha da alternativa mais favorável ao sistema, realizada pelo AltinaID. Esta escolha influencia o número total de entidades encontradas na CD, o que também implica que saídas diferentes podem ser avaliadas segundo diferentes conjuntos de EM da CD. Contudo, estas diferenças saldam-se sempre no favorecimento de cada sistema.

19.2.8 Emir

O módulo Emir pode ser considerado o homólogo do AvalIDa e do Vizir, mas para a avaliação da tarefa de classificação semântica, ao pontuar cada alinhamento segundo a classificação semântica das EM.

O Emir recebe os resultados gerados pelo AvalIDa, filtrados pelo Véus. A primeira linha desses resultados, que contém a informação sobre as categorias e/ou tipos usados no cenário da avaliação, é usada para efectuar o cálculo das várias medidas de avaliação correspondentes à classificação semântica.

Depois de avaliar o alinhamento em relação à classificação semântica, o Emir concatena o resultado no fim do alinhamento, usando um formato semelhante ao do AvalIDa. Considere-se o seguinte alinhamento hipotético gerado pelo AvalIDa:

```
<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL> ---> [<LOCAL TIPO="GEOGRAFICO">
Lisboa e Porto</LOCAL>]:[Parcialmente_Correcto_por_Excesso(0,6666; 0,3333)]
```

No seu processamento, o Emir retira a pontuação respeitante à tarefa de identificação (no exemplo dado, `[Correcto]`), e substitui-a por um novo resultado referente à tarefa de classificação semântica. Esse novo resultado é colocado no final do alinhamento, com dois pontos (`:`) como separador. Este resultado pode conter uma ou mais avaliações (uma por cada EM), e cada avaliação fica envolvida entre chavetas (`{}`).

O resultado da avaliação da classificação semântica, adicionado no final do alinhamento, contém quatro campos:

1. a lista de categorias que foram pontuadas como `correcto`, `espurio` ou `em_falta`;
2. a lista de tipos que foram pontuadas como `correcto`, `espurio` ou `em_falta`;
3. o valor da classificação semântica combinada (CSC), uma das quatro medidas de classificação semânticas adoptadas pelo HAREM (consulte-se a secção 18.3.2para mais detalhes sobre a CSC e o seu cálculo);

4. o valor do peso da EM. Para mais informação sobre o cálculo deste, consulte-se o capítulo 18.

No final, o alinhamento processado pelo Emir pode apresentar o seguinte aspecto:

```
<LOCAL TIPO="ADMINISTRATIVO">Lisboa</LOCAL> ---> [<LOCAL TIPO="GEOGRAFICO">
Lisboa e Porto</LOCAL>]:[{Categoria(Correcto:[LOCAL] Espúrio:[] Em_Falta:[])
Tipo(Correcto:[] Espúrio:[GEOGRAFICO] Em_Falta:[ADMINISTRATIVO]) CSC(1.0)
Peso(0.66)}]
```

No caso de alinhamentos **um para muitos**, o Emir escreve os vários resultados da avaliação da forma que se apresenta no exemplo abaixo, separados por vírgulas (,):

```
<LOCAL TIPO="ADMINISTRATIVO">Lisboa e Porto</LOCAL> --->
[<LOCAL TIPO="GEOGRAFICO">Lisboa</LOCAL>, <LOCAL TIPO="GEOGRAFICO">
Porto</LOCAL>]:[{Categoria(Correcto:[LOCAL] Espúrio:[] Em_Falta:[])
Tipo(Correcto:[] Espúrio:[GEOGRAFICO] Em_Falta:[ADMINISTRATIVO])
CSC(1.0) Peso(0.33)}, {Categoria(Correcto:[LOCAL] Espúrio:[]
Em_Falta:[]) Tipo(Correcto:[] Espúrio:[GEOGRAFICO] Em_Falta:
[ADMINISTRATIVO]) CSC(1.0) Peso(0.33)}]
```

No caso de haver vagueza na classificação semântica, ou seja, a EM possuir mais do que uma categoria ou tipo, estas são tratadas como se fossem uma classificação única, como exemplificado abaixo:

```
<ORGANIZACAO|ABSTRACCAO TIPO="SUB|IDEIA">Lisboa</ORGANIZACAO|
ABSTRACCAO> ---> [<LOCAL TIPO="ADMINISTRATIVO">Lisboa e Porto
e Faro e Braga</LOCAL>]:[{Categoria(Correcto:[] Espúrio:[LOCAL]
Em_Falta:[ORGANIZACAO|ABSTRACCAO]) Tipo(Correcto:[] Espúrio:[]
Em_Falta:[]) CSC(0.0) Peso(0.142)}]
```

Quando o Emir é executado **sem** a opção de cenário relativo, os alinhamentos espúrios são contabilizados pelo Emir, que considera todas as categorias e tipos como `espurio`. Um alinhamento como este:

```
<ESPURIO>DM-115CS</ESPURIO> ---> [<ABSTRACCAO TIPO="MARCA"
MORF="F,S">DM-115CS</ABSTRACCAO>]:[Espúrio]
```

é convertido pelo Emir (se não se optar pelo cenário relativo) para:

```
<ESPURIO>DM-115CS</ESPURIO> ---> [<ABSTRACCAO TIPO="MARCA"
MORF="F,S">DM-115CS</ABSTRACCAO>]:[{Categoria(Correcto:[]
Espúrio:[ABSTRACCAO] Em_Falta:[]) Tipo(Correcto:[] Espúrio:[]
Em_Falta:[]) CSC(0.0) Peso(0.0)}]
```

Da mesma forma que acontece com alinhamentos espúrios quando o Emir é executado **sem** a opção de cenário relativo, o Emir também considera e escreve as categorias e tipos `em_falta` quando as EM não foram identificadas, como se pode ver no seguinte exemplo:

```
<LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">Pinheiros</LOCAL> ---> [null]:[Em_Falta]
```

o alinhamento é convertido (se não se optar pelo cenário relativo) para:

```
<LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">Pinheiros</LOCAL> --->
[null]:[{Categoria(Correcto:[] Espúrio:[] Em_Falta:[LOCAL])
Tipo(Correcto:[] Espúrio:[] Em_Falta:[]) CSC(0.0) Peso(0.0)}]
```

No apêndice E.1 apresentam-se mais exemplos do processamento do Emir.

19.2.9 AltinaSEM

O módulo AltinaSEM, de um modo análogo aos módulos AltinaID e AltinaMOR, recebe os resultados do Emir e processa os alinhamentos marcados com etiquetas `<ALT>`, escolhendo as melhores alternativas para cada saída. Os critérios tomados em consideração na escolha da melhor alternativa estão descritos na página 18.2.4 do capítulo 18, e ao contrário do AltinaID, tomam em consideração os valores calculados pelo Emir para a tarefa de classificação semântica, no processo de selecção da melhor alternativa. A alternativa escolhida é escrita, enquanto que as restantes alternativas são eliminadas, tal como as etiquetas `<ALT>` e `<ALT`*n*`>`.

19.2.10 Ida2SEM

O módulo Ida2SEM é o avaliador global da tarefa de classificação semântica, ao calcular os valores das métricas, fornecendo dados para aferir o desempenho do sistema. Tal como o Ida2ID e Ida2MOR, a execução do Ida2SEM pode ser dividida em duas fases: i) todos os alinhamentos avaliados relativamente à classificação semântica são processados, procedendo-se a várias contagens; ii) os contadores são usados para calcular as métricas e gerar um relatório.

De seguida, reproduz-se um exemplo hipotético de um relatório gerado pelo Ida2SEM, que possui as seguintes informações:

1. O domínio da avaliação: quais as categorias e tipos a avaliar;
2. A avaliação referente à classificação semântica por categorias;
3. A avaliação referente à classificação semântica por tipos;
4. A avaliação referente à classificação semântica combinada;

5. A avaliação referente à classificação semântica plana.

```
Avaliação Global - Classificação Semântica por Categorias
Total de EMs classificadas na CD: 5004
Total de EMs classificadas pelo sistema: 5269
Total Correctos: 3120
Espúrios: 1866
Em Falta: 1832
Precisão: 0.5922527110682176
Abrangência: 0.6236170133130373
Medida F: 0.6075303289435293
Sobre-geração: 0.34937277663358923
Sub-geração: 0.36610711430855314

Avaliação Global - Classificação Semântica por Tipos
Total de EMs classificadas na CD: 3440
Total de EMs classificadas pelo sistema: 3448
Total Correctos: 2641
Espúrios: 599
Em Falta: 631
Precisão: 0.7660720776326169
Abrangência: 0.7678536406038555
Medida F: 0.7669618245288219
Sobre-geração: 0.17237410071942447
Sub-geração: 0.18343023255813953

Avaliação Global - Classificação Semântica Combinada
Valor máximo possível para a Classificação Semântica Combinada
   na CD: 8987.450000000072
Valor máximo possível para a Classificação Semântica Combinada
   do sistema: 7309.8648131094515
Precisão Máxima do Sistema: 0.7081119047925152
Abrangência Máxima na CD: 0.5759367002416341
Medida F: 0.6352214896681005

Avaliação Global - Classificação Semântica Plana
Total de EMs classificadas na CD: 23
Total de EMs classificadas pelo sistema: 23
Total Correctos: 1.8403361344537812
Espúrios: 18
Em Falta: 17
Precisão: 0.08001461454146874
```

```
Abrangência: 0.08001461454146874
Medida F: 0.08001461454146874
Sobre-geração: 0.782608695652174
Sub-geração: 0.7391304347826086
```

19.2.11 Vizir

O módulo Vizir faz a avaliação da tarefa de classificação morfológica, de uma forma análoga ao Emir na tarefa de classificação semântica, e ao AvalIDa na tarefa de identificação. Para tal, o Vizir pontua os alinhamentos cujas EM possuem atributos `MORF`.

O Vizir retira toda a informação semântica contida na EM, substituindo as categorias pela etiqueta `<EM>`, e eliminando os atributos `TIPO`. Esta etapa é ilustrada no seguinte exemplo abaixo, onde o alinhamento:

```
<LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Rússia</LOCAL> --->
[<LOCAL TIPO="ALARGADO" MORF="F,S">Rússia</EM>]
```

é convertido pelo Vizir na seguinte linha:

```
<EM MORF="F,S">Rússia</EM> ---> [<EM MORF="F,S">Rússia</EM>]
```

Em seguida, o Vizir, tal como o Emir, remove dos alinhamentos os resultados respeitantes à tarefa de identificação, substituindo-os por novos resultados referentes à tarefa de classificação morfológica. Esses resultados detalham as pontuações e valores para as três medidas usadas: Género, Número e Combinada. Os critérios de atribuição de pontuação e do respectivo valor para cada medida encontram-se detalhados na secção 18.4.2 deste livro.

O seguinte caso exemplifica o resultado da avaliação do Vizir:

```
<ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Lions Clube de
Faro</ORGANIZACAO> ---> [<PESSOA TIPO="GRUPOMEMBRO" MORF="F,S">
Lions Clube de Faro</PESSOA>]:[Correcto]
```

O alinhamento é convertido em:

```
<EM MORF="M,S">Lions Clube de Faro</EM> ---> [<EM MORF="F,S">
Lions Clube de Faro</EM>]:[(Género: Incorrecto 0) (Número:
Correcto 1) (Combinada: Incorrecto 0)]
```

Para mais exemplos do processamento do Vizir, consulte-se o apêndice E.2.

19.2.12 AltinaMOR

O módulo AltinaMOR, de um modo análogo ao AltinaID e ao AltinaSEM, recebe os resultados do Vizir e processa os alinhamentos marcados com etiquetas `<ALT>`, escolhendo as melhores alternativas para cada saída. Os critérios tomados em consideração na escolha da melhor alternativa estão descritos no capítulo 18, e tomam agora em consideração os valores calculados pelo Vizir para a tarefa de classificação morfológica.

19.2.13 Ida2MOR

O módulo Ida2MOR, de um modo análogo ao Ida2ID e ao Ida2SEM, calcula e gera os resultados globais para a tarefa de classificação morfológica. Para tal, o Ida2MOR processa os alinhamentos gerados pelo AltinaMOR, contando as pontuações calculadas. O relatório produzido pelo Ida2MOR apresenta os valores das métricas para as medidas de avaliação da tarefa de classificação morfológica: género, número, e combinada. Em seguida apresentamos um exemplo de um relatório do Ida2MOR:

```
RELATÓRIO DA AVALIAÇÃO DA CLASSIFICAÇÃO MORFOLÓGICA
Gerado em: 25 de Maio de 2005

Avaliação Global da Classificação Morfológica - Número

Total de classificações da CD: 111
Total de classificações do sistema : 92

Precisão: 0.940217391304348
Abrangência: 0.779279279279279
Medida F: 0.852216748768473
Sobre-especificação: 0
Sobre-geração: 0
Sub-geração: 0.171171171171171


Avaliação Global da Classificação Morfológica - Género

Total de classificações da CD: 88
Total de classificações do sistema : 92

Precisão: 0.652173913043478
Abrangência: 0.681818181818182
Medida F: 0.666666666666667
Sobre-especificação: 0.25
```

```
Sobre-geração: 0
Sub-geração: 0.215909090909091


Avaliação Global da Classificação Morfológica - Combinada

Total de classificações da CD: 111
Total de classificações do sistema : 92

Precisão: 0.652173913043478
Abrangência: 0.540540540540541
Medida F: 0.591133004926108
```

19.2.14 Sultão

O módulo Sultão tem por objectivo interpretar todos os relatórios globais gerados, e resumir os valores obtidos por todos os participantes na forma de tabelas, de modo a fornecer resultados comparativos da avaliação conjunta do HAREM. O Sultão é composto por três programas dedicados a cada tarefa de avaliação: o SultãoID, o SultãoMOR e o SultãoSEM, respectivamente para as tarefas de identificação, de classificação morfológica e de classificação semântica.

O Sultão precisa de ler os resultados dos vários sistemas segundo vários cenários para poder gerar os relatórios globais, pelo que o seu maior interesse é para os organizadores da avaliação conjunta. Ao resumir os resultados dos participantes, foi implementada no Sultão a opção de substituir o nome das saídas por pseudónimos, gerando também uma chave para poder desvendar os mesmos (Para conservar o anonimato dos resultados, esta chave deve naturalmente ser separada dos ficheiros, antes de serem divulgados).

As tabelas geradas pelo Sultão recorrem aos seguintes estilos:

1. os pseudónimos a **negrito** identificam as saídas consideradas oficiais, ou seja, as saídas enviadas durante a avaliação conjunta dentro do prazo estipulado;
2. os pseudónimos a *itálico* identificam os resultados no cenário selectivo escolhido para a saída em causa;
3. os valores a verde identificam os melhores para a métrica em questão.

A tabela 18.25 da secção 18.5.1 é um exemplo de tabelas geradas pelo Sultão.

19.2.15 Alcaide

O módulo Alcaide tem por objectivo gerar relatórios individuais para cada saída que participou no HAREM. Para tal, o Alcaide lê e processa os relatórios gerados pelo Sultão e

os relatórios gerados pelos módulos Ida2ID, Ida2MOR e Ida2SEM, organizando-os num único relatório composto por tabelas e gráficos, sub-dividido por tarefas, formas de avaliação, cenários, categorias, géneros textuais e variantes.

A saída do Alcaide consiste num relatório final em HTML, que resume o desempenho de uma saída, nas tarefas que esta se propôs realizar, nos diversos cenários e formas de avaliação discriminada por categoria, género textual e variante. Tal como o Sultão, o Alcaide é um módulo vocacionado para ser utilizado pela organização do HAREM, uma vez que os seus relatórios são apresentados de uma forma comparativa, que, para ser compilada, exige o acesso aos resultados dos outros participantes.

As tabelas 18.26 a 18.28 e as figuras 18.1 a 18.5 da secção 18.5.2 são exemplos de tabelas e gráficos gerados pelo Alcaide.

19.3 Comentários finais

Os programas aqui descritos foram desenvolvidos pelo primeiro autor (Véus, AlinhEM, AvalIDa, Ida2ID, AltinaID, Emir, AltinaSEM, Ida2SEM e Sultão), pelo segundo autor (Validador, Extractor e Alcaide) e pelo terceiro autor (Vizir, AltinaMOR e Ida2MOR), e testados exaustivamente pela quarta autora, com a ajuda dos primeiros.

Estes programas encontram-se acessíveis no sítio do HAREM, e a informação técnica para a sua utilização está patente no apêndice D.2.

Congratulamo-nos com o facto de existirem já alguns utilizadores que os usam rotineiramente, e esperamos que possam vir a ser usados, com poucas alterações, em futuras edições do HAREM.

Agradecimentos

Este capítulo foi escrito no âmbito da Linguateca, financiada pela Fundação para a Ciência e Tecnologia através do projecto POSI/PLP/43931/2001, co-financiado pelo POSI, e pelo projecto POSC 339/1.3/C/NAC.

Capítulo 20

Disponibilizando a <OBRA>Colecção Dourada</OBRA> do <ACONTECIMENTO> HAREM </ACONTECIMENTO> através do projecto <LOCAL | ORGANIZACAO | ABSTRACCAO> AC/DC </LOCAL | ORGANIZACAO | ABSTRACCAO>

Paulo Rocha e Diana Santos

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 20, p. 307–326, 2007.

Ao concertar dois projectos caros à Linguateca (o HAREM e o AC/DC) num único recurso, este capítulo tem dois objectivos distintos:

1. Disponibilizar a colecção dourada do HAREM num formato mais amigável para a sua exploração por uma comunidade mais abrangente, e apresentar alguma informação quantitativa que permitirá avaliar a dificuldade subjacente ao Primeiro HAREM;

2. Produzir documentação mais actualizada sobre o projecto AC/DC, descrevendo como codificar (e consequentemente usar) outro tipo de informação (a que chamamos informação estrutural) a partir de uma colecção anotada, e cujo processo até agora nunca tinha sido descrito em pormenor.

Este capítulo começa por descrever brevemente o projecto AC/DC, explicando os motivos para disponibilizar a colecção dourada como um corpus. De seguida, é feita uma pequena introdução ao formalismo subjacente ao AC/DC, para explicar as opções tomadas na codificação da colecção dourada (ilustradas com exemplos de procuras não triviais no âmbito do corpus CDHAREM). O capítulo termina por uma descrição quantitativa da colecção dourada (e das colecções douradas parciais que foram usadas em 2005 e 2006), de forma a contribuir para uma caracterização e medição rigorosas do problema que os sistemas tentaram resolver no HAREM.

20.1 O projecto AC/DC

O projecto AC/DC, Acesso a Corpora/Disponibilização de Corpora (Santos e Bick, 2000; Santos e Sarmento, 2003) é um projecto que pretende facilitar o acesso a corpora em português, tanto para o utilizador casual, como para o investigador na área. O AC/DC disponibiliza todos os corpora que a Linguateca possui num ponto único de acesso , num formato pensado para ser usado por seres humanos.

Este projecto teve início em 1998, e o número de corpora disponibilizados tem crescido sustentadamente desde essa data; actualmente, é possível consultar no sítio do AC/DC (`http://www.linguateca.pt/ACDC/`) cerca de vinte corpora, através de uma interface simples e padronizada. Estes corpora, na sua maioria criados por entidades exteriores à Linguateca, abrangem vários géneros textuais e proveniências, e incluem alguns de grande dimensão, nomeadamente o CETEMPúblico (Rocha e Santos, 2000) com mais de 180 milhões de palavras de texto jornalístico em português europeu, e o Corpus NILC/São Carlos, com mais de 32 milhões de palavras em português do Brasil, bem como outros corpora de menor dimensão mas geralmente com mais informação linguística associada. Embora não fazendo estritamente parte do AC/DC, convém referir que também o COMPARA (Frankenberg-Garcia e Santos, 2002), um corpus paralelo de textos literários em português e inglês, e a Floresta Sintá(c)tica (Bick et al., 2007) se podem considerar continuadores do

AC/DC, no sentido de que resultam de uma estratégia de enriquecimento deste, mantendo a filosofia original.

Note-se que os corpora do AC/DC permitem também a criação de outros recursos, como é exemplo a própria Colecção HAREM, em cuja compilação vários corpora do AC/DC foram empregues, ou a colecção dourada usada nas Morfolimpíadas (Santos et al., 2003; Costa et al., 2007).

Cremos poder afirmar que o projecto AC/DC tem cumprido a sua missão, ao registar cerca de 6.000 acessos mensais em Abril de 2007, totalizando cerca de 250.000 acessos desde o seu início.

20.1.1 A criação de um corpus novo no AC/DC

Os corpora, como simples conjunto de textos, só permitem realizar consultas simples, como, por exemplo, verificar as concordâncias de uma determinada unidade no corpus e quantas vezes ocorre. Assim, de modo a permitir consultas mais elaboradas, os corpora do AC/DC são enriquecidos com informação adicional relevante.

Em primeiro lugar, os corpora são anotados gramaticalmente com o analisador sintáctico PALAVRAS (Bick, 2000), que adiciona informação complementar, tal como o lema ou a categoria gramatical de cada palavra existente nos corpora, o género ou o tempo verbal, ou a função sintáctica dos vários constituintes.

De igual modo, aplicam-se a todos os corpora procedimentos sistemáticos e rigorosos de atomização e separação de frases em português[1]. São também geradas listas de formas e lemas presentes nos corpora.

Além disso, alguns corpora são marcados com anotações adicionais, como por exemplo o período de tempo a que se referem, o país de origem ou a fonte dos textos, permitindo restringir as procuras a uma subsecção do corpus. As anotações utilizadas pelo corpus da CD do HAREM são descritas na secção 20.2.[2]

20.1.2 IMS-CWB, o sistema subjacente

Os corpora são compilados usando o IMS Corpus Workbench ou IMS-CWB[3](Christ et al., 1999; Evert, 2005), que se revelou robusto e eficiente para os nossos propósitos (Santos e Ranchhod, 1999). O IMS-CWB é detentor de uma linguagem poderosa de interrogação de corpora através do seu módulo *Corpus Query Processor* (CQP), permitindo codificar a informação associada a um corpus de duas formas complementares: atributos estruturais e atributos posicionais.

[1] No sítio do AC/DC pode ser encontrada informação mais detalhada sobre os critérios de separação em frases e sobre as ferramentas usadas para essa tarefa.

[2] Para mais informação sobre os outros corpora, consulte as páginas do AC/DC.

[3] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`

Os **atributos estruturais** referem-se às etiquetas usadas no texto para marcar ou delimitar um subconjunto desse texto. No exemplo abaixo, as etiquetas `PESSOA` e `OBRA` são transformadas em atributos estruturais homónimos.

```
Entre as propostas mais ousadas, decidiu-se pedir ao <PESSOA TIPO="CARGO"
MORF="M,S"> Presidente da República </PESSOA> que proponha um referendo
sobre a <OBRA TIPO="PUBLICACAO" MORF="F,S"> Lei do Financiamento </OBRA>
```

Aos atributos estruturais podem ser associados valores, como por exemplo, `<DOCID id=¨899¨>`. Estes valores podem ser usados na restrição de uma consulta, mas não podem ser alvo de consultas de distribuição.[4]

Os **atributos posicionais** contêm valores que se atribuem a cada unidade no texto. Usando como exemplo o início da frase anterior e o atributo posicional `pos` (categoria gramatical, do inglês *part-of-speech*), obtemos a seguinte linha de texto:

```
Entre/PRP as/DET propostas/N mais/ADV ousadas/ADJ, ...
```

Uma descrição mais completa da sintaxe usada no IMS-CWB pode ser encontrada em `http://acdc.linguateca.pt/acesso/anotacao.html`. Recomendamos vivamente a sua leitura, visto que reúne informação considerável sobre o uso específico do PALAVRAS como anotador no AC/DC e sobre o pós-processamento a que a anotação deste é sujeita. Essa página também remete para vários outros locais de ensino do CQP. Mencione-se, a propósito, que a anotação do PALAVRAS também é a base da parte portuguesa do COMPARA (Santos e Inácio, 2006; Inácio e Santos, 2006) e da Floresta Sintá(c)tica (Afonso et al., 2002; Afonso, 2006), ambas revistas posteriormente.

20.2 Disponibilizando a CD do HAREM como corpus

Apesar de as CD estarem publicamente disponíveis do sítio do HAREM desde o fim das respectivas avaliações conjuntas[5], optámos por disponibilizá-las igualmente como um recurso no âmbito do AC/DC, facilitando assim o acesso à riqueza de informação associada à colecção e adicionando informação morfossintáctica. Tal permite um número de pesquisas na colecção que não seriam fáceis ou práticas de efectuar por um utilizador menos experimentado.

O corpus resultante, denominado CDHAREM, é então formado pelo texto das duas CD usadas nas duas avaliações conjuntas do HAREM, acrescido de toda a informação anexa a esse texto e da sua anotação gramatical.

[4] Ao contrário das concordâncias, onde se pede o texto, nas consultas de distribuição (ou consultas agregadas, em terminologia de bases de dados) pretende-se habitualmente saber a quantidade de vezes que um dado fenómeno ocorre, e qual a distribuição quantitativa dos elementos que satisfazem a procura em questão.

[5] Em `http://poloxldb.linguateca.pt/harem.php?l=coleccaodourada`

Procura: "**Lisboa**".
Distribuicao de **em**
Corpus: Corpus CD HAREM, 0.1

40 casos.
Distribuição
Houve 11 valores diferentes de **em**.

Lisboa	28
Universidade_de_Lisboa	3
Metropolitano_de_Lisboa	1
Universidade_Nova_de_Lisboa	1
Grande_Lisboa	1
Emissores_Associados_de_Lisboa	1
10h00_de_Lisboa	1
Hotel_Lisboa_Plaza	1
Governadora_Civil_do_Distrito_de_Lisboa	1
Instituto_Técnico_de_Lisboa	1
Departamento_de_Matemática_da_Universidade_de_Lisboa	1

Figura 20.1: Distribuição de uma palavra por EM.

20.2.1 Opções gerais de codificação

Na tabela 20.1 apresentamos, de forma condensada e para referência subsequente, a lista de conversões de atributos presente na CD para os formatos usados no AC/DC, com o objectivo de facilitar vários tipos de pesquisa, que nos parecerem especialmente relevantes neste contexto. Como norma geral, para o corpus CDHAREM, foram usadas letras maiúsculas para os atributos estruturais, e minúsculas para os atributos posicionais. A única excepção foram os atributos estruturais <p> e <s>, provenientes da separação de frases. Caso um atributo posicional não se encontre definido para uma determinada unidade, é-lhe atribuído o valor "0".

20.2.2 O atributo EM

O atributo estrutural EM, como o seu próprio nome indica, identifica uma EM, independentemente da sua classificação. A consulta seguinte encontra exclusivamente a EM *Porto*, excluindo assim os casos em que esta palavra faz parte de uma EM maior (por exemplo, *Porto Seguro*):

```
<EM> "Porto" </EM>
```

Na próxima consulta encontramos os casos em que a expressão *São Paulo* é parte de uma EM:

Tipo de atributo	Colecção dourada	Atributo estrutural	Valores	Atributo posicional	Valores
Delimitador de um documento	DOC	DOC	docid= genero= origem=	–	–
Identificação do documento da CD	DOCID...	–	–	–	HAREM-871-07800, etc.
Género de texto	GENERO...	–	–	–	Web, Técnico, etc.
País de origem do texto	ORIGEM...	–	–	–	PT, BR, etc.
Delimitador do texto de um documento	TEXTO	TEXTO	tam=	–	–
Entidade mencionada	LOCAL, PESSOA, etc...	EM	tam=	–	–
Categoria(s) a que pertence a palavra	**OBRA** TIPO="ARTE" MORF="M,S", etc.	LOCAL, PESSOA, etc.	–	categoria,	PESSOA, LOCAL, etc...
Tipo(s) a que pertence a palavra	OBRA TIPO=**"ARTE"** MORF="M,S", etc.	–	–	tipo, local, pessoa, etc.	ADMNISTRATIVO, INSTITUICAO, etc.
Género e número da EM (revisto manualmente)	OBRA TIPO="ARTE" **MORF**="M,S", etc.	–	–	morf	M, S, F, P, etc.
Posição relativa na EM de uma palavra	–	–	–	prem	1,2,...,29
Delimitador de parágrafo	–	p	–	–	–
Delimitador de frase	–	s	–	–	–
Parte de uma anotação alternativa	<ALT> .... | ... </ALT>	ALT	num=	alt	P, M ou F, seguido da categoria da alternativa, ou de 0. POBRA, FPESSOA, M0, etc.

Tabela 20.1: Conversão de atributos da CD do HAREM para o corpus CDHAREM do AC/DC.

```
"São" "Paulo" within EM
```

O atributo EM é codificado no corpus juntamente com o tamanho (em unidades) da EM, como é ilustrado no exemplo abaixo:

```
<EM TAM=3>
<PESSOA>
Presidente
da
República
</PESSOA>
</EM>
```

Para identificar a EM à qual um termo pertence, pode ser usado o atributo posicional `em`. Este atributo assume como valor o texto da EM, com sublinhados a separar as unidades; no exemplo acima, a cada uma das unidades *Presidente*, *da* e *República* é atribuído o valor *Presidente_da_República*. Pode-se assim mais facilmente descobrir a que EM um termo pertence e quantas vezes, tal como no exemplo da Figura 20.1.

20.2.3 Atributos relativos às categorias e tipos das EM

Todas as categorias existentes na CD equivalem a um atributo estrutural distinto. Estes atributos podem ser usados para facilitar a procura de uma determinada categoria de EM; por exemplo, para obter todas as EM de categoria `OBRA`:

```
<OBRA> []* </OBRA>
```

ou todas as EM de três palavras que sejam simultaneamente `ORGANIZACAO` e `LOCAL`:

```
<ORGANIZACAO> <LOCAL> [] [] [] </LOCAL> </ORGANIZACAO>
```

Para facilitar as consultas, usam-se também atributos posicionais para identificar as categorias e tipos, apropriadamente chamados `categoria` e `tipo` respectivamente. O exemplo seguinte mostra os valores do atributo `categoria` para um excerto particular.

```
<s> As/0 ilhas/0 de/0 Cabo/LOCAL Verde/LOCAL foram/0 descobertas/0 por/0
navegadores/0 portugueses/0 em/0 Maio/TEMPO de/TEMPO 1460/TEMPO ,/0 sem/0
indícios/0 de/0 presença/0 humana/0 anterior/0 ./0 </s>
```

No caso de uma EM pertencer a múltiplas categorias ou tipos, eles são listados por ordem alfabética, separados por sublinhados (ver secção 20.3.1).

Além disso, foi definido um atributo posicional para cada uma das categorias, que assumem o valor do tipo correspondente à EM. Os atributos posicionais têm o mesmo

nome dos estruturais, mas em minúsculas (`local`, `pessoa`, etc.). Assim, podemos procurar a palavra *Lisboa* como parte do nome de uma organização mas não parte do nome de um local (o valor "0" implica que o campo não tem um valor definido):

```
[word="Lisboa" & organizacao!="0" & local="0"]
```

Assim como podemos identificar os casos em que à categoria `PESSOA` corresponde o tipo `CARGO` (independentemente de outros):

```
<PESSOA> [pessoa=".*CARGO.*"]+ </PESSOA>
```

Se se quisesse apenas os casos em que `CARGO` é o único tipo, empregar-se-ia a seguinte expressão de consulta:

```
<PESSOA> [pessoa="CARGO"]+ </PESSOA>
```

20.2.4 O atributo `prem` para compatibilizar contagens por palavras e por EM

Um atributo posicional importante que foi inserido no corpus CDHAREM é o atributo `prem` (posição relativa na EM), que identifica o número de ordem de uma palavra dentro de uma EM. O atributo `prem` assume o valor "0" no caso de a palavra não pertencer a nenhuma EM.

Podemos usar este atributo também para identificar os casos em que *São Paulo* é a parte final de uma EM maior:

```
[word="São" & prem!="1" & prem!="0"] "Paulo"
```

Ou, pelo contrário, a parte inicial de uma EM maior:

```
"São" "Paulo" [prem="3"]
```

Assim como obter os casos de *Porto* que não fazem parte de uma EM.

```
[word="Porto" & prem="0"]
```

A maior utilidade deste atributo é permitir restringir as consultas de distribuição apenas às EM, e que devem ser feitas apenas sobre a primeira palavra de cada EM (ou seja, em que o valor de `prem` seja igual a 1), para que as outras palavras da EM não influenciem o resultado (senão, uma EM com cinco palavras contaria cinco vezes).

20.2.5 Atributos relativos ao texto

As etiquetas que delimitam documentos da CD (<DOC> e </DOC>) e os respectivos textos (<TEXTO> e </TEXTO>) foram convertidas no CDHAREM em atributos estruturais. à etiqueta <DOC> foi adicionada a informação constante das etiquetas <DOCID>, <GENERO> e <ORIGEM>, que não foram incluídas no corpus; à etiqueta <TEXTO> foi adicionado o tamanho do excerto, como se pode ver no exemplo abaixo.

```
<DOC docid=HAREM-871-07800 genero=Web origem=PT>
<TEXTO TAM=279>
```

Foram adicionados ainda outros três atributos posicionais com informação constante nas etiquetas removidas, e relativos ao documento propriamente dito:

- docid, a identificação do documento na colecção, no formato especificado no capítulo 19;
- genero, o tipo de texto, que pode ter um dos seguintes valores: Jornalístico, Web, CorreioElectrónico, Entrevista, Expositivo, Literário, Político, Técnico;
- origem, dado pelo código ISO do país de origem do texto: PT (Portugal), BR (Brasil), AO (Angola), MZ (Moçambique), CV (Cabo Verde), MO (Macau), IN (Índia) ou TL (Timor-Leste)[6].

Estes atributos posicionais, gerados a partir das etiquetas homónimas, podem ser usados, por exemplo, para identificar todas as pessoas assinaladas em texto jornalístico brasileiro:

```
<PESSOA> [origem="BR" & genero="Jorn.*"]* </PESSOA>
```

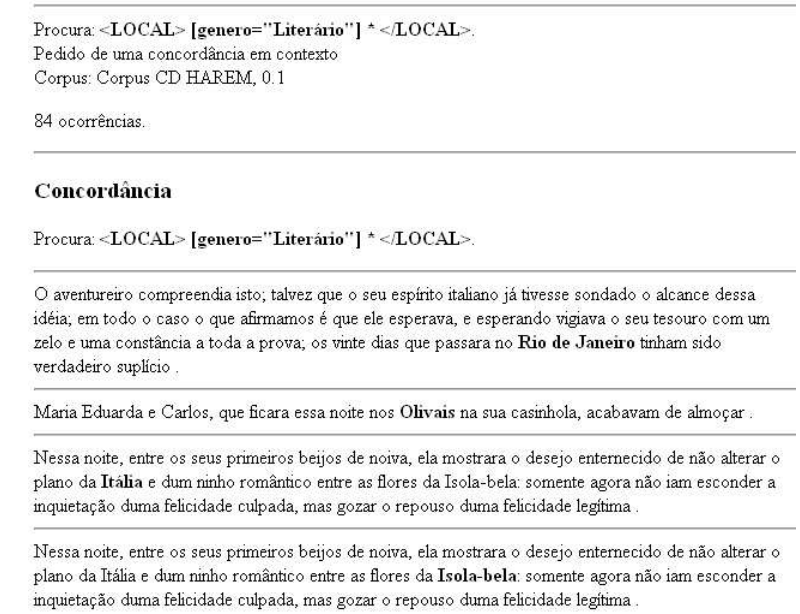
Escolhendo a distribuição das EM por categoria, podemos ver a distribuição das EM em texto técnico (note-se o uso de prem para que só uma palavra de cada EM seja contabilizada):

```
[genero="Técnico" & prem="1"]
```

Refinando ainda mais esta consulta, podemos seleccionar a distribuição por tipo apenas das EM da categoria COISA em texto técnico:

```
[genero="Técnico" & prem="1" & coisa!="0"]
```

[6] Embora existam textos de São Tomé e Príncipe (ST) e da Guiné-Bissau (GW) na colecção do HAREM, estes não aparecem nas colecções douradas.



Procura: <LOCAL> [genero="Literário"] * </LOCAL>.
Pedido de uma concordância em contexto
Corpus: Corpus CD HAREM, 0.1

84 ocorrências.

Concordância

Procura: <LOCAL> [genero="Literário"] * </LOCAL>.

O aventureiro compreendia isto; talvez que o seu espírito italiano já tivesse sondado o alcance dessa idéia; em todo o caso o que afirmamos é que ele esperava, e esperando vigiava o seu tesouro com um zelo e uma constância a toda a prova; os vinte dias que passara no **Rio de Janeiro** tinham sido verdadeiro suplício .

Maria Eduarda e Carlos, que ficara essa noite nos **Olivais** na sua casinhola, acabavam de almoçar .

Nessa noite, entre os seus primeiros beijos de noiva, ela mostrara o desejo enternecido de não alterar o plano da **Itália** e dum ninho romântico entre as flores da Isola-bela: somente agora não iam esconder a inquietação duma felicidade culpada, mas gozar o repouso duma felicidade legítima .

Nessa noite, entre os seus primeiros beijos de noiva, ela mostrara o desejo enternecido de não alterar o plano da Itália e dum ninho romântico entre as flores da **Isola-bela**: somente agora não iam esconder a inquietação duma felicidade culpada, mas gozar o repouso duma felicidade legítima .

Figura 20.2: Exemplo de concordância: locais referidos em texto literário (excerto)

20.2.6 Atributos relativos à classificação morfológica

A informação morfológica da CD do HAREM foi mantida no CDHAREM com a ajuda do atributo posicional `morf`. Desta forma, podemos por exemplo procurar todas as referências a pessoas do sexo feminino na CDHAREM:

```
<PESSOA> [tipo="INDIVIDUAL" & morf="F,S"]+ </PESSOA>
```

ou pedir a distribuição por género e número da categoria dos acontecimentos.

```
<ACONTECIMENTO> []
```

20.2.7 Atributos relativos à anotação sintáctica do AC/DC

Foram também adicionados atributos estruturais relativos aos parágrafos (`<p>`) e às frases (`<s>`). Podemos assim, por exemplo, pedir ao serviço AC/DC todas as frases contendo a palavra *Luanda*.

```
<s> []* "Luanda" []* </s>
```

Por fim, existe a informação gramatical acrescentada pelo analisador sintáctico PALAVRAS. Esta informação é gerada automaticamente e não foi, até agora, revista manualmente – para avaliações parciais do desempenho do PALAVRAS, veja-se Bick (2000), Santos e Gasperin (2002) ou Santos e Inácio (2006) – mas permite fazer consultas poderosas

desde que se tome esse facto em consideração. Um exemplo pode ser a distribuição das EM por função sintáctica:

```
[prem="1"]
```

ou das EM da categoria PESSOA como sujeito de um verbo de locução:

```
<PESSOA> [func="SUBJ"]* </PESSOA> [lema="dizer|afirmar|relatar"]
```

Pode-se também combinar numa consulta atributos de fontes diferentes, ou seja, atributos posicionais vindos do HAREM e da anotação gramatical automática, como o demonstra a seguinte procura de EM precedidas por um adjectivo:

```
[pos="ADJ" & prem="0"] [prem="1"]
```

20.3 Vagueza

Como várias vezes referido neste livro e noutras publicações (Santos et al., 2006), a codificação explícita da vagueza é um dos pontos fortes do HAREM.

20.3.1 Vagueza na classificação (categorias ou tipos com |)

Um total de 271 EM (2,9% do total das EM da CD) apresentavam anotações alternativas (66 entre tipos da mesma categoria, 202 entre duas categorias distintas e 3 entre três categorias), embora contendo exactamente as mesmas palavras. Nestes casos, as anotações foram mantidas e as EM foram assinaladas no CDHAREM com todas as suas categorias e tipos.

Casos como:

```
<PESSOA|ORGANIZACAO TIPO="GRUPOCARGO|SUB" MORF="F,S">
Convenção
</PESSOA|ORGANIZACAO>
```

foram, em termos de atributos estruturais, codificados como

```
<PESSOA TIPO="GRUPOCARGO" MORF="F,S">
<ORGANIZACAO TIPO="SUB" MORF="F,S">
Convenção
</ORGANIZACAO>
</PESSOA>
```

Assim sendo, apenas as dez categorias simples de EM estão codificadas directamente em atributos posicionais e estruturais. Para encontrar EM classificadas como pertencendo a múltiplas categorias, há várias maneiras possíveis de efectuar a consulta:

```
<PESSOA> <ORGANIZACAO> []
<ORGANIZACAO> <PESSOA> []
[pessoa!="0" & organizacao="0"]
[categoria="ORGANIZACAO-PESSOA"]
```

20.3.2 Vagueza na identificação: as etiquetas <ALT>

Uma vez que o formato usado para as etiquetas `<ALT>` leva à repetição dos textos das EM na CD, tivemos de proceder a algum processamento adicional de forma a codificar as anotações alternativas assinaladas com estas etiquetas.

Há um total de 122 etiquetas `<ALT>` na CD que foram identificadas na CDHAREM com o atributo posicional `alt`, contendo um valor diferente de 0.

De momento, codificámos a primeira alternativa, indicando o número de alternativas como valor do atributo estrutural `<ALT>`, bem como o valor de `alt` à categoria ou categorias das alternativas, iniciada por P M ou F (princípio, meio e fim). Quando a alternativa fosse nula (não pertencesse a EM), considerámos 0 como nome da categoria. Quando o princípio, meio e fim coincidissem, marcámos sempre primeiro o princípio, seguido de meio e só em último lugar do fim.

Seguem alguns exemplos ilustrativos:

```
... no jogo <ALT> <ACONTECIMENTO TIPO="EVENTO" MORF="M,S"> Académica-Benfica
</ACONTECIMENTO> | <PESSOA TIPO="GRUPOMEMBRO"> Académica </PESSOA> - <PESSOA
TIPO="GRUPOMEMBRO"> Benfica </PESSOA> </ALT>.

<ALT num=2>
<ACONTECIMENTO>
Académica PPESSOA
-         P0
Benfica   PPESSOA
</ACONTECIMENTO>
</ALT>

<ALT> Governo de <PESSOA TIPO=INDIVIDUAL>Cavaco Silva</PESSOA> | <ORGANIZACAO|
PESSOA TIPO=ADMINISTRACAO|GRUPOIND> Governo de Cavaco Silva</ALT>

<ALT num=2>
Governo   PORGANIZACAO
de        MORGANIZACAO
<PESSOA>
Cavaco    MORGANIZACAO
```

```
Silva     FORGANIZACAO
</PESSOA>
</ALT>

Um pouco de <ALT> HISTÓRIA | <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">
HISTÓRIA </ABSTRACCAO> </ALT>

<ALT num=2>
HISTÓRIA PABSTRACCAO
</ALT>
```

Em termos de procuras possibilitadas pelo AC/DC, além de podermos observar que sequências alternativas foram consideradas `<ALT>` nas CD:

```
<ALT> []+ </ALT>
```

podemos também localizar os casos em que a palavra *Governo* faz parte, na CD, de uma EM alternativa à assinalada no corpus:

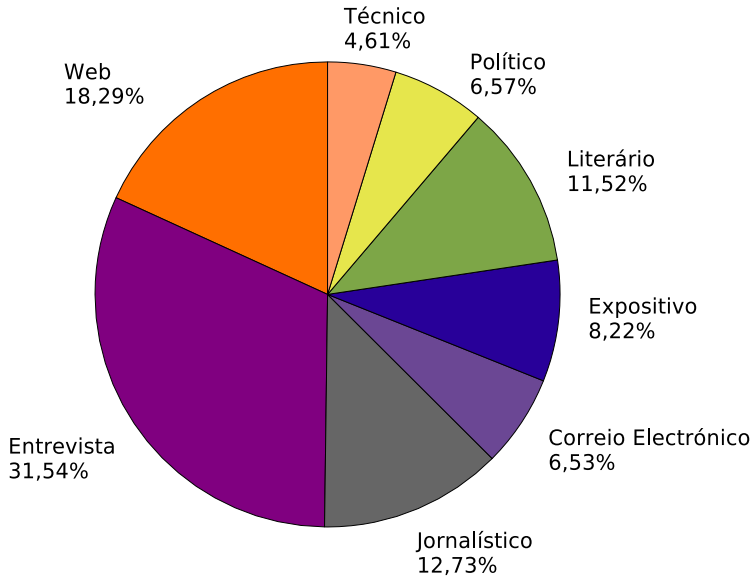
```
[word="Governo" & alt!="0"]
```

	Documento	Parágrafo	Frase
Média de número de entidades por	33,65	1,71	1,05
Mediana do número de entidades por	30	1	0
Número máximo de EM num	205	9	9
Número mínimo de EM num	2	0	0
Unidades textuais com 0 EM	0%	40,5%	50,3%
Unidades textuais com 1 EM	0%	25,4%	24,6%
Unidades textuais com 2 EM	1,9%	13,8%	12,4%
Unidades textuais com 3 EM	0%	7,0%	5,6%
Unidades textuais com 4 EM	2,3%	4,1%	3,1%
Unidades textuais com 5 a 9 EM	7,8%	6,9%	3,7%
Unidades textuais com 10 a 19 EM	24,0%	2,0%	0,3%
Unidades textuais com 20+EM	63,6%	0,4%	0,1%

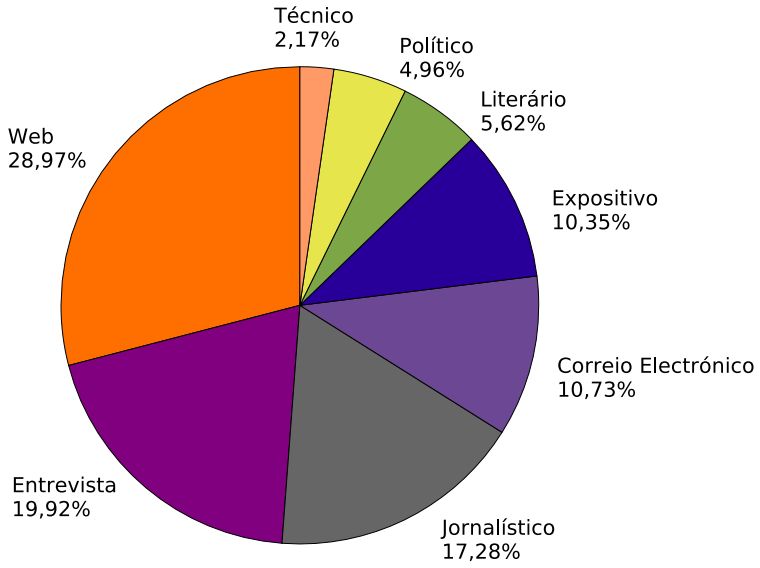
Tabela 20.2: Distribuição da quantidade de EM por unidades de texto.

20.4 Dados quantitativos

Segundo as normas de atomização do projecto AC/DC, o CDHAREM contém 154.863 unidades (133.569 das quais palavras, 86,3%), incluindo 8.976 EM que abrangem 17.206 unidades (16.821 das quais palavras, 97,2%).



(a) Em função do número de unidades.



(b) Em função do número de EM.

Figura 20.3: Distribuição por género dos termos existentes nas CD.

Quanto a EM, o CDHAREM apresenta um total de 8.967 EM (menos 463 que as CD originais, devido à nossa escolha relativa aos ALT), distribuídas por 8.184 frases (incluindo 990 fragmentos), agrupadas em 5.062 parágrafos e oriundas de 257 documentos distintos. Na Tabela 20.2, encontra-se uma distribuição quantitativa das EM por texto, por parágrafo e por frase.

Como mencionado acima, os documentos da CD foram classificados como pertencentes a oito géneros distintos de texto. A Figura 20.3(a) mostra a repartição dos textos da CD em função do número de unidades, enquanto que a Figura 20.3(b) mostra a repartição em função do número de EM, elucidando as diferenças em termos de densidade de EM em função do género literário: certos géneros são mais ricos (ou mais pobres) em EM do que outros.

Como se pode ver na Figura 20.4, as categorias de EM mais frequentes são `LOCAL` e `PESSOA`, que entre si cobrem quase metade das EM.

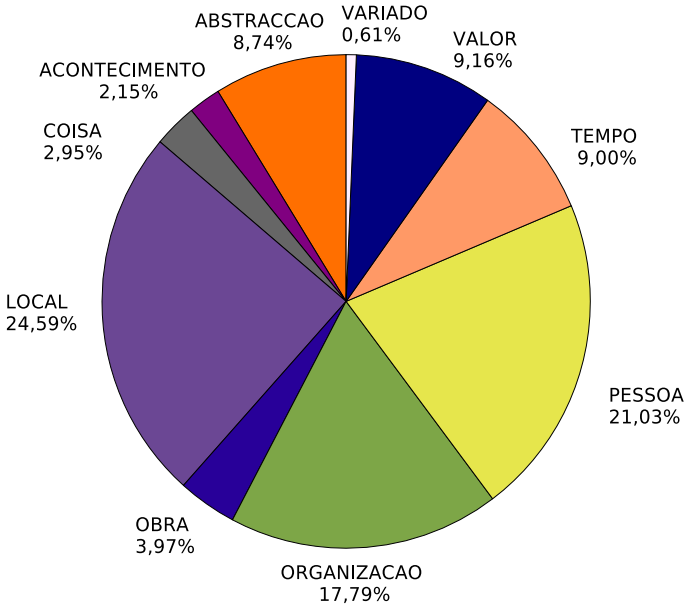


Figura 20.4: Distribuição das categorias semânticas de EM nas CD (sem peso).

As Figuras 20.5 e 20.6 mostram a relação entre as diferentes categorias de EM e os diversos géneros de texto.

Uma análise semelhante é feita em termos de variante, mas dado que a contribuição de textos em português não oriundos nem de Portugal nem do Brasil foi ínfima, considerámos apenas estas duas variantes na análise apresentada nas Tabelas 20.3 e 20.4 (correspondente assim a 251 textos, 150.041 unidades e 8.339 EM).

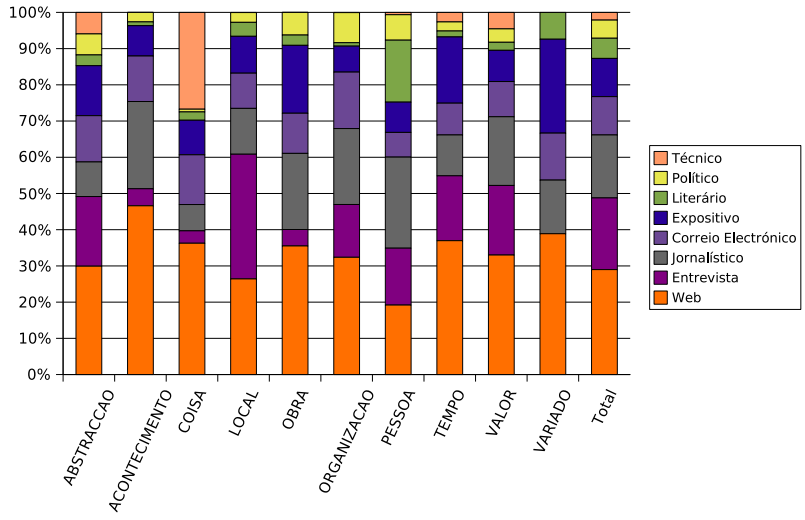


Figura 20.5: Distribuição das categorias semânticas de EM por género textual nas CD (sem peso).

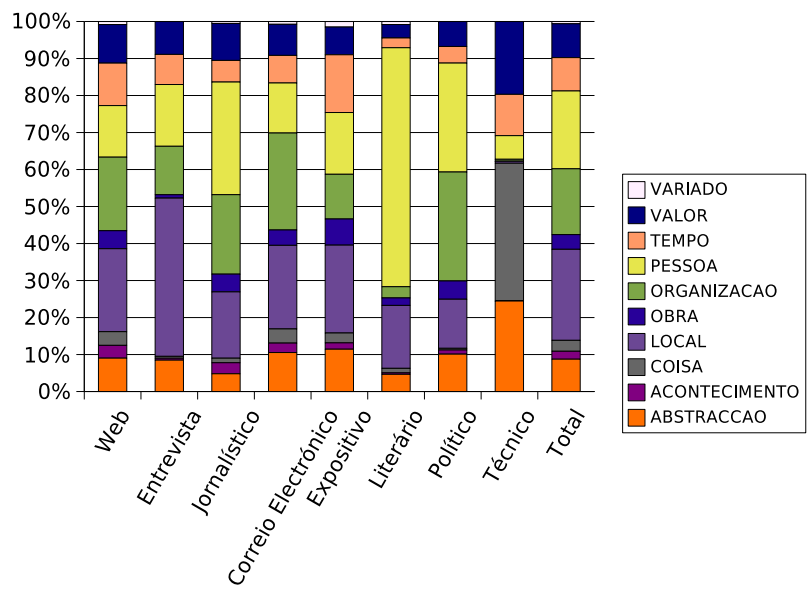


Figura 20.6: Distribuição do género textual das categorias semânticas de EM nas CD (sem peso).

A Tabela 20.5 apresenta a distribuição das categorias de entidades mencionadas na colecção dourada, repetindo em forma tabular a informação da figura 20.3.

A Tabela 20.6 apresenta o tamanho das entidades mencionadas em número de pala-

Categoria	Brasil	%	Portugal	%
ABSTRACCAO	364	49%	372	51%
ACONTECIMENTO	90	48%	96	52%
COISA	156	60%	104	40%
LOCAL	1.099	53%	987	47%
OBRA	147	46%	174	54%
ORGANIZACAO	785	51%	747	49%
PESSOA	920	51%	898	49%
TEMPO	349	45%	423	55%
VALOR	446	56%	354	44%
VARIADO	41	77%	12	23%
Total	4397	51%	4167	49%

Tabela 20.3: Distribuição das categorias semânticas por variante na CD (contando independentemente todas as classificações: EM pertencentes a múltiplas categorias são contadas para cada categoria).

Categoria	EM sem peso		EM com peso	
	Brasil	%	Portugal	%
ABSTRACCAO	7,6%	9,2%	7,9%	8,7%
ACONTECIMENTO	1,9%	1,8%	1,9%	2,3%
COISA	3,6%	2,7%	3,6%	2,5%
LOCAL	24,9%	23,3%	25,0%	23,9%
OBRA	3,2%	3,9%	3,2%	4,1%
ORGANIZACAO	17,4%	18,1%	17,7%	17,8%
PESSOA	21,5%	21,8%	21,4%	21,3%
TEMPO	8,2%	10,2%	8,0%	10,3%
VALOR	10,6%	8,5%	10,4%	8,7%
VARIADO	1,0%	0,3%	0,9%	0,3%

Tabela 20.4: Distribuição por variante das categorias semânticas na CD ; "EM sem peso" contam cada EM por cada categoria a que pertence; "EM com peso" contabilizam cada EM uma única vez atribuindo uma fracção a cada uma das suas categorias.

vras. Como se pode ver, mais de metade das EM contêm uma única palavra. A EM mais comprida (o título de uma palestra) contém 29 palavras.

Na Tabela 20.7 apresenta-se o tamanho médio das EM em número de palavras e a percentagem de EM simples (i.e., contendo uma única palavra), por categoria, e por cada variante. Todas as categorias de EM têm uma moda de 1 palavra, com excepção da categoria ACONTECIMENTO, onde a moda é de 3 palavras.

A Tabela 20.8 mostra a distribuição morfológica das EM em geral, e a Tabela 20.9 a mesma por categoria semântica. É interessante constatar a maioria esmagadora de entidades singulares.

A Tabela 20.10 mostra diferentes vertentes, permitindo uma primeira quantificação da

Categoria	CD 2005	CD 2006	Total	%
ABSTRACCAO	449	326	775	8,7%
ACONTECIMENTO	128	63	191	2,2%
COISA	82	180	262	3,0%
LOCAL	1.286	895	2.181	24,6%
OBRA	222	130	352	4,0%
ORGANIZACAO	956	622	1.578	17,8%
PESSOA	1.029	836	1.865	21,0%
TEMPO	434	364	798	9,0%
VALOR	484	328	812	9,2%
VARIADO	40	14	54	0,6%
Total	5.110	3.758	8.868	100,0%

Tabela 20.5: Distribuição das categorias de EM na CD.

Nº palavras	CD 2005	CD 2006	Total	%	Exemplo
1	2.769	2.052	4.821	54,3%	*Brasil*
2	1.049	888	1.937	21,8%	*São Paulo*
3	706	421	1.127	12,7%	*Universidade do Minho*
4	255	178	433	4,9%	*Rua 25 de Março*
5	165	94	259	2,9%	*25 de Abril de 1974*
6	48	36	84	0,9%	*Governador do Rio Grande do Norte*
7	46	22	68	0,8%	*26ª jornada da II Divisão de Honra*
8	20	12	32	0,4%	*Lei Antitruste (nº 8.884 / 94)*
9	19	12	31	0,3%	*Band of Gypsies: Live at the Fillmore East*
10+	38	43	81	0,9%	
Total	5.115	3.758	8.873	100,0%	

Tabela 20.6: Tamanho em número de palavras das EM.

Categoria	Texto completo			Textos brasileiros			Textos portugueses		
	Nº unid. médio por EM	EM simples (%)	EM de 6 ou mais palavras	Nº unid. médio por EM	EM simples (%)	EM de 6 ou mais palavras	Nº unid. médio por EM	EM simples (%)	EM de 6 ou mais palavras
ABSTRACCAO	2,2	51%	5%	2,7	46%	8%	1,3	56%	1%
ACONTECIMENTO	3,7	20%	16%	4,0	26%	20%	3,4	17%	11%
COISA	1,4	72%	<1%	1,5	71%	1%	1,3	73%	0%
LOCAL	1,7	68%	2%	1,8	61%	2%	1,5	74%	1%
OBRA	3,4	26%	13%	3,8	22%	17%	3,1	33%	11%
ORGANIZACAO	2,2	57%	6%	2,0	61%	4%	2,4	54%	9%
PESSOA	2,0	41%	2%	1,9	44%	1%	2,0	38%	2%
TEMPO	1,7	69%	<1%	1,7	67%	<1%	1,7	71%	<1%
VALOR	1,7	46%	<1%	1,8	43%	1%	1,7	50%	0%
VARIADO	1,9	69%	6%	1,9	71%	7%	2,2	58%	0%
Todas as categorias	2,0	54%	3%	2,0	53%	3%	1,9	55%	3%

Tabela 20.7: Informação sobre o tamanho das EM em número de palavras por categoria

	S	P	?	Total
M	3713	214	0	3.927
F	2565	83	0	2.648
?	543	1	94	638
Total	6.821	298	94	7.213
	Sem classificação			1.655

Tabela 20.8: Informação morfológica sobre as EM em geral

Categoria	M	F	?	S	P	?	s/class.
ABSTRACCAO	292 (38%)	418 54%)	54 (7%)	686 (89%)	59 (8%)	19 (2%)	11 (1%)
ACONTECIMENTO	102 (53%)	76 (40%)	13 (7%)	174 (91%)	16 (8%)	1 (<1%)	0 (0%)
COISA	183 (70%)	41 (16%)	33 (13%)	198 (75%)	38 (15%)	21 (8%)	5 (2%)
LOCAL	978 (45%)	750 (34%)	352 (16%)	2022 (93%)	46 (2%)	12 (1%)	101 (5%)
OBRA	188 (53%)	98 (26%)	58 (16%)	301 (85%)	20 (6%)	18 (5%)	13 (4%)
ORGANIZACAO	695 (44)	819 (52%)	58 (4%)	1524 (97%)	44 (3%)	4 (<1%)	6 (<1%)
PESSOA	1384 (74%)	431 (23%)	48 (3%)	1798 (96%)	61 (3%)	4 (<1%)	2 (<1%)
TEMPO	75 (9%)	13 (2%)	2 (<1%)	83 (10%)	7 (1%)	0 (0%)	708 (89%)
VARIADO	23 (43%)	5 (9%)	20 (37%)	30 (56%)	3 (6%)	15 (28%)	6 (11%)
Todas as categorias	44,5%	29,9%	7,1%	76,9%	3,6%	1,0%	18,5%

Tabela 20.9: Informação morfológica sobre as EM por categoria semântica.

dificuldade associada à tarefa descrita pela colecção dourada do HAREM, em particular:

- o número de palavras em maiúscula na colecção e quantas faziam parte de uma EM;
- o número de unidades pertencentes a EM que fazem parte de EM distintas (excluindo números e sinais de pontuação);
- o número de EM que tiveram diferentes classificações em contexto (dentre as EM que aparecem mais do que uma vez);
- o número de palavras (independentemente de estarem em maiúsculas ou minúsculas) que aparecem na colecção tanto fora como dentro de EM (excluindo números e sinais de pontuação);
- quantas palavras pertencentes a EM têm categorias distintas (excluindo números e sinais de pontuação).

20.5 Observações finais

A conversão da CD do HAREM para o corpus CDHAREM do AC/DC teve como principais objectivos produzir um recurso de maior qualidade e de mais fácil acesso, e disponibilizar uma ferramenta que permita preparar, com mais conhecimento empírico do problema,

Questão	Valores absolutos	%
Palavras em maiúscula	5.191 em 14.705	35,3%
Palavras distintas pertencentes a várias EM	1.655 em 5.453	30,6%
EM que ocorrem mais do que uma vez e com várias interpretações	360 em 4996	7,2%
Palavras distintas dentro de EM que também aparecem fora	1.337 em 4.455	30,0%
Palavras pertencentes a EM de categorias distintas	862 em 4.455	20,8%

Tabela 20.10: Dificuldade da tarefa reflectida na CD do HAREM

próximas edições do HAREM, permitindo medições mais rigorosas do(s) problema(s) que se pretende(m) resolver.

Ao converter a CD num formato mais acessível a linguistas, esperamos também provocar um maior interesse na comunidade linguística sobre o problema de reconhecimento de entidades mencionadas, assim como aproximar projectos como a Floresta Sintá(c)tica e o COMPARA de iniciativas como o HAREM ou as Morfolimpíadas.

Por outro lado, ao desenvolver um esquema que, de certa forma, combina as escolhas da tarefa partilhada do CoNLL (Sang, 2002; Sang e Meulder, 2003), baseadas em palavras – donde, atributos posicionais em formato CQP, e do MUC/HAREM, baseadas em atributos estruturais, mais uma vez usando terminologia do CQP – esperamos poder congregar uma comunidade alargada em redor de uma representação combinada do problema de REM, permitindo comparações finas e informadas entre diferentes abordagens de REM.

Agradecimentos

Estamos muito gratos ao Nuno Cardoso por nos ter facultado as figuras e tabelas constantes do presente capítulo, reproduzidas ou recalculadas da sua apresentação no PROPOR 2006 e na sua tese.

Este capítulo foi escrito no âmbito da Linguateca, integralmente financiado pela Fundação para a Ciência e Tecnologia através do projecto projecto POSC 339/1.3/C/NAC.

Apêndices

Apêndice A

Resultados do Primeiro HAREM

Os resultados completos dos dois eventos do Primeiro HAREM foram publicados (anonimizados) no sítio do HAREM, de onde se encontram ainda acessíveis, e, depois de termos obtido autorização de publicação dos resultados com o nome dos sistemas, pedida para análise mais completa destes na tese de Cardoso (2006a), outras apresentações foram calculadas e publicadas.

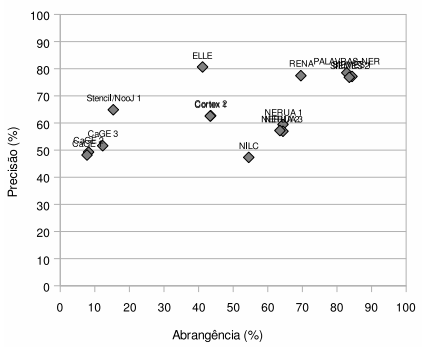
Aqui repetimos apenas os resultados principais das três tarefas (identificação, classificação morfológica e classificação semântica) para o cenário global, absoluto, quer para o primeiro evento (Figura A.1) quer para o Mini-HAREM (Figura A.4).

Para dar uma ideia do tipo de flexibilidade e de resultados obtidos, apresentamos também os melhores resultados por categoria, para o primeiro evento (Figura A.2) e para o Mini-HAREM (Figura A.5), em que apenas colocamos na figura os resultados do melhor sistema para cada categoria.

Da mesma forma, apresentamos os resultados dos vencedores, por género textual, também para os dois eventos (Figura A.3 e Figura A.6).

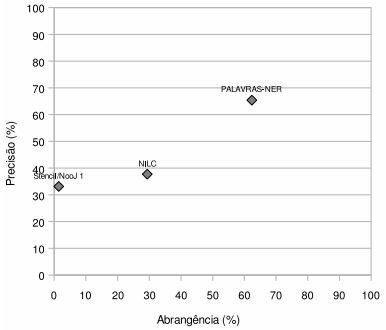
Em ambos os casos escolhemos os cenários totais absolutos para as classificações semântica e morfológica, e usámos a medida CSC para a primeira (ver secção 18.3) e a medida combinada para a segunda (ver secção 18.4).

Saída	Precisão (%)	Abrangência (%)	Medida F
PALAVRAS-NER	78,50	82,84	0,8061
SIEMÊS 1	77,15	84,35	0,8059
SIEMÊS 2	76,85	83,56	0,8006
RENA	77,43	69,57	0,7329
NERUA 1	59,45	64,39	0,6182
NERUA 3	56,95	64,39	0,6044
NERUA 2	57,21	63,51	0,6020
ELLE	80,64	41,15	0,5450
Cortex 2	62,68	43,51	0,5136
Cortex 1	62,55	43,36	0,5122
NILC	47,32	54,50	0,5066
Stencil/NooJ 1	64,87	15,33	0,2480
CaGE 3	51,61	12,28	0,1984
CaGE 2	49,28	8,19	0,1405
CaGE 1	48,18	7,74	0,1334



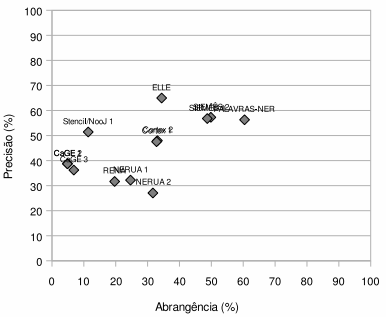
(a) Tarefa de identificação.

Saída	Precisão (%)	Abrangência (%)	Medida F
PALAVRAS-NER	78,50	82,84	0,8061
SIEMÊS 1	77,15	84,35	0,8059
SIEMÊS 2	76,85	83,56	0,8006



(b) Tarefa de classificação morfológica.

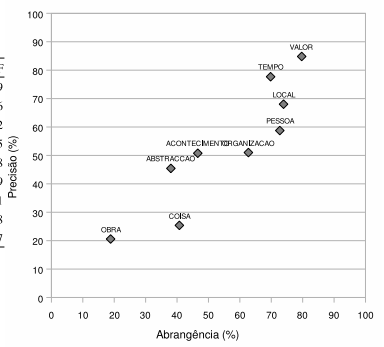
Saída	Precisão (%)	Abrangência (%)	Medida F
PALAVRAS-NER	56,3	60,42	0,5829
SIEMÊS 2	57,28	49,85	0,5330
SIEMÊS 1	56,79	48,73	0,5245
ELLE	64,98	34,41	0,4499
Cortex 2	47,95	33,09	0,3916
Cortex 1	47,54	32,81	0,3882
NERUA 2	27,06	31,66	0,2918
NERUA 1	32,2	24,64	0,2792
RENA	31,66	19,66	0,2426
Stencil/NooJ 1	51,42	11,37	0,1862
CaGE 3	36,22	6,85	0,1152
CaGE 1	38,78	4,87	0,0866
CaGE 2	38,7	4,85	0,0862



(c) Tarefa de classificação semântica.

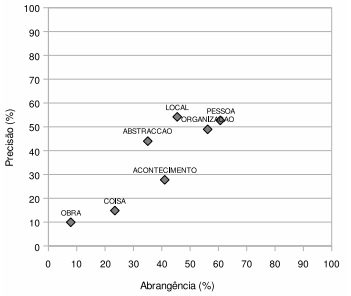
Figura A.1: Resultados globais para o primeiro evento do Primeiro HAREM.

Categoria	Saída	Precisão (%)	Abrangência (%)	Medida F
PESSOA	PALAVRAS-NER	58,75	72,72	0,6499
ORGANIZACAO	PALAVRAS-NER	51,01	62,72	0,5626
TEMPO	PALAVRAS-NER	77,68	69,79	0,7352
LOCAL	SIEMÊS 2	68,03	73,91	0,7085
OBRA	PALAVRAS-NER	20,58	18,85	0,1968
ACONTECIMENTO	SIEMÊS 2	50,76	46,61	0,4859
ABSTRACCAO	PALAVRAS-NER	45,43	38,04	0,4141
COISA	PALAVRAS-NER	25,38	40,74	0,3128
VALOR	PALAVRAS-NER	84,82	79,69	0,8217



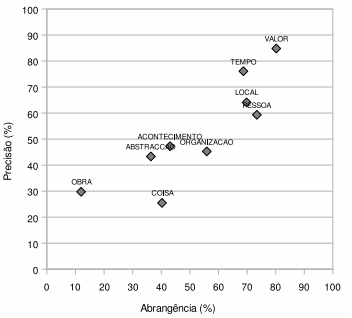
(a) Tarefa de identificação.

Categoria	Saída	Precisão (%)	Abrangência (%)	Medida F
PESSOA	PALAVRAS-NER	52,71	60,67	0,5641
ORGANIZACAO	PALAVRAS-NER	48,95	56,21	0,5233
LOCAL	PALAVRAS-NER	54,18	45,47	0,4944
OBRA	PALAVRAS-NER	9,94	7,90	0,0880
ACONTECIMENTO	PALAVRAS-NER	27,75	41,06	0,3311
ABSTRACCAO	PALAVRAS-NER	43,99	35,08	0,3903
COISA	PALAVRAS-NER	14,84	23,46	0,1818



(b) Tarefa de classificação morfológica.

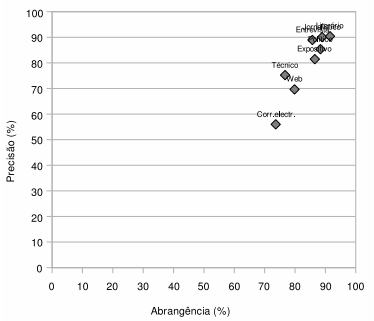
Categoria	Saída	Precisão (%)	Abrangência (%)	Medida F
PESSOA	PALAVRAS-NER	59,33	73,43	0,6563
ORGANIZACAO	PALAVRAS-NER	45,26	55,92	0,5003
TEMPO	PALAVRAS-NER	76,09	68,71	0,7221
LOCAL	SIEMÊS 2	64,09	69,83	0,6683
OBRA	SIEMÊS 2	29,75	11,96	0,1706
ACONTECIMENTO	SIEMÊS 2	47,26	43,05	0,4506
ABSTRACCAO	PALAVRAS-NER	43,33	36,37	0,3955
COISA	PALAVRAS-NER	25,47	40,25	0,3120
VALOR	PALAVRAS-NER	84,81	80,21	0,8245



(c) Tarefa de classificação semântica.

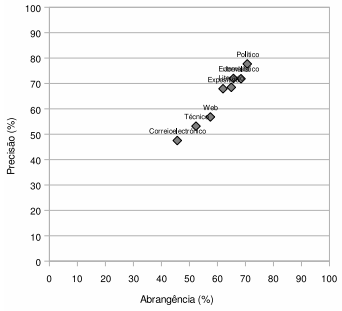
Figura A.2: Melhores resultados por categoria para o primeiro evento do Primeiro HAREM.

Género textual	Saída	Precisão (%)	Abrangência (%)	Medida F
Jornalístico	SIEMÊS 1	90,08	88,92	0,8950
Literário	PALAVRAS-NER	90,42	91,53	0,9097
Expositivo	PALAVRAS-NER	81,47	86,52	0,8392
Político	PALAVRAS-NER	85,35	88,38	0,8684
Web	SIEMÊS 1	69,65	79,85	0,7440
Entrevista	PALAVRAS-NER	88,96	85,72	0,8731
Corr.electr.	SIEMÊS 1	55,99	73,65	0,6362
Técnico	PALAVRAS-NER	75,23	76,72	0,7597



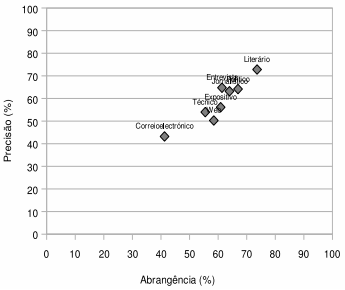
(a) Tarefa de identificação.

Género Textual	Saída	Precisão (%)	Abrangência (%)	Medida F
Jornalístico	PALAVRAS-NER	71,84	68,42	0,7008
Literário	PALAVRAS-NER	68,47	64,95	0,6667
Expositivo	PALAVRAS-NER	67,91	61,99	0,6481
Político	PALAVRAS-NER	77,69	70,65	0,7400
Web	PALAVRAS-NER	56,80	57,48	0,5714
Entrevista	PALAVRAS-NER	71,97	65,72	0,6870
Correioelectrónico	PALAVRAS-NER	47,51	45,66	0,4656
Técnico	PALAVRAS-NER	53,17	52,34	0,5276



(b) Tarefa de classificação morfológica.

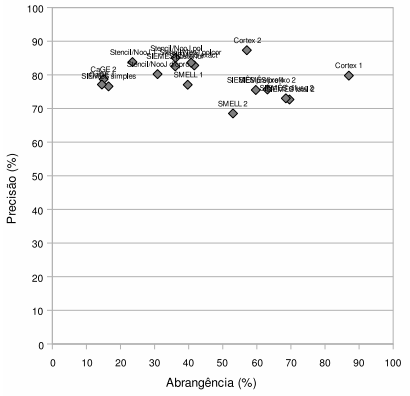
Género Textual	Saída	Precisão (%)	Abrangência (%)	Medida F
Jornalístico	PALAVRAS-NER	63,19	63,97	0,6358
Literário	PALAVRAS-NER	72,81	73,63	0,7322
Expositivo	PALAVRAS-NER	56,12	60,89	0,5841
Político	PALAVRAS-NER	64,17	66,96	0,6554
Web	PALAVRAS-NER	50,26	58,49	0,5406
Entrevista	SIEMÊS 2	64,75	61,35	0,6301
Correioelectrónico	SIEMÊS 2	43,2	41,25	0,4220
Técnico	PALAVRAS-NER	53,97	55,5	0,5472



(c) Tarefa de classificação semântica.

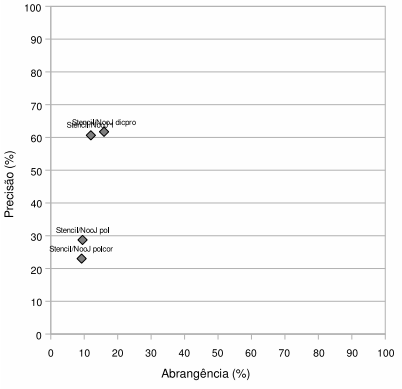
Figura A.3: Melhores resultados por género textual para o primeiro evento do Primeiro HAREM.

Saída	Precisão (%)	Abrangência (%)	Medida F
Cortex 1	79,77	87,00	0,8323
Cortex 3	79,77	87,00	0,8323
SIEMÊS total 2	72,68	69,58	0,7110
SIEMÊS total 1	72,68	69,58	0,7110
SIEMÊS difuso 2	73,03	68,52	0,7071
SIEMÊS difuso 1	73,03	68,52	0,7071
Cortex 2	87,33	57,07	0,6903
SIEMÊS prefixo 2	75,65	63,11	0,6881
SIEMÊS prefixo 4	75,50	59,70	0,6668
SMELL 2	68,53	53,00	0,5977
SIEMÊS exact	82,80	41,65	0,5542
Stencil/NooJ polcor	83,61	40,72	0,5476
SMELL 1	77,06	39,72	0,5242
Stencil/NooJ pol	85,06	36,17	0,5075
SIEMÊS posterior	82,52	36,07	0,5019
Stencil/NooJ dicpro	80,22	30,87	0,4458
Stencil/NooJ 1	83,80	23,51	0,3672
SIEMÊS simples	76,58	16,49	0,2713
CaGE 2	78,73	15,06	0,2529
CaGE 1	77,13	14,54	0,2447



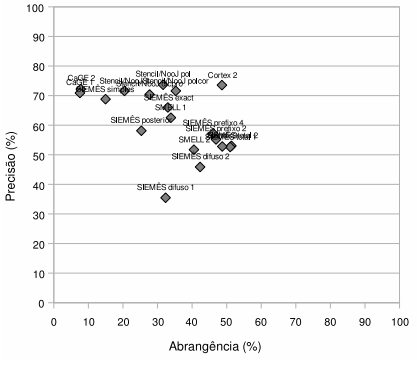
(a) Tarefa de identificação.

Saída	Precisão (%)	Abrangência (%)	Medida F
Stencil/NooJ dicpro	61,72	15,93	0,2532
Stencil/NooJ 1	60,64	11,98	0,2001
Stencil/NooJ pol	28,72	9,50	0,1428
Stencil/NooJ polcor	23,02	9,23	0,1317



(b) Tarefa de classificação morfológica.

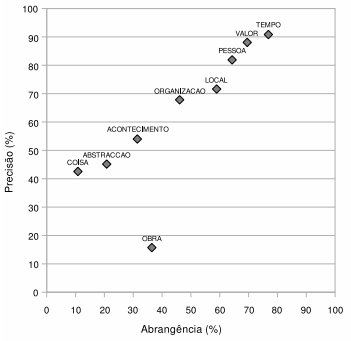
Saída	Precisão (%)	Abrangência (%)	Medida F
Cortex 2	73,55	48,63	0,5855
SIEMÊS total 2	53,02	51,38	0,5219
SIEMÊS total 1	52,63	51,01	0,5181
SIEMÊS prefixo 4	57,25	46,08	0,5106
SIEMÊS prefixo 2	55,17	46,92	0,5071
Cortex 3	52,80	48,73	0,5068
Stencil/NooJ polcor	71,59	35,29	0,4727
SMELL 2	51,72	40,53	0,4545
Stencil/NooJ pol	73,68	31,61	0,4424
SIEMÊS difuso 2	45,88	42,31	0,4403
SIEMÊS exact	66,00	32,99	0,4399
SMELL 1	62,53	33,87	0,4394
Stencil/NooJ dicpro	70,38	27,71	0,3977
SIEMÊS posterior	58,12	25,33	0,3528
SIEMÊS difuso 1	35,48	32,33	0,3383
Stencil/NooJ 1	71,58	20,44	0,3180
SIEMÊS simples	68,79	14,97	0,2458
CaGE 2	72,20	7,81	0,1409
CaGE 1	70,89	7,55	0,1365



(c) Tarefa de classificação semântica.

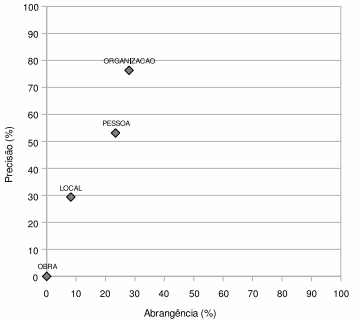
Figura A.4: Resultados globais para o Mini-HAREM.

Categoria	Saída	Precisão (%)	Abrangência (%)	Medida F
PESSOA	Cortex 3	81,90	64,29	0,7203
ORGANIZACAO	Cortex 3	67,82	46,09	0,5488
TEMPO	Cortex 3	90,83	76,86	0,8327
LOCAL	SMELL 2	71,62	58,92	0,6465
OBRA	SIEMÊS total 1	15,73	36,45	0,2197
ACONTECIMENTO	SMELL 1	54,00	31,42	0,3972
ABSTRACCAO	SIEMÊS total 2	45,16	20,80	0,2848
COISA	SIEMÊS prefixo 2	42,57	10,82	0,1725
VALOR	SIEMÊS posterior	88,05	69,53	0,7770



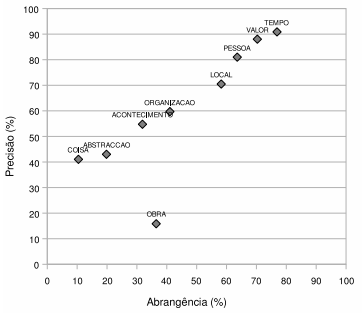
(a) Tarefa de identificação.

Categoria	Saída	Precisão (%)	Abrangência (%)	Medida F
PESSOA	Stencil/NooJ dicpro	53,12	23,39	0,3248
ORGANIZACAO	Stencil/NooJ dicpro	76,33	28,00	0,4097
LOCAL	Stencil/NooJ dicpro	29,38	8,21	0,1283
OBRA	Stencil/NooJ dicpro	0	0	0
COISA	Stencil/NooJ dicpro	0	0	0
ABSTRACCAO	Stencil/NooJ dicpro	0	0	0
COISA	Stencil/NooJ dicpro	0	0	0



(b) Tarefa de classificação morfológica.

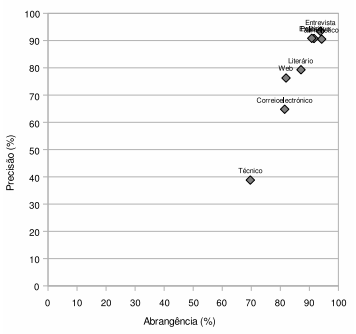
Categoria	Saída	Precisão (%)	Abrangência (%)	Medida F
PESSOA	Cortex 3	81,00	63,58	0,7124
ORGANIZACAO	Cortex 3	59,70	41,05	0,4865
TEMPO	Cortex 3	90,87	76,89	0,8330
LOCAL	SMELL 2	70,48	58,22	0,6376
OBRA	SIEMÊS total 1	15,85	36,46	0,2209
ACONTECIMENTO	SMELL 2	54,76	31,86	0,4028
ABSTRACCAO	SIEMÊS total 2	43,02	19,82	0,2713
COISA	SIEMÊS prefixo 2	41,05	10,43	0,1664
VALOR	SIEMÊS posterior	88,02	70,31	0,7817



(c) Tarefa de classificação semântica.

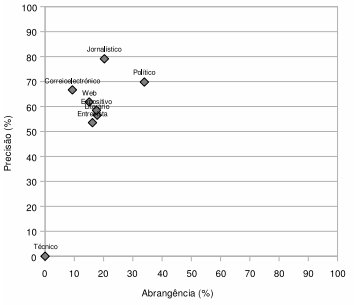
Figura A.5: Melhores resultados por categoria para o Mini-HAREM.

Género textual	Saída	Precisão (%)	Abrangência (%)	Medida F
Jornalístico	Cortex 3	90,52	94,24	0,9234
Literário	Cortex 3	79,29	87,12	0,8302
Expositivo	Cortex 3	90,76	91,59	0,9117
Político	Cortex 3	90,83	90,83	0,9080
Web	Cortex 3	76,26	81,97	0,7901
Entrevista	Cortex 3	93,40	93,79	0,9359
Correioelectrónico	Cortex 3	64,80	81,50	0,7220
Técnico	Cortex 3	38,81	69,67	0,4985



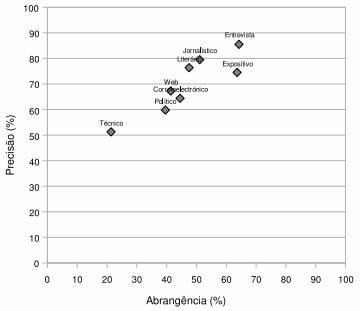
(a) Tarefa de identificação.

Género textual	Saída	Precisão (%)	Abrangência (%)	Medida F
Jornalístico	Stencil/NooJ dicpro	79,12	20,28	0,3229
Literário	Stencil/NooJ dicpro	56,52	17,81	0,2708
Expositivo	Stencil/NooJ dicpro	58,50	17,57	0,2702
Político	Stencil/NooJ dicpro	69,81	33,94	0,4568
Web	Stencil/NooJ dicpro	61,82	15,08	0,2424
Entrevista	Stencil/NooJ dicpro	53,53	16,20	0,2487
Correioelectrónico	Stencil/NooJ dicpro	66,67	9,34	0,1638
Técnico	Stencil/NooJ dicpro	0	0	0



(b) Tarefa de classificação morfológica.

Género textual	Saída	Precisão (%)	Abrangência (%)	Medida F
Jornalístico	Cortex 2	79,51	51,10	0,6221
Literário	Cortex 2	76,39	47,54	0,5861
Expositivo	SIEMÊS total 1	74,53	63,63	0,6865
Político	Cortex 2	59,81	39,57	0,4763
Web	Cortex 2	67,23	41,38	0,5123
Entrevista	Cortex 2	85,51	64,26	0,7338
Correioelectrónico	Cortex 2	64,39	44,50	0,5263
Técnico	SIEMÊS exact	51,27	21,35	0,3015



(c) Tarefa de classificação semântica.

Figura A.6: Melhores resultados por género textual para o Mini-HAREM.

Apêndice B

Lista de entidades classificadas no ensaio pré-HAREM

As tabelas B.1 e B.2 listam as entidades identificadas por pelo menos um anotador e a classificação que cada anotador lhes atribuiu, para o CETEMPúblico e o o CETENFolha, respectivamente. As categorias utilizadas nos quadros são mnemónicas das originalmente utilizadas pelos anotadores.

As tabelas B.3 e B.4 listam as entidades para as quais não houve acordo quanto à segmentação e respectiva classificação, para o CETEMPúblico e o o CETENFolha, respectivamente. As categorias utilizadas nos quadros são mnemónicas das originalmente utilizadas pelos anotadores.

A negrito encontra-se a maior sequência identificada; a itálico destaca-se as entidades que ficaram com outras encaixadas.

O fundo cinzento destaca nas tabelas B.1 e B.2 as entidades numéricas e temporais, e nas tabelas B.3 e B.4 as diferentes segmentações de uma mesma sequência do texto propostas pelos anotadores.

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[7 e Meio]		NPr+org	NPr+emp	NPr+lug	NPr+llaz	NPr+lug	NPr+org	NPr+lug	NPr+lug
[Algarve]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Albufeira]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Lisboa]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Londres]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Dublin]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Faro]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Portimão]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[dos 60]								Temp+data	
[dos 70]								Temp+data	
[Calypso]	NPr+lug	NPr+org	NPr+emp	NPr+lug	NPr+llaz	NPr+lug	NPr+org	NPr+lug	
[Locomia]	NPr+lug	NPr+org	NPr+emp	NPr+lug	NPr+llaz	NPr+lug	NPr+org	NPr+lug	
[2,5 milhões]					Num+mon	Num+mon		Num+din	Num+mon
[municípios]	NPr+org								
[Executivo]	NPr+org	NPr+out	NPr+org		NPr+inst		NPr+org	NPr+org	NPr+org
[câmaras]	NPr+org								
[autarquias]	NPr+org								
[GAT]	NPr+out	NPr+org	NPr+org	NPr+org	NPr+inst	NPr+org	NPr+org	NPr+org	NPr+org
[GAT]	NPr+out	NPr+org	NPr+org	NPr+org	NPr+inst	NPr+org	NPr+org	NPr+org	NPr+org
[Castro Verde]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+pess
[800 Km2]						Num+medida			
[Logitech]	NPr+org	NPr+org	NPr+emp	NPr+org	NPr+emp	NPr+lug	NPr+org	NPr+org	NPr+org
[Basileia]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Suíça]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Audioman]	NPr+out	NPr+prod	NPr+prod	NPr+pess	NPr+marProd	NPr+lug	NPr+out	NPr+eqpmt	NPr+eqpmt
[Steve d'Averio]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[director de marketing]	NPr+pess								
[Europa]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Logitech]	NPr+org	NPr+org	NPr+emp	NPr+org	NPr+emp	NPr+lug	NPr+org	NPr+org	NPr+org
[Audioman]		NPr+prod	NPr+prod	NPr+pess	NPr+marProd	NPr+obj	NPr+out	NPr+eqpmt	NPr+eqpmt
[Suíça]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[sete meses]						Temp+data		Temp+data	
[290 francos suíços]					Num+mon	Num+mon		Num+din	Num+mon
[28 contos]					Num+mon			Num+din	Num+mon
[Junqueiro]	NPr+pess		NPr+pess	NPr+org	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[João Cravinho]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Jorge Sampaio]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Estado]			NPr+org			NPr+org	NPr+org	NPr+org	NPr+org

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[Moçambique]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Junqueiro]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[IGAT]	NPr+out	NPr+org	NPr+org??	NPr+org	NPr+inst	NPr+org	NPr+org	NPr+org	NPr+org
[um mês]								Temp+data	
[João Pedro Henriques]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Museu do Ar]	NPr+org	NPr+org	NPr+org	NPr+lug	NPr+lcul	NPr+lug	NPr+org	NPr+org	NPr+org
[Portugal]	NPr+lug	NPr+lug	NPr+tema	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Japão]	NPr+lug	NPr+lug	NPr+tema	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Brasil]	NPr+lug	NPr+lug	NPr+tema	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[África]	NPr+lug	NPr+lug	NPr+tema	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Far-West]	NPr+lug	NPr+out	NPr+tema	NPr+lug	NPr+lug1	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Portugal]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Barcelona]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Port Aventura]	NPr+out	NPr+out	NPr+org	NPr+lug	NPr+lcul	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Seis milhões]					Num+mon	Num+pess		Num+din	
[dez anos]						Temp+data		Temp+data	
[Presidente da República]	NPr+pess		NPr+cg	NPr+pess	NPr+carFun	NPr+cgPub	NPr+pess	NPr+cg	
[seis milhões de contos]					Num+mon	Num+mon		Num+din	Num+mon
[Força Aérea]	NPr+org	NPr+org	NPr+org	NPr+org	NPr+inst	NPr+org	NPr+org	NPr+org	NPr+org
[Cameron Hall]	NPr+out	NPr+pess	NPr+??	NPr+pess	NPr+emp	NPr+lug	NPr+org	NPr+org	NPr+pess
[Rendimento Mínimo Garantido]	NPr+out	NPr+out	NPr+tit	NPr+out	NPr+prodMon	NPr+org	NPr+out	NPr+progGov	NPr+org
[RMG]	NPr+out	NPr+out	NPr+tit	NPr+out	NPr+prodMon	NPr+org	NPr+out	NPr+progGov	
[7.777 famílias]						Num+pess			
[26.668 pessoas]						Num+pess			
[PÚBLICO]	NPr+org	NPr+org	NPr+MCS	NPr+org	NPr+lcul		NPr+org	NPr+jorn	NPr+org
[RMG]	NPr+out	NPr+out	NPr+tit	NPr+out	NPr+prodMon	NPr+org	NPr+out	NPr+progGov	NPr+org
[36 por cento]					Num+perc	Num+perc		Num+perc	
[PÚBLICO]	NPr+org	NPr+org	NPr+MCS	NPr+org	NPr+lcul		NPr+org	NPr+jorn	NPr+org
[Paulo Pedroso]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[RMG]	NPr+out	NPr+out	NPr+tit	NPr+out	NPr+prodMon	NPr+lug	NPr+out	NPr+progGov	NPr+org
[Adriano Pimpão]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Pimpão]	NPr+pess	NPr+pess	NPr+pess		NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Estaline]	NPr+pess	NPr+pess	NPr+pess	NPr+out	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Pacto Molotov-Ribbentrop]	NPr+out	NPr+out	NPr+acont	NPr+out	NPr+doc	NPr+doc	NPr+out	NPr+doc	NPr+doc
[Ieltsin]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[cem mil pessoas]						Num+pess			
[Rússia]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug[	NPr+lug	NPr+lug	NPr+lug
[Rússia]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[Ocidente]	NPr+lug		NPr+org	NPr+lug	NPr+lug1	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[O Emigrante]	NPr+out	NPr+obra	NPr+tit	NPr+out	NPr+oCine	NPr+cultl	NPr+obra	NPr+tit	NPr+filme
[José]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Ram]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess		NPr+pess	NPr+pess	NPr+pess
[3000 anos]						Temp+data		Temp+data	Temp+data
[Egipto]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Chahine]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[bastonário dos advogados]	NPr+pess								
[Ahmed al-Khawaga]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[José]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Corão]	NPr+out	NPr+obra	NPr+tit	NPr+out	NPr+oLit	NPr+cult	NPr+obra	NPr+livro	NPr+livro
[Egipto]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[França]	NPr+lug	NPr+pess	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[The Music of Chance]	NPr+out	NPr+obra	NPr+tit	NPr+out	NPr+oLit	NPr+cult	NPr+obra	NPr+tit	NPr+tit
[Paul Auster]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Paul Auster]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Nashe]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Pozzi]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Faber and Faber]	NPr+pess	NPr+org	NPr+tit	NPr+org			NPr+org	NPr+tit	

Tabela B.1: Lista das entidades em comum identificadas por pelo menos um dos anotadores no CETEMPúblico.

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[PT]		NPr+org		NPr+org	NPr+ptdPol	NPr+org	NPr+org	NPr+org	NPr+org
[Gilberto Dimenstein]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[hoje]								Temp+data	
[77%]					Num+perc	Num+perc		Num+perc	Num+perc
[PT]	NPr+out	NPr+org	NPr+org	NPr+org	NPr+ptdPol	NPr+org	NPr+org	NPr+org	NPr+org
[tempos na ditadura]								Temp+data	
[PT]	NPr+out	NPr+org	NPr+org	NPr+org	NPr+ptdPol	NPr+org	NPr+org	NPr+org	NPr+org
[agora]								Temp+data	
[Lula]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[PT]	NPr+out	NPr+org	NPr+org	NPr+org	NPr+ptdPol	NPr+org	NPr+org	NPr+org	NPr+org
[Confissões]	NPr+out	NPr+obra			NPr+progTV	NPr+cult	NPr+obra	NPr+tit	NPr+prog
[Portugal]	NPr+lug	NPr+lug	NPr+país	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[dia 13]					Temp+data	Temp+data		Temp+data	Temp+data
[Confissões de Adolescente]	NPr+out	NPr+obra	NPr+tit	NPr+out	NPr+progTV	NPr+cult	NPr+obra	NPr+tit	NPr+prog
[Cultura]	NPr+org	NPr+org	NPr+MCS	NPr+org	NPr+lcul	NPr+lug	NPr+org	NPr+org	NPr+emiss
[TF1]		NPr+org	NPr+MCS	NPr+org	NPr+lcul	NPr+lug	NPr+org	NPr+org	NPr+emiss
[Manchete]		NPr+org	NPr+MCS		NPr+lcul	NPr+lug	NPr+org	NPr+org	NPr+emiss
[Câmera Manchete]	NPr+out	NPr+obra	NPr+tit	NPr+out	NPr+progTV	NPr+cult	NPr+org	NPr+tit	NPr+prog
[quarta-feira]					Temp+data	Temp+data		Temp+data	
[22h30]					Temp+hora	Temp+hora		Temp+hora	Temp+hora
[Rede Manchete]	NPr+org	NPr+org	NPr+MCS	NPr+org	NPr+lcul	NPr+lug	NPr+org	NPr+org	NPr+emiss
[Ronaldo Rosas]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Sônia Pompeu]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Ewaldo Ruy]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Primeiro Mundo]	NPr+out		NPr+reg	NPr+lug		NPr+lug	NPr+out	NPr+lug	NPr+lug
[Maurício]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Maurício]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Carlão]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Paulão]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[hoje]								Temp+data	NPr+pess
[Giovane]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	
[anteontem à noite]								Temp+data	
[Free shops]							NPr+org		NPr+lug
[LX 810]	NPr+out	NPr+marca	NPr+mod	NPr+pess	NPr+marProd	NPr+obj	NPr+out		
[Epson]	NPr+org	NPr+org	NPr+marca	NPr+org	NPr+emp	NPr+org	NPr+org	NPr+org	NPr+marca
[Miami]		NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[FSE]	NPr+out	NPr+out	NPr+??	NPr+org	NPr+prodMon	NPr+org	NPr+out	NPr+fundo	NPr+org
[Fernando Henrique Cardoso]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[FSE]		NPr+out	NPr+??	NPr+org	NPr+prodMon	NPr+org	NPr+out	NPr+fundo	NPr+org
[Sérgio Danese]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+org	NPr+pess	NPr+pess
[ontem]								Temp+data	
[Rubens Ricúpero]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[TSE]		NPr+out	NPr+??	NPr+lug	NPr+inst	NPr+org	NPr+out	NPr+org	NPr+org
[R$ 334,9 milhões]					Num+mon	Num+mon		Num+din	Num+din
[Congresso]	NPr+org	NPr+out	NPr+orgn		NPr+inst	NPr+lug	NPr+org	NPr+org	NPr+org
[um dia]						Temp+data			
[TSE]	NPr+out	NPr+out	NPr+??	NPr+lug	NPr+inst		NPr+out	NPr+org	NPr+org
[um dia]								Temp+data	
[CDI]	NPr+out	NPr+out	NPr+??	NPr+out	NPr+doc	NPr+org	NPr+out	NPr+org	NPr+org
[Telê]		NPr+org		NPr+org	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[São Paulo]	NPr+lug	NPr+org	NPr+org	NPr+lug	NPr+gpDesp	NPr+lug	NPr+org	NPr+equi	NPr+equi
[Folha]		NPr+org			NPr+lcul	NPr+cult	NPr+org	NPr+jorn	NPr+jorn
[Telê]		NPr+org		NPr+org	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[dois árbitros]						Num+pess			
[70]						Num+pess			
[CPI]	NPr+out	NPr+out	NPr+??		NPr+gpTrab	NPr+org	NPr+org	NPr+org	NPr+org
[Fifa]	NPr+org	NPr+org	NPr+orgn	NPr+org	NPr+assoc	NPr+org	NPr+org	NPr+org	NPr+org
[CBF]	NPr+org	NPr+org	NPr+orgn	NPr+org	NPr+assoc	NPr+org	NPr+org	NPr+org	NPr+org
[ontem]								Temp+data	
[Benedito Vieira Pereira]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[49]						Num+idade			
[C]			NPr+mod						
[hoje]								Temp+data	
[Prandi]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	
[Charade]	NPr+out	NPr+marca	NPr+mod	NPr+out	NPr+marProd	NPr+obj	NPr+auto	NPr+auto	NPr+carro
[Suzuki Swift]	NPr+out	NPr+marca	NPr+mod	NPr+out	NPr+marProd	NPr+obj	NPr+auto	NPr+auto	NPr+carro
[Twingo]	NPr+out	NPr+marca	NPr+mod	NPr+out	NPr+marProd	NPr+obj	NPr+auto	NPr+auto	NPr+carro
[Renault]	NPr+org	NPr+org	NPr+marca	NPr+org	NPr+emp	NPr+lug	NPr+org	NPr+org	NPr+emp
[Caparelli]	NPr+pess	NPr+pess	NPr+mod	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Herbert Berger]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[diretor-superintendente da empresa]	NPr+pess								
[Charade]	NPr+out	NPr+marca	NPr+mod	NPr+out	NPr+marProd	NPr+obj	NPr+auto	NPr+auto	NPr+carro
[Applause]	NPr+out	NPr+marca	NPr+mod	NPr+out	NPr+marProd	NPr+obj	NPr+auto	NPr+auto	NPr+carro
[Daihatsu]	NPr+org	NPr+org	NPr+marca	NPr+org	NPr+emp	NPr+lug	NPr+org	NPr+org	NPr+emp
[Corinthians]	NPr+org	NPr+org	NPr+equi	NPr+org	NPr+gpDesp	NPr+lug	NPr+org	NPr+equi	NPr+equi

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[dia 17]					Temp+data	Temp+data		Temp+data	Temp+data
[CBF]	NPr+out	NPr+org	NPr+org	NPr+org	NPr+assoc	NPr+org	NPr+org	NPr+org	NPr+org
[Vila]		NPr+out	NPr+??		NPr+gpDesp	NPr+lug	NPr+org	NPr+lug	NPr+lug
[Neto]	NPr+pess	NPr+pess	NPr+pess		NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Telecine]		NPr+out	NPr+cine	NPr+org	NPr+lcul	NPr+cult	NPr+org	NPr+canal	NPr+canal
[20h30]					Temp+hora			Temp+hora	Temp+hora
[Schwarzenegger]	NPr+pess	NPr+pess		NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[HBO]	NPr+out	NPr+org	NPr+??	NPr+lug	NPr+lcul	NPr+lug	NPr+org	NPr+canal	NPr+canal
[HBO]		NPr+org	NPr+??	NPr+lug	NPr+lcul	NPr+lug	NPr+org	NPr+canal	NPr+canal
[Exterminador do Futuro 2 -- O Julgamento Final]	NPr+out	NPr+obra	NPr+tit	NPr+out	NPr+oCine	NPr+cult	NPr+obra	NPr+tit	NPr+filme
[Schwarzenegger]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Alexandre Cardoso]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[21]						Num+idade			
[Topeira]	NPr+out	NPr+pess	NPr+alc	NPr+pess	NPr+pess	NPr+lug	NPr+pess	NPr+pess	NPr+apel
[20 anos]						Temp+data		Temp+data	Temp+data
[Souza]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	
[30 de julho de 93]					Temp+data	Temp+data		Temp+data	Temp+data
[Fifa]	NPr+org	NPr+org	NPr+org	NPr+org	NPr+assoc		NPr+org	NPr+org	NPr+org
[SÍLVIO LANCELLOTTI]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess		NPr+org	NPr+pess	NPr+pess
[Fifa]	NPr+org	NPr+org	NPr+org	NPr+org	NPr+assoc	NPr+org	NPr+org	NPr+org	NPr+org
[seis meses depois]						Temp+data		Temp+data	
[Copa]	NPr+out	NPr+even	NPr+acont		NPr+even		NPr+org		NPr+org
[João Havelange]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Havelange]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[África]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Ásia]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Europa]	NPr+lug	NPr+lug	NPr+org	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Havelange]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Antonio Matarrese]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Ambulim]	NPr+out	NPr+org	NPr+org	NPr+lug	NPr+inst	NPr+lug	NPr+org	NPr+org	NPr+lug
[Nova York]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Brasil]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Áustria]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Estados Unidos]	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Harrison Pope]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Alfredo Volpi]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Volpi]	NPr+pess	NPr+pess		NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[Fukushima]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Pérsio]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Raimo]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Douchez]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Volpi]		NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Cícero]	NPr+pess	NPr+pess		NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Quércia]	NPr+pess	NPr+pess	NPr+??	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Quércia]	NPr+pess	NPr+pess	NPr+??	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Romário]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[Romário]	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess	NPr+pess

Tabela B.2: Lista das entidades em comum identificadas por pelo menos um dos anotadores no CETENFolha.

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[Ministério do Planeamento e Administração do Território]	NPr+org	NPr+org		NPr+lug+INST	NPr+inst	NPr+org	NPr+org	NPr+org	NPr+org
[Ministério do Planeamento]			NPr+org						
[Administração do Território]			NPr+org						
[membro do Governo]	NPr+pess								
[Governo]		NPr+out	NPr+org		NPr+inst		NPr+org	NPr+org	NPr+org
[secretário de Estado]	NPr+pess *[secretário de [Estado]]*				NPr+cgFun	NPr+pess			
[Estado]	NPr+org		NPr+org				NPr+org	NPr+org	NPr+org
[organismos do Estado]	NPr+org								
[Estado]		NPr+out	NPr+org		NPr+inst		NPr+org	NPr+org	NPr+org
[Jardim Zoológico de Lisboa]	NPr+org *[Jardim Zoológico de [Lisboa]]*	NPr+org	NPr+org	NPr+lug+INST	NPr+llaz				NPr+lug
[Jardim Zoológico]							NPr+org	NPr+lug	
[Lisboa]	NPr+lug					NPr+lug	NPr+lug	NPr+lug	
[major Carlos Barbosa]	NPr+pess								
[Carlos Barbosa]		NPr+pess	NPr+pess	NPr+pess+hum	NPr+antrop	NPr+pess	NPr+pess	NPr+pess	NPr+pess
[relações públicas da Força Aérea]	NPr+org *[relações públicas da [Força Aérea]]*								
[Força Aérea]	NPr+org	NPr+org	NPr+org	NPr+org+org	NPr+inst	NPr+org	NPr+org	NPr+org	NPr+org
[pouco mais de um mês]								Temp+data	
[um mês]						Temp+data			
[Comissão Nacional do RMG]	NPr+org *[Comissão Nacional do [RMG]]*	NPr+org	NPr+org		NPr+gpTrab				
[Comissão Nacional]				NPr+org+org		NPr+org	NPr+org	NPr+org	NPr+org
[RMG]	NPr+out			NPr+out+semtit		NPr+org	NPr+out	NPr+progGov	NPr+org
[30 de Março]			Temp+data		Temp+data	Temp+data		Temp+data	Temp+data
[Março]							NPr+data		
[dia 1 de Julho]					Temp+data				Temp+data
[1 de Julho]			Temp+data			Temp+data		Temp+data	
[Julho]							NPr+data		

Entidades	AS	CM	DS	EB	Lab	LO	Prib	RM	VM
[presidente da Comissão Nacional do RMG]	NPr+pess *[presidente da [Comissão Nacional do [RMG]]]*								
[Comissão Nacional do RMG]	NPr+org *[Comissão Nacional do [RMG]]*	NPr+org	NPr+org	NPr+org+org	NPr+gpTrab	NPr+org	NPr+org	NPr+org	NPr+org
[RMG]	NPr+out			NPr+out+semtit		NPr+org	NPr+out		
[pouco mais de um mês]								Temp+data	
[um mês]						Temp+data			
[secretário de Estado do Desenvolvimento Regional]	NPr+pess				NPr+cgFun	NPr+cgPub	NPr+pess		
[Estado do Desenvolvimento Regional]			NPr+cg+pessoa					NPr+org	NPr+org
[Desenvolvimento Regional]				NPr+out+even					
[Museu da Segunda Guerra Mundial]	NPr+org *[Museu da [Segunda Guerra Mundial]]*	NPr+org	NPr+org+museu	NPr+lug+INST	NPr+lcul	NPr+lug	NPr+org	NPr+lug	NPr+org
[Segunda Guerra Mundial]	NPr+out								
[ministro dos Negócios Estrangeiros da Alemanha]	NPr+pess *[ministro dos Negócios Estrangeiros da [Alemanha]]*		NPr+cg+pessoa						
[ministro dos Negócios Estrangeiros]					NPr+cgFun	NPr+cgPub	NPr+pess		
[Negócios Estrangeiros da Alemanha]				NPr+out+genre					
[Negócios Estrangeiros]								NPr+org	
[Alemanha]	NPr+lug	NPr+lug			NPr+top	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Presidente russo]	NPr+pess								
[Presidente]			NPr+cg+pessoa		NPr+cgFun	NPr+cgPub	NPr+pess		
[filho de Jacob]	Pr+pess *[filho de [Jacob]]*								
[Jacob]	NPr+pess	NPr+pess	NPr+pess	NPr+pess+hum	NPr+antrop	NPr+pess	NPr+pess	NPr+pess	NPr+pess

Tabela B.3: Lista das entidades em que não houve acordo quanto à sua segmentação no CETEMPúblico.

Entidades	[AS]	[CM]	DS	EB	Lab	LO	Prib	RM	VM
[BRASÍLIA Pesquisa Datafolha]	NPr+out *[[BRASÍLIA] Pesquisa Datafolha]*								
[BRASÍLIA]	NPr+lug	NPr+lug	NPr+lug	NPr+lug+civ	NPr+top	NPr+lug	NPr+lug	NPr+lug	NPr+lug
[Pesquisa Datafolha]			NPr+estudo						NPr+org
[Datafolha]		NPr+org		NPr+out+semtit	NPr+lcul	NPr+cult	NPr+org	NPr+org	
[Governo Fernando Henrique Cardoso]	NPr+org *[Governo [Fernando Henrique Cardoso]]*		NPr+org+gov	NPr+org+org				NPr+org	NPr+org
[Governo]		NPr+out			NPr+inst		NPr+org		
[Fernando Henrique Cardoso]	NPr+pess	NPr+pess			NPr+antrop	NPr+pess	NPr+pess		
[TVI de Portugal]	NPr+org *[TVI de [Portugal]]*		NPr+MCS			NPr+lug			
[TVI]		NPr+org		NPr+org+org	NPr+lcul		NPr+org	NPr+org	NPr+emiss
[Portugal]	NPr+lug	NPr+lug		NPr+lug+civ	NPr+top		NPr+lug	NPr+lug	NPr+lug
[mais um dia]								Temp+data	
[um dia]						Temp+data			
[US$ 178]					Num+mon	Num+mon		Num+din	Num+din
[US$]			NPr+moe						
[US$ 422]					Num+mon	Num+mon		Num+din	Num+din
[US$]			NPr+moe						
[ministro da Fazenda]	NPr+pess *[ministro da [Fazenda]]*				NPr+carFun	NPr+cgPub	NPr+pess		
[Fazenda]	NPr+org		NPr+cg					NPr+org	NPr+org

[assessor de imprensa do Ministério da Fazenda]	NPr+pess *[assessor de imprensa do [Ministério da Fazenda]]*								
[Ministério da Fazenda]	NPr+org	NPr+org	NPr+orgn	NPr+lug+inst	NPr+inst	NPr+lug	NPr+org	NPr+org	NPr+org
[ministro da Fazenda]	NPr+pess *[ministro da [Fazenda]]*				NPr+carFun	NPr+cgPub			
[Fazenda]	NPr+org		NPr+??				NPr+pess	NPr+org	NPr+org
[R$ 452,7 milhões]					Num+mon			Num+din	Num+din

Entidades	[AS]	[CM]	DS	EB	Lab	LO	Prib	RM	VM
[R$]			NPr+moe						
[presidente da Cooper]	NPr+pess *[presidente da [Cooper]]*								
[Cooper]	NPr+org	NPr+org	NPr+emp	NPr+org+org	NPr+emp	NPr+org	NPr+org	NPr+org	NPr+org
[Honda Civic]	NPr+out *[[Honda] Civic]*	NPr+marca	NPr+mod	NPr+out+veic	NPr+marProd	NPr+obj	NPr+auto	NPr+auto	NPr+carro
[Honda]	NPr+org								
[US$ 30 mil]					Num+mon	Num+mon		Num+din	Num+din
[US$]			NPr+moe						
[JFK -- A PERGUNTA QUE NÃO QUER CALAR]		NPr+obra	NPr+tit	NPr+out+semtit	NPr+oCine			NPr+tit	NPr+pess
[JFK]							NPr+obra		
[5ª Conferência Internacional sobre Transtornos Alimentares]	NPr+out	NPr+even		NPr+out+even	NPr+even	NPr+lug	NPr+org	NPr+even	NPr+conf
[Conferência Internacional sobre Transtornos Alimentares]			NPr+org+conf						
[de 29 de abril a 1º de maio]					Temp+data	Temp+data			
[29 de abril a 1º de maio]									Temp+data
[29 de abril]								Temp+data	
[1º de maio]								Temp+data	
[Escola de Medicina de Harvard]	NPr+org *[Escola de Medicina de [Harvard]]*	NPr+org	NPr+org+esc	NPr+lug+inst	NPr+inst	NPr+lug			
[Escola de Medicina]							NPr+org	NPr+org	NPr+lug
[Harvard]	NPr+lug						NPr+lug	NPr+lug	

Tabela B.4: Lista das entidades em que não houve acordo quanto à sua segmentação no CETENFolha.

Apêndice C

Tabelas de valores *p*

		palavras	siemes1	siemes2	rena	nooj2no	nerua_em	nerua_cp	nerua_ct	malinche	elle	cortex2	cortex1	cortex3no	nilc	nooj1	cage3	cage2	cage1
		79,18%	0,0010	0,0010	**0,3150**	**0,0400**	0,0010	0,0010	0,0010	0,0010	**0,4670**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
palavras	Abr.	83,31%	**0,6510**	**0,0230**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,8119	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	76,75%	0,0010	**0,0130**	0,0010	**0,0180**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
siemes1	Abr.	83,11%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,7981	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	76,39%	0,0010	0,0010	0,0010	0,0030	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
siemes2	Abr.	82,29%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,7923	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	78,76%	0,0010	**0,2060**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,9840**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
rena	Abr.	68,17%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,7308	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	78,03%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,3190**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
nooj2no	Abr.	56,42%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0040	0,0010	0,0010	0,0010	0,0010
	MedF	0,6549	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	57,47%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0060	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
nerua_em	Abr.	62,68%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5996	0,0010	0,0010	0,0010	0,0010	0,0010	**1,0000**	0,0010	**0,0230**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	59,70%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
nerua_cp	Abr.	62,68%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,6115	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	57,67%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
nerua_ct	Abr.	61,87%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5970	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,0690**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	49,68%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0090	0,0010	**0,8440**	**0,0500**	**0,0640**
malinche	Abr.	71,95%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5878	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	78,74%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
elle	Abr.	39,16%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5231	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,4920**	0,0010	**0,4170**	0,0010	**0,0820**	0,0010	0,0010	0,0010	0,0010
	Prec.	64,64%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,0240**	0,0010	0,0010	0,0010	**0,2570**	0,0010	0,0010	0,0010
cortex2	Abr.	43,33%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,0180**	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5188	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,0190**	0,0010	0,0010	**0,3540**	0,0010	0,0010	0,0010	0,0010
	Prec.	64,57%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,2960**	0,0010	0,0010	0,0010
cortex1	Abr.	43,22%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5178	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,4280**	0,0010	0,0010	0,0010	0,0010
	Prec.	66,37%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
cortex3no	Abr.	85,21%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,7462	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	48,52%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,1110**	**0,5950**	**0,6650**
nilc	Abr.	54,35%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,5127	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	63,54%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
nooj1	Abr.	14,96%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,2421	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	49,84%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,0230**	**0,0320**
cage3	Abr.	10,86%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,1783	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	Prec.	48,05%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,5140**
cage2	Abr.	7,09%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,3740**
	MedF	0,1236	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	**0,3740**
	Prec.	48,14%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
cage1	Abr.	7,12%	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010
	MedF	0,1240	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010	0,0010

Tabela C.1: Valores de p para a tarefa de identificação do evento de 2005.

			palavras	siemes1	siemes2	rena	nooj2no	nerua_cp	nerua_ct	elle	cortex2	cortex1	nooj1	cage3	cage2	cage1
palavras	Prec.	58,22%		0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	58,68%		0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5845		0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes1	Prec.	49,23%			0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	**0,8348**	**0,8252**	**0,0714**	0,0001	0,0001	0,0001
	Abr.	45,96%			**0,0219**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4754			0,0005	0,0001	0,0001	0,0001	0,0001	0,0004	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes2	Prec.	50,21%				0,0001	0,0001	0,0001	0,0001	0,0001	**0,2269**	**0,1031**	**0,4985**	0,0001	0,0001	0,0001
	Abr.	46,96%				0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4853				0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
rena	Prec.	31,83%					0,0001	**0,9108**	0,0001	0,0001	0,0001	0,0001	0,0001	**0,0186**	0,0001	0,0001
	Abr.	18,21%					0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,2316					0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
nooj2no	Prec.	44,01%						0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	30,47%						0,0001	0,0029	**0,0371**	**0,0812**	**0,1155**	0,0001	0,0001	0,0001	0,0001
	MedF	0,3601						0,0001	0,0001	0,0001	0,0004	0,0012	0,0001	0,0001	0,0001	0,0001
nerua_cp	Prec.	31,90%							0,0001	0,0001	0,0001	0,0001	0,0001	**0,0123**	0,0001	0,0001
	Abr.	22,21%							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,2619							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
nerua_ct	Prec.	27,19%								0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	28,67%								0,0005	0,0010	0,0018	0,0001	0,0001	0,0001	0,0001
	MedF	0,2791								0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
elle	Prec.	64,18%									0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	32,59%									**0,7680**	**0,6374**	0,0001	0,0001	0,0001	0,0001
	MedF	0,4323									0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
cortex2	Prec.	49,37%										0,0001	**0,1108**	0,0001	0,0001	0,0001
	Abr.	32,28%										**0,8941**	0,0001	0,0001	0,0001	0,0001
	MedF	0,3904										**0,8255**	0,0001	0,0001	0,0001	0,0001
cortex1	Prec.	49,08%											**0,0586**	0,0001	0,0001	0,0001
	Abr.	32,08%											0,0001	0,0001	0,0001	0,0001
	MedF	0,3880											0,0001	0,0001	0,0001	0,0001
nooj1	Prec.	50,82%												0,0001	0,0001	0,0001
	Abr.	10,92%												0,0001	0,0001	0,0001
	MedF	0,1797												0,0001	0,0001	0,0001
cage3	Prec.	34,04%													0,0001	0,0001
	Abr.	5,81%													0,0006	0,0008
	MedF	0,0993													0,0010	0,0016
cage2	Prec.	37,94%														**0,5047**
	Abr.	4,37%														**0,9616**
	MedF	0,0784														**0,9592**
cage1	Prec.	38,04%														
	Abr.	4,39%														
	MedF	0,0787														

Tabela C.2: Valores de p para a tarefa de classificação semântica (na medida combinada) do evento de 2005.

		cortex3	cortex1rem	siemes_total2	siemes_total1	siemes_difuso2	siemes_difuso1	cortex2cem	siemes_prefixo2	siemes_prefixo4	smell2	siemes_exact	stencil_polcor	smell1	stencil_polcro	siemes_posterior	stencil_dicpro	stencil_1	siemes_simples	cage2	cage1
cortex3	Prec.	80,23%	**1,0000**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0022	0,0001	0,0001	0,0001	0,0001	0,0001	**0,0138**	0,0001
	Abr.	88,04%	**1,0000**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,8395	**1,0000**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
cortex1rem	Prec.	80,23%	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0012	0,0001	0,0002	0,0001	0,0001	0,0001	**0,0143**	0,0002
	Abr.	88,04%	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,8395	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_total2	Prec.	72,72%	**1,0000**		**0,0203**	0,0001	**0,0115**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	71,62%	**1,0000**		0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,7216	**1,0000**		0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_total1	Prec.	72,72%			**0,0213**	**0,0155**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	71,62%			0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,7216			0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_difuso2	Prec.	72,91%				**0,4991**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	70,18%				**0,4991**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,7152				**0,4991**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_difuso1	Prec.	72,93%					0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	70,18%					0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,7152					0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
cortex2cem	Prec.	87,39%						0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0023	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	56,69%						0,0001	**0,5324**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,6877						0,0001	0,0001	**0,0817**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_prefixo2	Prec.	75,33%							**0,1191**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0003	**0,0587**
	Abr.	65,04%							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,6981							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_prefixo4	Prec.	75,16%									0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	**0,0460**	0,0002	**0,0398**
	Abr.	61,81%									0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,6783									0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
smell2	Prec.	69,84%										0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	56,34%										0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,6236										0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_exact	Prec.	83,14%											**0,1740**	0,0001	0,0002	**0,8510**	0,0012	**0,1442**	0,0001	0,0001	0,0001
	Abr.	42,83%											**0,7337**	**0,0887**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5654											**0,5051**	**0,6871**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
stencil_polcor	Prec.	84,10%												0,0001	0,0001	**0,1360**	0,0001	**0,5468**	0,0001	0,0001	0,0001
	Abr.	43,04%												**0,1342**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5694												**0,2708**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
smell1	Prec.	78,33%													0,0001	0,0001	0,0028	0,0001	**0,0568**	**0,9739**	**0,0677**
	Abr.	43,92%													0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5629													0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
stencil_pol	Prec.	85,69%														0,0005	0,0001	0,0002	0,0001	0,0001	0,0001
	Abr.	38,94%														0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5355														0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_posterior	Prec.	83,04%															0,0019	**0,1097**	0,0001	0,0001	0,0001
	Abr.	36,43%															0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,5064															0,0001	0,0001	0,0001	0,0001	0,0001
stencil_dicpro	Prec.	80,59%																0,0001	0,0001	**0,0211**	0,0001
	Abr.	33,02%																0,0001	0,0001	0,0001	0,0001
	MedF	0,4684																0,0001	0,0001	0,0001	0,0001
stencil_1	Prec.	84,36%																	0,0001	0,0001	0,0001
	Abr.	25,56%																	0,0001	0,0001	0,0001
	MedF	0,3923																	0,0001	0,0001	0,0001
siemes_simples	Prec.	76,62%																		**0,1578**	**0,7947**
	Abr.	16,91%																		**0,1573**	**0,0370**
	MedF	0,2771																		**0,2207**	**0,0490**
cage2	Prec.	78,39%																			0,0001
	Abr.	16,05%																			0,0001
	MedF	0,2664																			0,0001
cage1	Prec.	76,96%																			
	Abr.	15,66%																			
	MedF	0,2603																			

Tabela C.3: Valores de p para a tarefa de identificação do evento de 2006.

			cortex3	siemes_total2	siemes_total1	siemes_difuso2	siemes_difuso1	cortex2cem	siemes_prefixo2	siemes_prefixo4	smell2	siemes_exact	stencil_polcor	smell1	stencil_pol	siemes_posterior	stencil_dicpro	stencil_1	siemes_simples	cage2	cage1
cortex3	Prec.	52,56%		0,0086	0,0014	0,0001	0,0001	0,0001	**0,1462**	0,0001	**0,3174**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	45,63%		0,0010	0,0017	0,0001	0,0001	**0,2465**	**0,9995**	**0,6016**	0,0003	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4885		**0,1826**	**0,2973**	0,0001	0,0001	0,0001	**0,6230**	**0,3596**	0,0009	0,0001	**0,4248**	0,0005	**0,0193**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_total2	Prec.	51,03%			0,0029	0,0001	0,0001	0,0001	0,0001	0,0001	**0,1605**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	48,53%			**0,8821**	0,0001	0,0001	**0,3638**	**0,0130**	0,0065	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4975			**0,6915**	0,0001	0,0001	0,0001	**0,4614**	**0,8512**	0,0001	0,0001	**0,0694**	0,0001	0,0003	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_total1	Prec.	50,78%				0,0001	0,0001	0,0001	0,0001	0,0001	**0,0736**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	48,38%				0,0001	0,0001	**0,4229**	**0,0198**	0,0088	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4955				0,0001	0,0001	0,0001	**0,6506**	**0,9470**	0,0001	0,0001	**0,0986**	0,0001	0,0003	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_difuso2	Prec.	43,25%					0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	38,97%					0,0001	0,0001	0,0001	0,0001	**0,0278**	0,0001	**0,0296**	0,0041	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4100					0,0001	0,0001	0,0001	0,0001	0,0001	**0,1834**	0,0001	0,0001	0,0001	0,0001	**0,9309**	0,0001	0,0001	0,0001	0,0001
siemes_difuso1	Prec.	34,16%						0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	30,53%						0,0001	0,0001	0,0001	0,0001	**0,3063**	0,0001	0,0003	**0,0360**	0,0001	**0,2330**	0,0001	0,0001	0,0001	0,0001
	MedF	0,3224						0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
cortex2cem	Prec.	73,49%							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	**0,2319**	0,0001	0,0001	**0,2566**	0,0001	0,0001	0,0001
	Abr.	47,37%							**0,0588**	**0,0131**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0066	0,0001	0,0001
	MedF	0,5761							0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_prefixo2	Prec.	53,43%								0,0001	**0,0294**	0,0001	0,0001	0,0001	**0,0255**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	45,63%								**0,6309**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4923								**0,6203**	0,0002	0,0001	**0,1704**	0,0002	0,0011	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_prefixo4	Prec.	55,22%									0,0001	0,0001	0,0001	0,0001	**0,6679**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	45,02%									0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4960									0,0001	0,0001	**0,0597**	0,0001	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
smell2	Prec.	51,94%										0,0001	0,0001	0,0001	0,0001	0,0005	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	41,05%										0,0001	0,0003	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4586										0,0019	**0,0714**	**0,2937**	**0,7548**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
siemes_exact	Prec.	63,45%											0,0001	**0,0112**	0,0001	0,0001	0,0001	**0,0169**	0,0003	0,0001	0,0076
	Abr.	32,15%											0,0003	0,0030	**0,3709**	0,0001	0,0041	0,0001	0,0001	0,0001	0,0001
	MedF	0,4267											0,0001	**0,0371**	0,0091	0,0001	**0,1046**	0,0001	0,0001	0,0001	0,0001
stencil_polcor	Prec.	70,49%												0,0001	0,0001	0,0001	0,0001	**0,0562**	**0,1873**	0,0001	0,0026	0,0001
	Abr.	36,12%												**0,3843**	**0,0928**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4776												0,0011	**0,1602**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
smell1	Prec.	61,21%													0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	35,22%													**0,0814**	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
	MedF	0,4471													**0,4763**	0,0001	0,0012	0,0001	0,0001	0,0001	0,0001
stencil_pol	Prec.	72,65%														0,0001	0,0001	0,0045	0,0001	0,0001	0,0001
	Abr.	33,12%														0,0001	0,0028	0,0001	0,0001	0,0001	0,0001
	MedF	0,4550														0,0001	0,0013	0,0001	0,0001	0,0001	0,0001
siemes_posterior	Prec.	54,92%															0,0001	0,0001	0,0001	0,0001	0,0001
	Abr.	23,72%															0,0001	**0,0719**	0,0001	0,0001	0,0001
	MedF	0,3313															0,0001	**0,7236**	0,0001	0,0001	0,0001
stencil_dicpro	Prec.	69,37%																0,0002	0,0001	**0,1133**	**0,0148**
	Abr.	29,00%																0,0001	0,0001	0,0001	0,0001
	MedF	0,4090																0,0001	0,0001	0,0001	0,0001
stencil_1	Prec.	71,23%																	0,0001	0,0024	0,0003
	Abr.	21,94%																	0,0001	0,0001	0,0001
	MedF	0,3354																	0,0001	0,0001	0,0001
siemes_simples	Prec.	65,56%																		**0,1986**	**0,5485**
	Abr.	14,57%																		0,0001	0,0001
	MedF	0,2385																		0,0001	0,0001
cage2	Prec.	67,50%																			0,0003
	Abr.	7,81%																			**0,7895**
	MedF	0,1401																			**0,7696**
cage1	Prec.	66,49%																			
	Abr.	7,65%																			
	MedF	0,1373																			

Tabela C.4: Valores de p para a tarefa de classificação semântica (na medida combinada) do evento de 2006.

Apêndice D

Documentação técnica da plataforma de avaliação

D.1 Instalação e configuração

Os módulos foram desenvolvidos por Nuno Seco, Nuno Cardoso e Rui Vilela, e encontram-se disponíveis no sítio do HAREM, em `http://poloxldb.linguateca.pt/harem.php?l=programas`. Qualquer investigador tem acesso livre a estes programas e pode usá-los para avaliar o desempenho do seu sistema de REM, e compará-lo com os resultados obtidos pelos outros sistemas em avaliações conjuntas passadas. Dado que o código fonte também foi incluído nos pacotes de distribuição, qualquer utilizador pode estender e melhorar os programas.

Visto que alguns módulos foram programados em Perl, e outros em Java, a plataforma está disponível através de dois pacotes:

`ferramentas_HAREM_java.jar`, o pacote de módulos programados em Java, nomeadamente os módulos AlinhEM, AvalIDa, Véus, Emir, AltinaID, AltinaSEM, Ida2ID, Ida2SEM e Sultão.

`ferramentas_HAREM_perl.tar.gz`, o pacote de módulos programados em Perl, nomeadamente os módulos Extractor, Vizir, AltinaMOR, Ida2MOR e Alcaide.

A versão 1.5 do Java e a versão 5.8 do Perl foram usadas no desenvolvimento dos módulos, em ambiente Linux, e segundo a codificação de caracteres `ISO-8859-1`. Não é necessário nenhum procedimento de instalação para executar os módulos desenvolvidos em Java, sendo contudo necessária a presença da *Java Virtual Machine (JVM)* para a sua execução. Para executar os módulos desenvolvidos em Perl, é primeiro necessário instalar os módulos. Para tal, executa-se os seguinte comando:

```
tar xfz ACMorf.tar.gz
perl Makefile.PL
make
make install
```

Na mesma directoria onde se encontra o ficheiro `ferramentas_HAREM_java.jar`, é obrigatório existir um ficheiro chamado `harem.conf`, que descreve os géneros textuais, variantes, categorias e tipos válidos para a avaliação. O apêndice D.3 inclui o ficheiro `harem.conf` usado no Mini-HAREM.

Para a execução de módulos programados em Java, é necessário especificar na linha de comandos o parâmetro `-Dfile.encoding=ISO-8859-1`, de modo a garantir que os ficheiros sejam processados utilizando codificação de caracteres correcta. Na execução de módulos programados em Perl, é necessário verificar se o ambiente de execução é de codificação `ISO-8859-1`. O Alcaide requer, além disso para a geração dos gráficos, os módulos Perl GD-2.28, GDGraph-1.43 e GDTextUtil-0.86 (as versões dos módulos referidas são as versões utilizadas e testadas).

Dentro do programa Alcaide, é também necessário configurar os seguintes parâmetros, antes da sua execução:

$directoria_identificacao - directoria com os relatórios do SultãoID

$directoria_morfologia - directoria com os relatórios do SultãoMOR

$directoria_semantica - directoria com os relatórios do SultãoSEM

$directoria_ida - directoria com os relatórios dos programas ida2ID, ida2MOR e ida2SEM. Esta directoria deverá manter a estrutura de directorias, ou seja, uma directoria com o nome da saída, e sobre esta uma directoria para cada tarefa (`identificacao`, `morfologia` ou `semantica`), e debaixo das directorias `morfologia` e `semantica`, directorias `absoluto` e `relativo`.

D.2 Utilização

D.2.1 Extractor

Para executar o Extractor, usa-se o seguinte comando:

```
perl extrairCDdasSubmissoes.pl -in FICHEIRO_ENTRADA
-out FICHEIRO_SAIDA -cdids FICHEIRO_CDIDS
```

`FICHEIRO_ENTRADA` corresponde ao ficheiro da saída do sistema REM, a partir do qual serão extraídos os documentos correspondentes à CD para um novo ficheiro,

FICHEIRO_SAIDA. Os identificadores dos documentos a retirar (que, normalmente, correspondem aos identificadores dos documentos da CD) são lidos do ficheiro FICHEIRO_CDIDS, que deve conter uma lista com os últimos cinco números de cada DOCID, um por cada linha (no exemplo HAREM-87J-07845, o valor a colocar seria 07845).

Nota: Os ficheiros de identificadores das CD de 2005 e de 2006 (FICHEIRO_CDIDS) estão incluídos no pacote ferramentas_HAREM_perl.tar.gz.

D.2.2 AlinhEM

Para executar o AlinhEM, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.Aligner -submissao FICHEIRO_SUBMISSAO
-cd FICHEIRO_CD [-etiquetas sim|nao] [-ignorar FICHEIRO_ATOMOS]
> FICHEIRO_ALINHEM
```

FICHEIRO_SUBMISSAO corresponde ao nome do ficheiro pré-processado pelo Extractor, e FICHEIRO_CD corresponde ao ficheiro da CD. O resultado do alinhamento é enviado para o *standard output*, pelo que se recomenda o redireccionamento da saída para um ficheiro. Esse ficheiro, o FICHEIRO_ALINHEM, será usado pelo AvalIDa.

O AlinhEM possui dois parâmetros adicionais que podem ser usados na linha de comandos:

etiquetas, que pode ter os valores sim ou nao. A sintaxe é -etiquetas [sim|nao]. A opção nao é usada por defeito. Ao especificar o valor sim, o AlinhEM produz as etiquetas numéricas para identificar os átomos.

ignorar, que recebe como valor o nome de um ficheiro que contém uma lista de átomos que serão ignorados pelo AlinhEM. A sintaxe é -ignorar FICHEIRO_ATOMOS. O ficheiro FICHEIRO_ATOMOS deve ser composto por uma lista de átomos, um por linha.

D.2.3 AvalIDa

Para executar o AvalIDa, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.IndividualAlignmentEvaluator -alinhamento
FICHEIRO_ALINHEM > FICHEIRO_AVALIDA
```

O ficheiro FICHEIRO_ALINHEM corresponde ao nome do ficheiro gerado pelo AlinhEM, que contém os alinhamentos com as etiquetas numéricas. O resultado é enviado para o *standard output*, pelo que se recomenda o redireccionamento da saída para um ficheiro. Esse ficheiro, o FICHEIRO_AVALIDA, será usado pelos módulos Véus, AltinaID, Vizir e Emir.

O AvalIDa requer obrigatoriamente a opção `-alinhamento`, para especificar o ficheiro gerado pelo AlinhEM, o `FICHEIRO_ALINHEM`.

D.2.4 Véus

Para executar o Véus, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.AlignmentFilter -alinhamento FICHEIRO_AVALIDA
[-categoria CATEGORIAS] [-genero GENERO_TEXTUAL] [-origem VARIANTE]
[-estilo muc|relax|harem] > FICHEIRO_VEUS
```

`FICHEIRO_AVALIDA` corresponde ao ficheiro gerado pelo AvalIDa. O Véus escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_VEUS`.

O Véus pode receber até cinco parâmetros de entrada. Só o parâmetro `-alinhamento` é obrigatório, sendo os restantes parâmetros facultativos. Estes parâmetros podem ser combinados de várias formas, de modo a obter o filtro desejado.

-alinhamento, que deve vir acompanhado do nome do ficheiro gerado pelo AvalIDa, `FICHEIRO_AVALIDA`.

-categoria, que especifica as categorias e/ou tipos que devem ser filtradas. O argumento do parâmetro, `CATEGORIAS`, é uma lista de categorias separadas por ':'. Por exemplo, a lista `'PESSOA:ORGANIZACAO:ABSTRACCAO'` faz com que o Véus escreva para o *standard output* todos os alinhamentos que contêm EM de qualquer uma das categorias `PESSOA`, `ORGANIZACAO` ou `ABSTRACCAO`. Note-se que basta existir apenas uma referência à categoria e/ou tipo num dado alinhamento (ou seja, tanto nas EM da CD como nas EM da saída) para que este seja considerado e escrito.

A restrição nos tipos é representada por uma lista de tipos entre parênteses imediatamente a seguir à respectiva categoria. Por exemplo, a lista `'PESSOA(CARGO,GRUPOMEMBRO):ORGANIZACAO'` filtra os alinhamentos para procurar EM de categorias `ORGANIZACAO` e `PESSOA`, sendo que só tipos `CARGO` e `GRUPOMEMBRO` é que são tidos em conta para a categoria `PESSOA`.

-genero, que especifica o(s) género(s) textual(is) a filtrar. Recebe uma lista de géneros separados por ':', ou então um único género textual. Os valores da lista devem estar mencionados na lista `GENEROS` do ficheiro `harem.conf`. Por exemplo, ao especificar `-genero Web`, o Véus escreve todos os alinhamentos de documentos de género textual Web.

-origem, que especifica a(s) variante(s) a filtrar. Recebe uma lista de variantes separadas por ':', ou então uma variante. Os valores da lista devem estar mencionados na lista `ORIGENS` do ficheiro `harem.conf`. Por exemplo, ao especificar `-origem PT`, o Véus filtra e escreve todos os alinhamentos de documentos da variante portuguesa.

-estilo, que pode ter um dos três valores seguintes: **muc**, **relax** e **harem**. Com o valor **muc**, o Véus retira todos os alinhamentos que geraram uma pontuação `parcialmente_correcto`, o que simula o cenário da avaliação dos MUC-6 e MUC-7, que não reconhecia este tipo de pontuação.Com o valor **relax**, o Véus aceita apenas no máximo uma pontuação `parcialmente_correcto` por cada de alinhamento a uma EM na CD. Ou seja, nos casos em que a EM na CD alinhe com várias EM da saída, ou uma EM da saída alinhe com várias EM da CD (gerando várias pontuações `parcialmente_correcto`), só o primeiro alinhamento é pontuado com `parcialmente_correcto`, enquanto que os restantes serão classificadas como `espurio` ou `em_falta`). Esta opção pode ser vista como uma restrição aos alinhamentos múltiplos. Finalmente, com a opção `harem`, todos os alinhamentos parcialmente correctos são considerados para avaliação.

D.2.5 AltinaID

Para executar o AltinaID, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.IdentificationAltAlignmentSelector -alinhamento
FICHEIRO_VEUS > FICHEIRO_ALTINAID
```

`FICHEIRO_VEUS` corresponde ao ficheiro gerado pelo Véus (ou, no caso de não se querer filtrar alinhamentos, pode-se usar o ficheiro gerado pelo AvalIDa). O AltinaID escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_ALTINAID`.

D.2.6 Ida2ID

Para executar o Ida2ID, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalIdentificationSelector -alinhamento
FICHEIRO_ALTINAID > FICHEIRO_IDA2ID
```

`FICHEIRO_ALTINAID` corresponde ao ficheiro gerado pelo AltinaID, ou seja, sem nenhuma alternativa `<ALT>`. O Ida2ID escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_IDA2ID`.

D.2.7 Emir

Para executar o Emir, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.SemanticAlignmentEvaluator -alinhamento
FICHEIRO_ALTINAID [-relativo sim] > FICHEIRO_EMIR
```

`FICHEIRO_ALTINAID` corresponde ao ficheiro gerado pelo AltinaID, ou seja, já sem nenhuma etiqueta `<ALT>`. O Emir escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_EMIR`.

O Emir aceita o parâmetro opcional `-relativo` com o valor **sim**, para assinalar ao Emir que a avaliação deve ser realizada segundo o cenário relativo (iso é, considerando apenas as EM identificadas como correctas ou parcialmente correctas pela saída). Se nada for especificado, o Emir avalia segundo um cenário absoluto (ou seja, considerando todas as EM da CD, incluindo as que não foram identificadas como correctas ou parcialmente correctas pelo sistema).

D.2.8 AltinaSEM

Para executar o AltinaSEM, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.SemanticAltAlignmentSelector -alinhamento
FICHEIRO_EMIR > FICHEIRO_ALTINASEM
```

`FICHEIRO_EMIR` corresponde ao ficheiro gerado pelo Emir. O AltinaSEM escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_ALTINASEM`.

D.2.9 Ida2SEM

Para executar o Ida2SEM, usa-se o seguinte comando:

```
java -Dfile.encoding=ISO-8859-1 -cp ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalSemanticSelector -alinhamento
FICHEIRO_ALTINASEM > FICHEIRO_IDA2SEM
```

`FICHEIRO_ALTINASEM` corresponde ao ficheiro gerado pelo AltinaSEM. O Ida2SEM escreve para o *standard output*, pelo que se recomenda o seu redireccionamento para um ficheiro, o `FICHEIRO_IDA2SEM`.

D.2.10 Vizir

Para executar o Vizir, usa-se o seguinte comando:

```
vizir.pl [-abs|-rel] -i FICHEIRO_VEUS|AVALIDA -o FICHEIRO_VIZIR
```

O parâmetro `-i` é obrigatório e especifica o ficheiro gerado pelo Véus ou pelo AvalIDa, `FICHEIRO_VEUS|AVALIDA`. O parâmetro `-o` especifica o ficheiro de escrita do Vizir, `FICHEIRO_VIZIR`. Caso esta opção não seja preenchida, é usado o nome do ficheiro `FICHEIRO_VEUS|AVALIDA`, acrescido da extensão `.vizir`.

O Vizir obriga a especificar o tipo de cenário a usar na avaliação. Para tal, é necessário optar por um dos seguintes parâmetros: `-abs`, para cenário absoluto que considera todas as EM para avaliação, ou `-rel`, para cenário relativo, que não considera as EM espúrias nem com classificação morfológica espúria.

D.2.11 AltinaMOR

Para executar o AltinaMOR, usa-se o seguinte comando:

```
altinamor.pl [-abs|-rel] -i FICHEIRO_VIZIR -o FICHEIRO_ALTINAMOR
```

O parâmetro `-i` é obrigatório e especifica o ficheiro gerado pelo Vizir, `FICHEIRO_VIZIR`. O parâmetro `-o` especifica o ficheiro de escrita do AltinaMOR, `FICHEIRO_ALTINAMOR`. Caso esta opção não seja especificada, é usado o nome do `FICHEIRO_VIZIR`, mais a extensão `.altmor`.

D.2.12 Ida2MOR

Para executar o Ida2MOR, usa-se o seguinte comando:

```
ida2mor.pl [-abs|-rel] -i FICHEIRO_ALTINAMOR -o FICHEIRO_IDA2MOR
```

O parâmetro `-i` é obrigatório e especifica o ficheiro gerado pelo AltinaMOR, `FICHEIRO_ALTINAMOR`. O parâmetro `-o` especifica o ficheiro criado pelo Ida2MOR, `FICHEIRO_IDA2MOR`. Caso esta opção não seja preenchida, é usado o nome do `FICHEIRO_ALTINAMOR`, acrescido da extensão `.ida2mor`.

D.2.13 Sultão

Para executar os três módulos do Sultão, omeadamente SultãoID, SultãoMOR e SultãoSEM, usam-se os seguintes comandos, respectivamente:

```
java -Dfile.encoding=ISO-8859-1 -jar ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalIdentificationReporter [-filtro FILTRO]
[-naooficiais LISTA_NAOOFICIAIS] [-depurar sim|nao]
[-saidas oficiais|naooficiais] > FICHEIRO_SULTAOID

java -Dfile.encoding=ISO-8859-1 -jar ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalMorphologyReporter [-filtro FILTRO]
[-naooficiais LISTA_NAOOFICIAIS] [-depurar sim|nao]
[-saidas oficiais|naooficiais] > FICHEIRO_SULTAOMOR

java -Dfile.encoding=ISO-8859-1 -jar ferramentas_HAREM_java.jar
pt.linguateca.harem.GlobalSemanticReporter [-filtro FILTRO]
[-naooficiais LISTA_NAOOFICIAIS] [-depurar sim|nao]
[-saidas oficiais|naooficiais] [-tipos sim|nao] > FICHEIRO_SULTAOSEM
```

O Sultão é executado com os seguintes parâmetros opcionais, que podem ser combinados entre si:

-filtro, que diz respeito aos ficheiros que deverão ser utilizados na geração dos relatórios, e recebe como valor o sufixo do ficheiro. Por exemplo, se usar no `FILTRO` o valor '`total.altid.ida2id`', o Sultão processa todos os ficheiros terminados com a extensão `total.altid.ida2id`. Se se pretende mais do que um padrão de ficheiros, pode-se utilizar uma lista de extensões separadas por ':', como por exemplo em `total.local.altid.ida2id:total.organizacao.altid.ida2id`.

-naooficiais, que indica ao Sultão quais os ficheiros que correspondem a saídas não oficiais entregues pelos participantes. O parâmetro recebe como valor o prefixo do ficheiro, que deve ter o nome da saída, como no seguinte exemplo:

`-naooficiais sistema1_nao_oficial:sistema4`

O exemplo indica que os ficheiros cujos nomes começam por `sistema1_nao_oficial` ou `sistema4` são para ser considerados não oficiais, e a sua entrada na tabela de resultados não vai ter o pseudónimo a negrito, mas sim a itálico.

-saidas, que indica ao Sultão as saídas que devem ser consideradas. A este parâmetro podem ser atribuídos dois valores: `oficiais` e `naooficiais`. No primeiro caso, só as saídas oficiais é que serão exportadas para o relatório final. No segundo, só as saídas não oficiais é que são consideradas. Se este parâmetro não for utilizado, todas as saídas são consideradas.

-depurar, que pode tomar os valores `sim` ou `nao`. Por defeito, o Sultão assume que a informação para depuração não é para ser colocada no relatório e que a anonimização é

para ser efectuada. Se o parâmetro for fornecido com o valor `sim`, então a anonimização não é efectuada e informação adicional é colocada no relatório final.

-tipos, parâmetro usado apenas no SultãoSEM, e que pode tomar os valores `sim` ou `nao`. Este parâmetro indica ao SultãoSEM se as tabelas referente à avaliação dos tipos devem ou não ser produzidas. Este opção existe uma vez que a avaliação dos tipos é sempre relativa (porque só se avaliam os tipos quando a categoria está correcta), logo os valores destas tabelas seriam sempre iguais na avaliação absoluta e relativa.

D.2.14 Alcaide

para executar o Alcaide, usa-se o seguinte comando:

```
perl alcaide.pl -sistema SISTEMA -run SAIDA -id ID -morf MORF
     -sem SEM -output SAIDA -workingdir DIRECTORIA
```

O Alcaide necessita obrigatoriamente dos seguintes parâmetros:

-sistema, com o nome do sistema que gerou a saída.

-run, com o nome da saída. Este nome deve ser exactamente igual ao nome da directoria que contém os relatórios de entrada, e também ao nome pelo qual começam os nomes dos ficheiros gerados pelos programas Ida2ID, Ida2MOR e Ida2SEM.

-id, que pode tomar o valor de 0 ou 1, assinala ao Alcaide que se pretende gerar tabelas da tarefa de identificação para o relatório individual.

-morf, que pode tomar o valor de 0 ou 1, assinala ao Alcaide que se pretende gerar tabelas da tarefa de classificação morfológica para o relatório individual.

-sem, que pode tomar o valor de 0 ou 1. Diz ao Alcaide que se pretende gerar tabelas da tarefa de classificação semântica para o relatório individual.

-output, que indica a directoria onde o Alcaide irá escrever o relatório. Esta directoria tem de conter uma subdirectoria chamada `images`, para armazenar as imagens que são criadas automaticamente pelo programa.

-workingdir, que designa a directoria raiz com os relatórios do Sultão, Ida2ID, Ida2MOR e Ida2SEM.

D.3 Ficheiro de configuração do HAREM, `harem.conf`

Neste apêndice, apresenta-se o ficheiro `harem.conf` usado no Mini-HAREM para definir as categorias e tipos válidos, bem como os géneros textuais e variantes autorizadas.

```
[ENTIDADES]
PESSOA:INDIVIDUAL,CARGO,GRUPOIND,GRUPOMEMBRO,MEMBRO,GRUPOCARGO
ORGANIZACAO:ADMINISTRACAO,EMPRESA,INSTITUICAO,SUB
TEMPO:DATA,HORA,PERIODO,CICLICO
LOCAL:CORREIO,ADMINISTRATIVO,GEOGRAFICO,VIRTUAL,ALARGADO
OBRA:ARTE,REPRODUZIDA,PUBLICACAO
ACONTECIMENTO:EFEMERIDE,ORGANIZADO,EVENTO
ABSTRACCAO:DISCIPLINA,ESTADO,ESCOLA,MARCA,PLANO,IDEIA,NOME,OBRA
COISA:CLASSE,SUBSTANCIA,OBJECTO,MEMBROCLASSE
VALOR:CLASSIFICACAO,QUANTIDADE,MOEDA
VARIADO:OUTRO

[GENEROS]
CorreioElectrónico
Entrevista
Expositivo
Jornalístico
Literário
Político
Técnico
Web

[ORIGENS]
AO
BR
CV
IN
MO
MZ
PT
TL
```

Apêndice E

Exemplos da invocação dos programas de avaliação

E.1 Exemplos do Emir

O seguinte exemplo ilustra o funcionamento do Emir, sobre os alinhamentos gerados pelo Véus sobre uma saída AvalIDa, de um documento hipotético. A entrada para o Emir possui os seguintes 9 alinhamentos:

```
#PESSOA=["GRUPOCARGO", "GRUPOMEMBRO"]; LOCAL=["GEOGRAFICO", "ALARGADO", "ADMINISTRATIVO",
"CORREIO"]; ORGANIZACAO=["INSTITUICAO", "ADMINISTRACAO", "EMPRESA", "SUB"]

HAREM-000-00000 PT Web

1. <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Freguesia de Itapecerica</LOCAL> ---> [<LOCAL
TIPO="ADMINISTRATIVO">Freguesia de Itapecerica pela Lei Provincial</LOCAL>]:
[Parcialmente_Correcto_por_Excesso(0.25; 0.75)]

2. <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Baú</LOCAL> ---> [null]:[Em_Falta]

3. <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Baú</LOCAL> ---> [<LOCAL TIPO="CORREIO" MORF="M,S">
Baú</LOCAL>]:[Correcto]

4. <ESPURIO>Porta da Esperança</ESPURIO> ---> [<LOCAL TIPO="GEOGRAFICO" MORF="F,S">Porta da
Esperança</LOCAL>]:[Espúrio]

5. <ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Estado Maior do Exército da República Federal da
Alemanha</ORGANIZACAO> ---> [<LOCAL TIPO="ADMINISTRATIVO">Estado Maior</LOCAL>, <LOCAL
TIPO="ADMINISTRATIVO">Alemanha</LOCAL>]:[Parcialmente_Correcto_por_Defeito(0.1111111111111111;
0.8888888888888888), Parcialmente_Correcto_por_Defeito(0.05555555555555555; 0.9444444444444444)]

6. <LOCAL|ORGANIZACAO TIPO="ALARGADO|EMPRESA" MORF="?,S">Planet Dance</LOCAL|ORGANIZACAO> --->
[<ORGANIZACAO TIPO="EMPRESA">Planet</ORGANIZACAO>]:[Parcialmente_Correcto_por_Defeito(0.25; 0.75)]

7. <PESSOA|ORGANIZACAO TIPO="GRUPOCARGO|SUB" MORF="M,S">Conselho de Administração</PESSOA|ORGANIZACAO>
---> [<ORGANIZACAO TIPO="ADMINISTRACAO">Conselho de Administração</ORGANIZACAO>]:[Correcto]

8. <ORGANIZACAO|LOCAL TIPO="INSTITUICAO|ALARGADO" MORF="F,S">Prisão de Caxias</ORGANIZACAO|LOCAL> --->
```

```
[null]:[Em_Falta]

9. <ORGANIZACAO|ORGANIZACAO TIPO="ADMINISTRACAO|SUB" MORF="M,S">Conselho Legislativo</ORGANIZACAO|ORGANIZACAO>
---> [<PESSOA TIPO="GRUPOCARGO">Presidentes da Knesset e do Conselho Legislativo</PESSOA>]
:[Parcialmente_Correcto_por_Excesso(0.14285714285714285; 0.8571428571428572)]
```

Após o processamento pelo Emir o resultado obtido é o seguinte:

```
#PESSOA=["GRUPOCARGO", "GRUPOMEMBRO"]; LOCAL=["GEOGRAFICO", "ALARGADO", "ADMINISTRATIVO", "CORREIO"];
ORGANIZACAO=["INSTITUICAO", "ADMINISTRACAO", "EMPRESA", "SUB"]

HAREM-000-00000 PT Web

1. <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Freguesia de Itapecerica</LOCAL> ---> [<LOCAL
 TIPO="ADMINISTRATIVO">Freguesia de Itapecerica pela Lei Provincial</LOCAL>]:[{Categoria(Correcto:[LOCAL]
Espúrio:[] Em_Falta:[]) Tipo(Correcto:[ADMINISTRATIVO] Espúrio:[] Em_Falta:[]) CSC(1.75) Peso(0.5)}]
Comentário: Estamos perante um alinhamento em que as EM foram correctamente classificadas tanto em relação
às categorias como os tipos. Note-se que o Emir só está interessado na classificação semântica do alinhamento
e não delimitação/identificação definida para as EM.

2. <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Baú</LOCAL> ---> [null]:[{Categoria(Correcto:[] Espúrio:[]
Em_Falta:[LOCAL]) Tipo(Correcto:[] Espúrio:[] Em_Falta:[]) CSC(0.0) Peso(0.0)}]
Comentário: Como não existe nenhuma EM identificada pelo sistema que alinhe com a EM da CD, a categoria LOCAL
é considerada em em falta. Note-se que como a categoria não foi correctamente atribuída, os tipos não são analisados.

3. <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Baú</LOCAL> ---> [<LOCAL TIPO="CORREIO" MORF="M,S">Baú</LOCAL>]:
[{Categoria(Correcto:[LOCAL] Espúrio:[] Em_Falta:[]) Tipo(Correcto:[] Espúrio:[CORREIO] Em_Falta:[ADMINISTRATIVO])
CSC(1.0) Peso(1.0)}]
Comentário: A categoria está correctamente atribuída, consequentemente os tipos são analisados. Como o tipo
considerado pelo sistema, CORREIO, não é o mesmo que está na CD, ADMINISTRATIVO, estes são considerados espúrios
e em falta respectivamente.

4. <ESPURIO>Porta da Esperança</ESPURIO> ---> [<LOCAL TIPO="GEOGRAFICO" MORF="F,S">Porta da Esperança</LOCAL>]:
[{Categoria(Correcto:[] Espúrio:[LOCAL] Em_Falta:[]) Tipo(Correcto:[] Espúrio:[] Em_Falta:[]) CSC(0.0) Peso(0.0)}]
Comentário: No caso de alinhamentos espúrios o Emir limita-se a considerar a categoria como espúria.

5. <ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Estado Maior do Exército da República Federal da Alemanha</ORGANIZACAO>
---> [<LOCAL TIPO="ADMINISTRATIVO">Estado Maior</LOCAL>, <LOCAL TIPO="ADMINISTRATIVO">Alemanha</LOCAL>]:[{Categoria
(Correcto:[] Espúrio:[LOCAL] Em_Falta:[ORGANIZACAO]) Tipo(Correcto:[] Espúrio:[] Em_Falta:[]) CSC(0.0)
Peso(0.2222222222222222)}, {Categoria(Correcto:[] Espúrio:[LOCAL] Em_Falta:[ORGANIZACAO]) Tipo(Correcto:[] Espúrio:[]
Em_Falta:[]) CSC(0.0) Peso(0.1111111111111111)}]
Comentário: No caso em que uma EM da CD é alinhada com mais do que uma EM identificada pelo sistema são gerados tuplos
de avaliação para cada EM identificada.

6. <LOCAL|ORGANIZACAO TIPO="ALARGADO|EMPRESA" MORF="?,S">Planet Dance</LOCAL|ORGANIZACAO> ---> [<ORGANIZACAO
 TIPO="EMPRESA">Planet</ORGANIZACAO>]:[{Categoria(Correcto:[ORGANIZACAO] Espúrio:[] Em_Falta:[]) Tipo(Correcto:[EMPRESA]
Espúrio:[] Em_Falta:[]) CSC(1.75) Peso(0.5)}]
Comentário: Quando uma EM na CD é etiquetada com mais do que uma categoria, e o sistema tenha optado por atribuir apenas
uma categoria, basta que o sistema acerte uma delas para se considerar correcta. O mesmo aplica-se aos tipos.

7. <PESSOA|ORGANIZACAO TIPO="GRUPOCARGO|SUB" MORF="M,S">Conselho de Administração</PESSOA|ORGANIZACAO> ---> [<ORGANIZACAO
 TIPO="ADMINISTRACAO">Conselho de Administração</ORGANIZACAO>]:[{Categoria(Correcto:[ORGANIZACAO] Espúrio:[] Em_Falta:[])
Tipo(Correcto:[] Espúrio:[ADMINISTRACAO] Em_Falta:[SUB]) CSC(1.0) Peso(1.0)}]
Comentário: Neste alinhamento o sistema consegui acertar uma das categorias mas errou na etiquetação dos tipos dessa
categoria. Note-se que só as categorias que pertencem à categoria correcta é que são consideradas.

8. <ORGANIZACAO|LOCAL TIPO="INSTITUICAO|ALARGADO" MORF="F,S">Prisão de Caxias</ORGANIZACAO|LOCAL> ---> [null]:
[{Categoria(Correcto:[] Espúrio:[] Em_Falta:[ORGANIZACAO|LOCAL]) Tipo(Correcto:[] Espúrio:[] Em_Falta:[]) CSC(0.0)
```

```
Peso(0.0)}]
Comentário: No caso de etiquetações compostas em que nenhuma das categorias foi identificada pelo sistema considera-se
a composição de etiquetas em falta.

9. <ORGANIZACAO|ORGANIZACAO TIPO="ADMINISTRACAO|SUB" MORF="M,S">Conselho Legislativo</ORGANIZACAO|ORGANIZACAO> --->
[<PESSOA TIPO="GRUPOCARGO">Presidentes da Knesset e do Conselho Legislativo</PESSOA>]:[{Categoria(Correcto:[]
Espúrio:[PESSOA] Em_Falta:[ORGANIZACAO]) Tipo(Correcto:[] Espúrio:[] Em_Falta:[]) CSC(0.0) Peso(0.2857142857142857)}]
Comentário: Quando a mesma categoria é utilizada mais do que uma vez de forma a permitir variar os tipos, só a categoria
individualmente é que é considerada em falta.
```

E.2 Exemplos do Vizir

Para clarificar o funcionamento do Vizir, os seguintes exemplos de resultados de alinhamentos processados pelos Veus foram escolhidos, e numerados para uma fácil referenciação:

```
1. <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL> --->
[<LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL>]:[Correcto]

2.<ABSTRACCAO TIPO="NOME" MORF="F,S">Escola Normal Livre de Agudos</ABSTRACCAO> --->
[<ORGANIZACAO TIPO="INSTITUICAO" MORF="F,S">Escola Normal Livre</ORGANIZACAO>]:
[Parcialmente_Correcto_por_Defeito(0.3; 0.7)]

3.<LOCAL TIPO="ALARGADO" MORF="M,S">Hotel Lisboa Plaza</LOCAL> --->
[<LOCAL TIPO="ALARGADO" MORF="M,S">Hotel Lisboa</LOCAL>, <PESSOA TIPO="INDIVIDUAL"
MORF="M,S">Plaza</PESSOA>]:[Parcialmente_Correcto_por_Defeito(0.3333333333333333;
0.6666666666666667), Parcialmente_Correcto_por_Defeito(0.16666666666666666; 0.8333333333333334)]

4.<COISA|COISA TIPO="CLASSE|OBJECTO" MORF="?,?">BATTENFELD</COISA|COISA> --->
[<COISA TIPO="CLASSE" MORF="M,S">BATTENFELD</COISA>]:[Correcto]

5.<ORGANIZACAO TIPO="SUB" MORF="F,S">Reportagem Local</ORGANIZACAO> --->
[<OBRA TIPO="REPRODUZIDA" MORF="?,?">a Reportagem Local</OBRA>]:
[Parcialmente_Correcto_por_Excesso(0.3333333333333333; 0.6666666666666667)]

6.<OBRA TIPO="PRODUTO" MORF="?,S">The Artic</OBRA> --->
[<OBRA TIPO="REPRODUZIDA" MORF="?,?">The Artic</OBRA>]:[Correcto]

7.<ESPURIO>História</ESPURIO> ---> [<ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">História</ABSTRACCAO>]:[Espúrio]

8.<LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">Pinheiros</LOCAL> ---> [null]:[Em_Falta]

9.<LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Brasil</LOCAL> --->
[<LOCAL TIPO="ADMINISTRATIVO">o Brasil</LOCAL>]:[Parcialmente_Correcto_por_Excesso(0.25; 0.75)]

10. <ORGANIZACAO TIPO="SUB" MORF="F,P">Relações Públicas</ORGANIZACAO> --->
[<ABSTRACCAO TIPO="DISCIPLINA" MORF="M,S">Relações Públicas</ABSTRACCAO>]:[Correcto]

11. <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Próximo Oriente</LOCAL> --->
[<ORGANIZACAO TIPO="INSTITUICAO" MORF="?,?">Próximo Oriente</ORGANIZACAO>]:[Correcto]
```

Após o processamento pelo Vizir (cenário absoluto) o resultado obtido é o seguinte:

```
1.<EM MORF="M,S">Portugal</EM> ---> [<EM MORF="M,S">Portugal</EM>]:
[(Género: Correcto 1) (Número: Correcto 1) (Combinada: Correcto 1)]
```

Comentário: Este caso está classificado como morfologicamente correcto,
o sistema também identificou correctamente a EM.

2.<EM MORF="F,S">Escola Normal Livre de Agudos</EM> ---> [<EM MORF="F,S">Escola Normal Livre</EM>]:
[(Género: Parcialmente Correcto 0.5) (Número: Parcialmente Correcto 0.5) (Combinada: Parcialmente Correcto 0.5)]
Comentário: Este caso está classificado como morfologicamente correcto.
Como a EM foi classificada como parcialmente correcta na identificação, foi
atribuída a pontuação parcialmente correcta para este caso.

3.<EM MORF="M,S">Hotel Lisboa Plaza</EM> ---> [<EM MORF="M,S">Hotel Lisboa</EM>, <EM MORF="F,S">Plaza</EM>]:
[(Género: Parcialmente Correcto 0.5) (Número: Parcialmente Correcto 0.5) (Combinada: Parcialmente Correcto 0.5)]
Comentário: Para a avaliação da classificação deste alinhamento, apenas conta a EM submetida pelo sistema,
cujo primeiro átomo (palavra Hotel) alinha com o primeiro átomo da EM na CD. Sendo assim,
apenas a 1ª EM é considerada, a 2ª EM não é considerada posteriormente para o total de EM do sistema.

4.<EM MORF="?,?">BATTENFELD</EM> ---> [<EM MORF="M,S">BATTENFELD</EM>]:
[(Género: Sobre especificado 0) (Número: Sobre especificado 0) (Combinada: Incorrecto 0)]
Comentário: Neste caso, a EM na CD não foi classificada morfologicamente. No entanto o
sistema classificou morfologicamente a EM, e sobre-especificou a classificação da EM.
Notar é atribuído para a pontuação combinada, o valor incorrecto.

5.<EM MORF="F,S">Reportagem Local</EM> ---> [<EM MORF="?,?">a Reportagem Local</EM>]:
[(Género: Em Falta 0) (Número: Em Falta 0) (Combinada: Em Falta 0)]
Comentário: O primeiro átomo da EM do sistema não combina com a EM da CD.

6.<EM MORF="?,S">The Artic</EM> ---> [<EM MORF="?,?">The Artic</EM>]:
[(Género: Correcto 1) (Número: Em Falta 0) (Combinada: Em Falta 0)]
Comentário: Tal como na CD, o sistema não foi chegou a nenhuma conclusão relativamente ao
género da EM na CD. Mas também não classificou a EM em relação ao número, como na CD está
classificado o número como singular, o sistema não classificou a EM em relação ao número.

7.<ESPURIO>História</ESPURIO> ---> [<EM MORF="F,S">História</EM>]:
[(Género: Espúrio 0) (Número: Espúrio 0) (Combinada: Espúrio 0)]
Comentário: O sistema classificou morfologicamente como uma EM, que não foi identificada como
sendo uma EM na CD. No cenário relativo este caso não seria avaliado, sendo descartado dos
resultados. Para o cenário absoluto, o sistema obtém a pontuação de espúrio para todos os campos.

8.<EM MORF="?,S">Pinheiros</EM> ---> [null]:
[(Género: Em Falta 0) (Número: Em Falta 0) (Combinada: Em Falta 0)]
Comentário: O sistema falhou em identificar a EM. Este caso não é contabilizado para o número
total de EM classificadas pelo sistema.

9.<EM MORF="M,S">Brasil</EM> ---> [<EM>o Brasil</EM>]:
[(Género: Em Falta 0) (Número: Em Falta 0) (Combinada: Em Falta 0)]
Comentário: O sistema não classificou morfologicamente a EM.
Este caso não é contabilizado pelo ida2mor para o número total de EM classificadas pelo sistema.

10.<EM MORF="F,P">Relações Públicas</EM> ---> [<EM MORF="M,S">Relações Públicas</EM>]:
[(Género: Incorrecto 0) (Número: Incorrecto 0) (Combinada: Incorrecto 0)]
Comentário: O sistema falhou em correctamente classificar morfologicamente a EM.

11.<EM MORF="M,S">Próximo Oriente</EM> ---> [<EM MORF="?,?">Próximo Oriente</EM>]:
[(Género: Em Falta 0) (Número: Em Falta 0) (Combinada: Em Falta 0)]
Comentário: O sistema falhou em determinar a classificação morfológica da EM.
Este caso é contabilizado pelo ida2mor para o número total de EM classificadas
pelo sistema.

Referências

(Afonso, 2006) Susana Afonso. Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica. 12 de Fevereiro de 2006. `http://www.linguateca.pt/Floresta/ArvoresDeitadas.doc`.

(Afonso et al., 2002) Susana Afonso, Eckhard Bick, Renato Haber e Diana Santos. Floresta sintá(c)tica: um treebank para o português. Em Anabela Gonçalves e Clara Nunes Correia, editores, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística*, APL 2001. Lisboa, Portugal. 2-4 de Outubro de 2002. p. 533–545.

(Agichtein e Gravano, 2000) Eugene Agichtein e Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. Em *Proceedings of the Fifth ACM Conference on Digital Libraries*. San Antonio, TX, EUA. 2-7 de Junho de 2000. p. 85–94.

(Almeida e Pinto, 1995) José João Almeida e Ulisses Pinto. Jspell – um módulo para análise léxica genérica de linguagem natural. Em *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*. Évora, Portugal. 6-8 de Outubro de 1995. p. 1–15.

(Almeida e Simões, 2006a) José João Almeida e Alberto Manuel Simões. Publishing multilingual ontologies: a quick way of obtaining feedback. Em Bob Martens e Milena Dobreva, editores, *Digital spectrum : integrating technology and culture : proceedings of the International Conference on Electronic Publishing*, ELPUB2006. Bansko, Bulgária. Junho de 2006. p. 373–374.

(Almeida e Simões, 2006b) José João Almeida e Alberto Manuel Simões. T2O - Recycling Thesauri into a Multilingual Ontology. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'2006. Génova, Itália. 22-28 de Maio de 2006. p. 1466–1471.

(Alves e Almeida, 2006) Edgar Alves e José João Almeida. Manual de utilizador do RENA. Relatório técnico. Universidade do Minho, Departamento de Informática. Julho de 2006.

(Amitay et al., 2004) Einat Amitay, Nadav Har'El, Ron Sivan e Aya Soffer. Web-a-Where: Geotagging Web content. Em Mark Sanderson, Kalervo Järvelin, James Allan e Peter Bruza, editores, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04. Sheffield, Reino Unido. 25-29 de Julho de 2004. p. 273–280.

(Appelt et al., 1995) Douglas E. Appelt, Jerry R. Hobbs, John Bear, David Israel, Megumi Kameyama, Andy Kehler, David Martin, Karen Myers e Marby Tyson. SRI International FASTUS system MUC-6 test results and analysis. Em *Proceedings of the 6th Message Understanding Conference*, MUC-6. Columbia, MD, EUA. 6-8 de Novembro de 1995. p. 237–248.

(Arévalo et al., 2002) Montserrat Arévalo, Xavier Carreras, Lluís Màrquez, Toni Martí, Lluís Padró e Maria José Simon. A proposal for wide-coverage Spanish named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*. 28:63–80. Maio de 2002.

(Baptista et al., 2006) Jorge Baptista, Fernando Batista, Nuno Mamede e Cristina Mota. Npro: um novo recurso para o processamento computacional do português. Em Joaquim Barbosa e Fátima Oliveira, editores, *Textos seleccionados do XXI Encontro da Associação Portuguesa de Linguística*. 2006.

(Ben-Kiki et al., 2005) Oren Ben-Kiki, Clark Evans e Brian Ingerson. YAML specification. 2005. `http://yaml.org/spec/`.

(Ben-Kiki et al., 2006) Oren Ben-Kiki, Clark Evans e Brian Ingerson. YAML cookbook. 2006. `http://yaml4r.sourceforge.net/cookbook/`.

(Bick, 2000) Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Tese de doutoramento. Universidade de Aarhus. Aarhus University Press. Novembro de 2000.

(Bick, 2003) Eckhard Bick. Multi-level NER for Portuguese in a CG framework. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*. Springer. Berlin/Heidelberg. 2003. p. 118–125.

(Bick, 2004) Eckhard Bick. A Named Entity Recognizer for Danish. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation*. Lisboa, Portugal. 26-28 de Maio de 2004. p. 305–308.

(Bick, 2006a) Eckhard Bick. Functional Aspects in Portuguese NER. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oli-

veira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*. Springer. Berlin/Heidelberg. 2006. p. 80–89.

(Bick, 2006b) Eckhard Bick. Functional Aspects on Portuguese NER. Encontro do HAREM. Porto, Portugal. Apresentação. 15 de Julho de 2006. `http://www.linguateca.pt/documentos/HAREM2006Bick.pdf`.

(Bick et al., 2007) Eckhard Bick, Diana Santos, Susana Afonso e Rachel Marchi. Floresta Sintá(c)tica: Ficção ou realidade? Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. Lisboa, Portugal. 20 de Março de 2007. p. 291–300.

(Bikel et al., 1997) Daniel M. Bikel, Scott Miller, Richard Schwartz e Ralph Weischedel. Nymble: a high performance learning name-finder. Em *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLP'97. Washington DC, DC, EUA. 31 de Março a 3 de Abril de 1997. p. 194–201.

(Bikel et al., 1999) Daniel M. Bikel, Richard Schwartz e Ralph Weischedel. An algorithm that learns what's in a name. *Machine Learning*. 34(1-3):211–231. Fevereiro de 1999.

(Black et al., 1998) William J. Black, Fabio Rinaldi e David Mowatt. FACILE: Description of the NE system used for MUC-7. Em *Proceedings of the 7th Message Understanding Conference*, MUC-7. Fairfax,VI, EUA. 29 de Abril a 1 de Maio de 1998.

(Blume, 2005) Matthias Blume. Automatic Entity Disambiguation: Benefits to NER, Relation Extraction, Link Analysis, and Inference. Em *Proceedings of 2005 International Conference on Intelligence Analysis*. McLean, VA, EUA. 2-4 de Maio de 2005.

(Bontcheva et al., 2002) Kalina Bontcheva, Hamish Cunningham, Valentin Tablan, Diana Maynard e Oana Hamza. Using GATE as an Environment for Teaching NLP. Em *Proceedings of the ACL'02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Filadélfia, PA, EUA. Julho de 2002.

(Borthwick, 1999) Andrew Borthwick. *A Maximum Entropy Approach to Named Entity Recognition*. Tese de doutoramento. Universidade de Nova Iorque, EUA. Setembro de 1999.

(Borthwick et al., 1998) Andrew Borthwick, John Sterling, Eugene Agichtein e Ralph Grishman. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. Em *Proceedings of the 6th Workshop on Very Large Corpora, WVLC-98*. Montreal, Quebec, Canadá. 15-16 de Agosto de 1998.

(Brin, 1998) Sergey Brin. Extracting Patterns and Relations from the World Wide Web. Em *Workshop on the Web and Database*, WebDB'98. Valência, Espanha. 27-28 de Março de 1998. p. 172–183.

(Buckley e Voorhees, 2000) Chris Buckley e Ellen M. Voorhees. Evaluating evaluation measure stability. Em Nicholas J. Belkin, Peter Ingwersen e Mun-Kew Leong, editores, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2000. Atenas, Grécia. 24-28 de Julho de 2000. p. 33–40.

(Buitelaar, 1998) Paul Buitelaar. CoreLex: An Ontology of Systematic Polysemous Classes. Em *Proceedings of International Conference on Formal Ontology in Information Systems, FOIS'98*. Trento, Itália. 6-8 de Junho de 1998.

(Burns, 1991) Linda Claire Burns. *Vagueness: An Investigation into Natural Languages and the Sorites Paradox*. Kluwer Academic Publishers. Dordrecht. 1991.

(Cardoso, 2006a) Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Tese de mestrado. Faculdade de Engenharia da Universidade do Porto. Outubro de 2006. Republicado como DI/FCUL TR-06-26, Departamento de Informática, Universidade de Lisboa, Novembro 2006.

(Cardoso, 2006b) Nuno Cardoso. HAREM e MiniHAREM: Uma análise comparativa. Encontro do HAREM. Porto, Portugal. Apresentação. 15 de Julho de 2006. `http://www.linguateca.pt/documentos/encontroHAREM_cardoso.pdf`.

(Carreras e Padró, 2002) Xavier Carreras e Lluís Padró. A flexible distributed architecture for natural language analyzers. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Espanha. 29-31 de Maio de 2002. p. 1813–1817.

(Carreras et al., 2002) Xavier Carreras, Lluís Màrques e Lluís Padró. Named Entity Extraction using AdaBoost. Em Dan Roth e Antal van den Bosch, editores, *Proceedings of CoNLL-2002, the 6th Conference on Natural Language Learning*. Taipé, Formosa. 31 de Agosto a 1 de Setembro de 2002. p. 167–170.

(Carreras et al., 2003a) Xavier Carreras, Lluís Màrquez e Lluís Padró. Named entity recognition for Catalan using only Spanish resources and unlabelled data. Em *10th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'03. Budapeste, Hungria. Abril de 2003. p. 43–50.

(Carreras et al., 2003b) Xavier Carreras, Lluís Màrquez e Lluís Padró. A Simple Named Entity Extractor using AdaBooMst. Em Walter Daelemans, Miles Osborne, Walter Daelemans e Miles Osborne, editores, *Proceedings of the Conference on Computational Natural Language Learning*, CoNLL-2003. Edmonton, Canadá. 31 de Maio a 1 de Junho de 2003. p. 152–155.

(Chaves et al., 2005) Marcirio Silveira Chaves, Mário J. Silva e Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. Em C. A. Heuser, editor, *Proceedings do 20º Simpósio Brasileiro de Banco de Dados*, SBBD. Uberlândia, MG, Brasil. 3-7 de Outubro de 2005. p. 40–54.

(Chinchor, 1992) Nancy Chinchor. The Statistical Significance of MUC-4 Results. Em *Proceedings of the 4th Conference on Message Understanding*, MUC-4. McLean, VI, EUA. 16-18 de Junho de 1992. p. 30–50.

(Chinchor, 1995) Nancy Chinchor. Statistical significance of MUC-6 results. Em *Proceedings of the 6th Message Understanding Conference*, MUC-6. Columbia, MD, EUA. 6-8 de Novembro de 1995. p. 39–43.

(Chinchor, 1998a) Nancy Chinchor. Statistical Significance of MUC-7 Results. Em *Proceedings of the 7th Message Understanding Conference*, MUC-7. Fairfax,VI, EUA. 29 de Abril a 1 de Maio de 1998.

(Chinchor e Marsh, 1998) Nancy Chinchor e Elaine Marsh. MUC-7 Named Entity Task Definition (version 3.5). Em *Proceedings of the 7th Message Understanding Conference*, MUC-7. Fairfax,VI, EUA. 29 de Abril a 1 de Maio de 1998.

(Chinchor, 1998b) Nancy A. Chinchor. Overview of MUC-7/MET-2. Em *Proceedings of the 7th Message Understanding Conference*, MUC-7. Fairfax,VI, EUA. 29 de Abril a 1 de Maio de 1998.

(Christ et al., 1999) Oliver Christ, Bruno M. Schulze, Anja Hofmann e Esther Koenig. The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual (CQP V2.2). Relatório técnico. Universidade de Estugarda. 16 de Agosto de 1999. `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/`.

(Cohen e Sarawagi, 2004) William W. Cohen e Sunita Sarawagi. Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. Em *Proceedings of KDD-04, the 10th International Conference on Knowledge Discovery and Data Mining*. Seattle, WA, EUA. 22-25 de Agosto de 2004.

(Costa et al., 2007) Luís Costa, Paulo Rocha e Diana Santos. Organização e resultados morfolímpicos. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. Lisboa, Portugal. 20 de Março de 2007. p. 15–33.

(Cruse, 2004) Alan Cruse. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press. Oxford. 2004.

(Cunningham, 2005) Hamish Cunningham. Information Extraction, Automatic. Em *Encyclopedia of Language and Linguistics*. Elsevier. 2ª edição. 2005. p. 665–677.

(Cunningham et al., 2002) Hamish Cunningham, Diana Maynard, Kalina Bontcheva e Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. Em *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02. Filadélfia, PA, EUA. Julho de 2002.

(Curran e Clark, 2003) James R. Curran e Stephen Clark. Language independent NER using a maximum entropy tagger. Em Walter Daelemans e Miles Osborne, editores, *Proceedings of the Conference on Computational Natural Language Learning*, CoNLL-2003. Edmonton, Canadá. 31 de Maio a 1 de Junho de 2003. p. 164–167.

(Daelemans et al., 2003) Walter Daelemans, Jakub Zavrel, Ko van der Sloot e Antal van den Bosch. TiMBL: Tilburg Memory-Based Learner. Relatório Técnico ILK 03-10. Universidade de Tilburg. 2003.

(Day et al., 1997) David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok e Patricia Robinson. Mixed-Initiative Development of Language Processing Systems. Em *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLP'97. Washington DC, DC, EUA. 31 de Março a 3 de Abril de 1997. p. 88–95.

(Delboni, 2005) Tiago M. Delboni. *Expressões de posicionamento como fonte de contexto geográfico na Web*. Tese de doutoramento. Universidade Federal de Minas Gerais. 2005.

(Densham e Reid, 2003) Ian Densham e James Reid. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. Em *Proceedings of the HLT-NAACL 2003 Workshop on the Analysis of Geographic References*. Edmonton, Canadá. 27 de Maio a 1 de Junho de 2003.

(Dietterich, 2000) Thomas G. Dietterich. Ensemble methods in machine learning. Em J. Kittler e F. Roli, editores, *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21-23, 2000 Proceedings*. Springer. Nova Iorque, NY, EUA. 2000. p. 1–15.

(Doddington et al., 2004) George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel e Ralph Weischedel. The Automatic Content Extraction (ACE) Program. Tasks, Data and Evaluation. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation*. Lisboa, Portugal. 26-28 de Maio de 2004. p. 837–840.

(Douthat, 1998) Aaron Douthat. The Message Understanding Conference Scoring Software User's Manual. Em *Proceedings of the 7th Message Understanding Conference*, MUC-7. Fairfax,VI, EUA. 29 de Abril a 1 de Maio de 1998.

(Efron, 1981) Bradley Efron. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. *Biometrika*. 81(3):589–599. Dezembro de 1981.

(Ellis, 1993) John M. Ellis. *Language, Thought and Logic*. Northwestern University Press. Evanston, IL, EUA. 1993.

(Evert, 2005) Stefan Evert. The CQP Query Language Tutorial (CWB version 2.2.b90). Relatório técnico. Universidade de Estugarda. 10 de Julho de 2005.

(Fairon, 1999) Cédrick Fairon. Parsing a Web site as a corpus. Em Cédrick Fairon, editor, *Analyse lexicale et syntaxique: Le système INTEX*. John Benjamins Publishing. Amsterdão, Países Baixos. 1999. p. 327–340.

(Ferrández et al., 2005) Óscar Ferrández, Zornitsa Kozareva, Andrés Montoyo e Rafael Muñoz. NERUA: sistema de detección y clasificación de entidades utilizando aprendizaje automático. *Sociedad Española para el Procesamiento del Lenguaje Natural*. 35:37–44. 2005.

(Ferrández et al., 2006) Óscar Ferrández, Antonio Toral e Rafael Muñoz. Fine tuning features and post-processing rules to improve Named Entity Recognition. Em Christian Kop, Günther Fliedl, Heinrich C. Mayr e Elisabeth Métais, editores, *Processing and Information Systems, 11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006, Klagenfurt, Austria, May 31 - June 2, 2006, Proceeding*. Springer. Berlin/Heidelberg. 2006. p. 176–185.

(Fillmore, 1968) Charles J. Fillmore. The case for case. Em Emmon Bach e Robert T. Harms, editores, *Universals in Linguistic Theory*. Holt, Rinehart and Winston. Londres. 1968. p. 1–88.

(Florian et al., 2003) Radu Florian, Abe Ittycheriah, Hongyan Jing e Tong Zhang. Named Entity Recognition through Classifier Combination. Em Walter Daelemans e Miles Osborne, editores, *Proceedings of the Conference on Computational Natural Language Learning*, CoNLL-2003. Edmonton, Canadá. 31 de Maio a 1 de Junho de 2003. p. 168–171.

(Frankenberg-Garcia e Santos, 2002) Ana Frankenberg-Garcia e Diana Santos. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução*. IX(1):61–79. 2002.

(Friburger, 2002) Nathalie Friburger. *Reconnaissance automatique des noms propres. Application à la classification automatique de textes journalistiques*. Tese de doutoramento. Universidade François Rabelais, Tours, França. 2 de Dezembro de 2002.

(Gale et al., 1992) William A. Gale, Kenneth W. Church e David Yarowsky. One Sense Per Discourse. Em *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. Harriman, NY, EUA. 23-26 de Fevereiro de 1992. p. 233–237.

(Gey et al., 2006) Frederic Gey, Ray Larson, Mark Sanderson, Hideo Joho e Paul Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. Em Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini e Maarten de Rijke, editores, *Acessing Multilingual information Repositories:6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Vienna, Austria, September 2005. Revised Selected papers*. Springer. Berlin/Heidelberg. 2006. p. 908–919.

(Gey et al., 2007) Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio M. Di Nunzio e Nicola Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*. Springer. Berlin / Heidelberg. 2007. p. 852–876.

(Ginsberg, 1987) Matthew Ginsberg. *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann. Los Altos, CA, EUA. 1987.

(Gomes e Silva, 2006) Daniel Gomes e Mário J. Silva. Modelling Information Persistence on the Web. Em *Proceedings of the 6th International Conference on Web Engineering, ICWE 2006*. Palo Alto, CA, EUA. 11-14 de Julho de 2006. p. 193–200.

(Good, 2000) Philip Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer. Nova Iorque, NY, EUA. 2ª edição. 2000.

(Grishman e Sundheim, 1995) Ralph Grishman e Beth Sundheim. Design of the MUC-6 Evaluation. Em *Proceedings of the 6th Message Understanding Conference*, MUC-6. Columbia, MD, EUA. 6-8 de Novembro de 1995. p. 413–422.

(Grishman e Sundheim, 1996) Ralph Grishman e Beth Sundheim. Message Understaning Conference 6: A Brief History. Em *Proceedings of the 16th International Conference on Computational Linguistics, COLING 96*. Copenhaga, Dinamarca. 5-9 de Agosto de 1996. p. 466–471.

(Gross, 1975) Maurice Gross. *Méthodes en Syntaxe - Régime des constructions complétives*. Hermann. Paris, França. 1975.

(Guthrie et al., 2004) Louise Guthrie, Roberto Basili, Eva Hajicova e Frederick Jelinek, editores. *Workshop proceedings of LREC 2004: Beyond Entity Recognition - Semantic Labelling for NLP Tasks*. ELRA. Lisboa, Portugal. 25 de Maio de 2004.

(Harman, 1993) Donna Harman. Overview of the First TREC Conference. Em Robert Korfhage, Edie Rasmussen e Peter Willett, editores, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 93*. Pittsburgh, PA, EUA. 27 de Junho a 1 de Julho de 1993. p. 36–47.

(Harpring, 1997) Patricia Harpring. Proper words in proper places: The thesaurus of geographic names. *MDA Information*. 2(3):5–12. 1997.

(Hill, 2000) Linda L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. Em José Luis Borbinha e Thomas Bake, editores, *Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000, Proceedings*, ECDL-00. Springer. Berlin/Heidelberg. 2000. p. 280–291.

(Hill et al., 1999) Linda L. Hill, James Frew e Qi Zheng. Geographic names: the implementation of a gazetteer in a georeferenced digital library. *D-Lib Magazine*. 5(1). Janeiro de 1999.

(Hirschman, 1998) Lynette Hirschman. The evolution of evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*. 12(4):281–305. 1998.

(Holte et al., 1989) Robert C. Holte, Liane Acker e Bruce W. Porter. Concept Learning and the Problem of Small Disjuncts. Em N. S. Sridharan, editor, *Proceedings of the Eleventh Joint International Conference on Artificial Intelligence*, IJCAI. Detroit, MI, EUA. Agosto de 1989. p. 813–818.

(Hughes e Cresswell, 1968) George E. Hughes e Maxwell J. Cresswell. *An Introduction to Modal Logic*. Methuen & Co., Ltd. Londres. 1968.

(Inácio e Santos, 2006) Susana Inácio e Diana Santos. Syntactical Annotation of COMPARA: Workflow and First Results. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*. Springer. Berlin/Heidelberg. 2006. p. 256–259.

(Japkowicz, 2003) Nathalie Japkowicz. Class Imbalances: Are we Focusing on the Right Issue? Em Nitesh Chawla, Nathalie Japkowicz e Aleksander Kolcz, editores, *Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II*. Washington DC, DC, EUA. Agosto de 2003.

(Joachims, 1999) Thorsten Joachims. Transductive inference for text classification using support vector machines. Em Saso Dzeroski e Ivan Bratko, editores, *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML 1999. Bled, Eslovénia. Junho de 1999. p. 200–209.

(Joachims, 2002) Thorsten Joachims. *Learning to Classify Text using Support Vector Machines: Methods Theory and Algorithms*. Kluwer Academic Publishers. Norwell, MA, EUA. Maio de 2002.

(Johannessen et al., 2005) Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick e Dorte Haltrup. Named Entity Recognition for the Mainland Scandinavian Languages. *Literary and Linguistic Computing*. 20(1):91–102. 2005.

(Jones et al., 2004) Cristopher B. Jones, Alia I. Abdelmoty, David Finch, Gaihua Fu e Subodh Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Em Max J. Egenhofer, Christian Freska e Harvey Miller, editores, *Geographic Information Science, Third International Conference, GIScience 2004, Adelphi, MD, USA, October 20-23, 2004, Proceedings*. Springer. Berlin/Heidelberg. 2004. p. 125–139.

(Jones e Bates, 1977) Karen Sparck Jones e R. G. Bates. Research on Automatic Indexing 1974-1976. Relatório técnico. Computer Laboratory, University of Cambridge. 1977.

(Kamp e Reyle, 1993) Hans Kamp e Uwe Reyle. *From Discourse to Logic: an Introduction to Model Theoretic Semantics of Natural language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Press. Dordrecht. 1993.

(Koehn, 2004) Philip Koehn. Statistical significance tests for machine translation evaluation. Em Dekang Lin e Dekai Wu, editores, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2004. Barcelona, Espanha. Julho de 2004. p. 388–395.

(Kohler, 2003) Janet Kohler. *Analysing search engine queries for the use of geographic terms*. Tese de doutoramento. Universidade de Sheffield. 2003.

(Kornai e Sundheim, 2003) Andras Kornai e Beth Sundheim, editores. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*. Association for Computational Linguistics. Morristown, NJ, EUA. 27 de Maio a 1 de Junho de 2003.

(Kozareva et al., 2007) Zornitsa Kozareva, Óscar Ferrández, Andrés Montoyo, Rafael Muñoz, Armando Suárez e Jaime Gómez. Combining data-driven systems for improving Named Entity Recognition. *Data & Knowledge Engineering*. 61(3):449–466. 2007.

(Krupka e Hausman, 1998) George R. Krupka e Kevin Hausman. IsoQuest Inc.: Description of the NetOwlTM extractor system as used for MUC-7. Em *Proceedings of the 7th Message Understanding Conference*, MUC-7. Fairfax,VI, EUA. 29 de Abril a 1 de Maio de 1998.

(Lakoff, 1987) George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press. Chicago & Londres. 1987.

(Lakoff e Johnson, 1980) George Lakoff e Mark Johnson. *Metaphors We Live By*. University of Chicago Press. Chicago & Londres. 1980.

(Lansing, 2001) Jeff Lansing. Geoparser Service Specification 0.71. Relatório Técnico OGC-01-035. Open Geospatial Consortium. Março de 2001.

(Leidner, 2004) Jochen L. Leidner. Towards a Reference Corpus for Automatic Toponym Resolution Evaluation. Em *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference*, GIR 2004. Sheffield, Reino Unido. 25-29 de Julho de 2004.

(Leidner et al., 2003) Jochen L. Leidner, Gail Sinclair e Bonnie Webber. Grounding Spatial Named Entities for Information Extraction and Question Answering. Em *Proceedings of the HLT-NAACL 2003 Workshop on the Analysis of Geographic References*. Edmonton, Canadá. 27 de Maio a 1 de Junho de 2003. p. 31–38.

(Leveling e Hartrumpf, 2006) Johannes Leveling e Sven Hartrumpf. On metonymy recognition for Geographic IR. Em *Proceedings of the Workshop on Geographic Information Retrieval held at the 29th Annual International ACM SIGIR Conference*, GIR 2006. Seattle, WA, EUA. Agosto de 2006.

(Leveling e Veiel, 2006) Johannes Leveling e Dirk Veiel. University of Hagen at GeoCLEF 2006: Experiments with metonymy recognition in documents. Em Alessandro Nardi, Carol Peters e José Luís Vicedo, editores, *Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop*. 2006. s/pp.

(Li et al., 2002) Huifeng Li, Rohini Srihari, Cheng Niu e Wei Li. Location normalization for information extraction. Em *Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002*. Taipé, Formosa. 24 de Agosto a 1 de Setembro de 2002. p. 1–7.

(Li, 1992) Wentian Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*. 38(6):1842–1845. 1992.

(Lin e Hauptmann, 2005) Wei-Hao Lin e Alexander Hauptmann. Revisiting the Effect of Topic Set Size on Retrieval Experiment Error. Em Ricardo A. Baeza-Yates, Nivio Ziviani,

Gary Marchionini, Alistair Moffat e John Tait, editores, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005*. Salvador, Brasil. 15-19 de Agosto de 2005. p. 637–638.

(Ling e Li, 1998) Charles X. Ling e Chenghui Li. Data mining for direct marketing: Problems and solutions. Em Rakesh Agrawal, Paul E. Stolorz e Gregory Piatetsky-Shapiro, editores, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD-98. Nova Iorque, NY, EUA. 31 de Agosto de 1998. p. 73–79.

(Madrigal et al., 2003) Víctor J. Díaz Madrigal, José Troyano e Fernando Enríquez. Aplicacion de Modelos de Markov y Maquinas SVM al Reconocimiento de Entidades. Em *Actas de las X Jornadas de la CAEPIA y V de la TTIA*, CAEPIA'2003. San Sebastián, Espanha. 11-14 de Novembro de 2003. p. 55–58.

(Malouf, 2002) Robert Malouf. Markov models for language-independent named entity recognition. Em Dan Roth e Antal van den Bosch, editores, *Proceedings of CoNLL-2002, the 6th Conference on Natural Language Learning*. Taipé, Formosa. 31 de Agosto a 1 de Setembro de 2002. p. 187–190.

(Mamede et al., 2003) Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores. *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*. Springer. Berlin/Heidelberg. 2003.

(Mandl et al., 2007) Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker e Xing Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. Em Alessandro Nardi e Carol Peters, editores, *Working Notes for the CLEF 2007 Workshop*. 2007. s/pp.

(Manov et al., 2003) Dimitar Manov, Atanas Kiryakov, Borislav Popov, Kalina Bontcheva, Diana Maynard e Hamish Cunningham. Experiments with geographic knowledge for information extraction. Em *Proceedings of the HLT-NAACL 2003 Workshop on the Analysis of Geographic References*. Edmonton, Canadá. 27 de Maio a 1 de Junho de 2003.

(Marcelino, 2005) Isabel Marcelino. Documentação do ELLE. 29 de Setembro de 2005. `http://www.linguateca.pt/Equipa/isabel/Documentacao_ELLE.pdf`.

(Markert e Nissim, 2002) Katja Markert e Malvina Nissim. Towards a corpus annotated for metonymies: the case of location names. Em Manuel González Rodríguez, Carmen Paz Suárez Araujo, Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Espanha. 29-31 de Maio de 2002. p. 1385–1392.

(Martins et al., 2005) Bruno Martins, Mário J. Silva e Marcirio Silveira Chaves. Challenges and Resources for Evaluating Geographical IR. Em *Proceedings of the 2005 Workshop On Geographic Information Retrieval, GIR 2005, Bremen, Germany, November 4, 2005*. Bremen, Alemanha. Outubro de 2005. p. 31–34.

(Martins et al., 2006) Bruno Martins, Marcirio Chaves e Mario J. Silva. O sistema CaGE para Reconhecimento de referências geográficas em textos na língua portuguesa. Encontro do HAREM. Porto, Portugal. Apresentação. 15 de Julho de 2006. `http://www.linguateca.pt/documentos/CaGEHAREM.ppt`.

(Martins et al., 2007) Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade e Mário J. Silva. The University of Lisbon at GeoCLEF 2006. Em Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke e Maximilian Stempfhuber, editores, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*. Springer. Berlin / Heidelberg. 2007. p. 986–994.

(Maynard et al., 2002) Diana Maynard, Hamish Cunningham, Kalina Bontcheva e Marin Dimitrov. Adapting a Robust Multi-genre NE System for Automatic Content Extraction. Em D. R. Scott, editor, *Artificial Intelligence: Methodology, Systems, and Applications, 10th International Conference, AIMSA 2002, Varna, Bulgaria, September 4-6, 2002, Proceedings*. Springer. Berlin/Heidelberg. 2002. p. 264–273.

(Maynard et al., 2003a) Diana Maynard, Kalina Bontcheva e Hamish Cunningham. Towards a semantic extraction of named entities. Em *Recent Advances in Natural Language Processing, RANLP*. Borovets, Bulgária. 10-12 de Setembro de 2003.

(Maynard et al., 2003b) Diana Maynard, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham e Yorick Wilks. MUSE: a MUlti-Source Entity recognition system. Em apreciação pela revista Computers and the Humanities. 2003. `http://gate.ac.uk/sale/muse/muse.pdf`.

(McDonald, 1996) David D. McDonald. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. Em Branimir Boguraev e James Pustejovsky, editores, *Corpus Processing for Lexical Acquisition*. MIT Press. Cambridge, MA & Londres. 1996. p. 21–39.

(Merchant et al., 1996) Roberta Merchant, Mary Ellen Okurowski e Nancy Chinchor. The Multilingual Entity Task (MET) Overview. Em *Proceedings of TIPSTER Text Program (Phase II)*. Vienna, VI, EUA. Maio de 1996. p. 449–451.

(Mikheev et al., 1999) Andrei Mikheev, Marc Moens e Claire Grover. Named Entity Recognition without Gazetteers. Em *Proceedings of EACL'99: Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, Noruega. 8-12 de Junho de 1999. p. 1–8.

(Mitsumori et al., 2004) Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi e Hirohumi Doi. Boundary correction of protein names adapting heuristic rules. Em Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Fifth International Conference*, CICLing 2004. Springer. Berlin/Heidelberg. 2004. p. 172–175.

(Moore et al., 2002) David S. Moore, George P. McCabe, William M. Duckworth e Stanley L. Sclove. *The Practice of Bussiness Statistics: Using Data for Decisions*. W. H. Freeman. Novembro de 2002.

(Morgan, 2006) William Morgan. Statistical Hypothesis Tests for NLP. 16 de Fevereiro de 2006. `http://nlp.stanford.edu/local/talks/sigtest.pdf`.

(Mota e Moura, 2003) Cristina Mota e Pedro Moura. ANELL: A Web System for Portuguese Corpora Annotation. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*. Springer. Berlin/Heidelberg. 2003. p. 184–188.

(Mota et al., 2007) Cristina Mota, Diana Santos e Elisabete Ranchhod. Avaliação de reconhecimento de entidades mencionadas: princípio de AREM. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. Lisboa, Portugal. 20 de Março de 2007. p. 161–176.

(Nissim et al., 2004) Malvina Nissim, Colin Matheson e James Reid. Recognising Geographical Entities in Scottish Historical Documents. Em *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference*, GIR 2004. Sheffield, Reino Unido. 25-29 de Julho de 2004.

(Noreen, 1989) Eric W. Noreen. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley & Sons. Nova Iorque, NY, EUA. 1989.

(Oliveira e Ribeiro, 2003) José N. Oliveira e Óscar Ribeiro. Knowledge renovator - Requirements Specification. Relatório técnico. Universidade do Minho, Departamento de Informática. 2003. IKF-P partner - IKF (E!2235).

(Olligschlaeger e Hauptmann, 1999) Andreas M. Olligschlaeger e Alexander G. Hauptmann. Multimodal Information Systems and GIS: The Informedia Digital Video Library. Em *Proceedings of the 1999 ESRI User Conference*. San Diego, CA, EUA. 26-30 de Julho de 1999.

(Palmer e Day, 1997) David D. Palmer e David S. Day. A Statistical Profile of the Named Entity Task. Em *Proceedings of the Fifth ACL Conference for Applied Natural Language Processing*, ANLP'97. Washington DC, DC, EUA. Abril de 1997. p. 190–193.

(Papineni et al., 2001) Kishore Papineni, Salim Roukos, Todd Ward e Wei-Jing Zhuw. BLEU: a Method for Automatic Evaluation of Machine Translation. Relatório Técnico RC22176 (W0109-022). Computer Science IBM Research Division, T.J.Watson Research Center. 17 de Setembro de 2001. `http://domino.watson.ibm.com/library/CyberDig.nsf/Home`. Republicado em ACL'02.

(Pasca, 2004) Marius Pasca. Acquisition of categorized named entities for web search. Em *Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management*. Washington DC, DC, EUA. 8-13 de Novembro de 2004. p. 137–145.

(Paumier, 2002) Sébastien Paumier. Manuel d'utilisation du logiciel Unitex. Relatório técnico. Universidade de Marne-la-Vall. Julho de 2002. `http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf`.

(Petasis et al., 2000) Georgios Petasis, Alessandro Cucchiarelli, Paola Velardi, Georgios Paliouras, Vangelis Karkaletsis e Constantine D. Spyropoulos. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. Em Nicholas J. Belkin, Peter Ingwersen e Mun-Kew Leong, editores, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2000. Atenas, Grécia. 24-28 de Julho de 2000. p. 128–135.

(Petasis et al., 2004) Georgios Petasis, Vangelis Karkaletsis, Claire Grover, Benjamin Hachey, Maria-Teresa Pazienza, Michele Vindigni e Jose Coch. Adaptive, Multilingual Named Entity Recognition in Web Pages. Em *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI 2004. Valência, Espanha. 22-27 de Agosto de 2004. p. 1073–1074.

(Platt, 1999) John C. Platt. Fast training of support vector machines using sequential minimal optimization. Em B. Schölkopf, C.J.C. Burges e A.J. Smola, editores, *Advances in Kernel Methods -Support Vector Learning*. MIT Press. Cambridge, MA, EUA. 1999. p. 185–208.

(Purves e Jones, 2004) Ross Purves e Christopher B. Jones. Workshop on Geographic Information Retrieval. *SIGIR Forum*. 38(1). 2004.

(Pustejovsky, 1994) James Pustejovsky. Semantic Typing and Degrees of Polymorphism. Em Carlos Martin-Vide, editor, *Current Issues in Mathematical Linguistics*. Elsevier. Amesterdão, Holanda. 1994. p. 221–238.

(Pustejovsky, 1995) James Pustejovsky. *The Generative Lexicon*. MIT Press. Cambridge, MA, EUA. 1995.

(Rauch et al., 2003) Erik Rauch, Michael Bukatin e Kenneth Baker. A confidence-based framework for disambiguating geographic terms. Em *Proceedings of the HLT-NAACL 2003 Workshop on the Analysis of Geographic References*. Edmonton, Canadá. 27 de Maio a 1 de Junho de 2003.

(Riezler e Maxwell III, 2005) Stefan Riezler e John T. Maxwell III. On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. Em *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization, MTSE 2005*. Ann Arbor, MI, EUA. Junho de 2005. p. 57–64.

(Riloff, 1996) Ellen Riloff. Automatically Generating Extraction Patterns from Untagged Text. Em *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96*. Portland, OR, EUA. 4-8 de Agosto de 1996. p. 1044–1049.

(Rocha e Santos, 2007) Paulo Rocha e Diana Santos. CLEF: Abrindo a porta à participação internacional em avaliação de RI do português. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. Lisboa, Portugal. 20 de Março de 2007. p. 143–158.

(Rocha e Santos, 2000) Paulo Alexandre Rocha e Diana Santos. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada*, PROPOR 2000. Atibaia, SP, Brasil. 19-22 de Novembro de 2000. p. 131–140.

(Sakai, 2006) Tetsuya Sakai. Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document. Em Hwee Tou Ng, Mun-Kew Leong, Min-Yen Kan e Donghong Ji, editores, *Proceedings of Third Asia Information Retrieval Symposium*, AIRS 2006. Springer. Nova Iorque, NY, EUA. 2006. p. 374–389.

(Sang, 2002) Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. Em Dan Roth e Antal van den Bosch, editores, *Proceedings of CoNLL-2002, the 6th Conference on Natural Language Learning*. Taipé, Formosa. 31 de Agosto a 1 de Setembro de 2002. p. 155–158.

(Sang e Meulder, 2003) Erik F. Tjong Kim Sang e Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Em Walter Daelemans e Miles Osborne, editores, *Proceedings of the Conference on Computational Natural Language Learning*, CoNLL-2003. Edmonton, Canadá. 31 de Maio a 1 de Junho de 2003. p. 142–147.

(Santos, 1940) Delfim Santos. *Conhecimento e realidade*. Tese de doutoramento. Universidade de Coimbra. Lisboa, Portugal. 1940.

(Santos, 2007a) Diana Santos, editor. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. Lisboa, Portugal. 20 de Março de 2007.

(Santos, 1997) Diana Santos. The importance of vagueness in translation: Examples from English to Portuguese. *Romansk Forum*. 5:43–69. Junho de 1997. Versão bilingue revista em TradTerm 5.1, Revista do centro interdepartamental de tradução e terminologia, FFLCH - Universidade de São Paulo, 1998, "A relevância da vagueza para a tradução, ilustrada com exemplos de inglês para português", pp.41-70 / "The relevance of vagueness for translation: Examples from English to Portuguese", pp. 71-78.

(Santos, 1999) Diana Santos. Processamento computacional da língua portuguesa: Documento de trabalho. Versão base de 9 de Fevereiro de 1999; revista a 13 de Abril de 1999. 1999. http://www.linguateca.pt/branco/index.html.

(Santos, 2000) Diana Santos. O projecto Processamento Computacional do Português: Balanço e perspectivas. Em Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada*, PROPOR 2000. Atibaia, SP, Brasil. 19-22 de Novembro de 2000. p. 105–113.

(Santos, 2002) Diana Santos. Um centro de recursos para o processamento computacional do português. *DataGramaZero - Revista de Ciência da Informação*. 3(1). Fevereiro de 2002. http://www.dgz.org.br/fev02/Art_02.htm.

(Santos, 2006a) Diana Santos. HAREM: the first evaluation contest for Named Entity Recognition in Portuguese. Palestra convidada no IST. Lisboa, Portugal. 24 de Fevereiro de 2006. http://www.linguateca.pt/documentos/SantosISTFev2006.pdf.

(Santos, 2006b) Diana Santos. Reconhecimento de entidades mencionadas. Palestra convidada na PUC. Rio de Janeiro, Brasil. 18 de Maio de 2006. http://www.linguateca.pt/Diana/download/SantosPalestraPUCRio2006.pdf.

(Santos, 2006c) Diana Santos. Resumo da actividade da Linguateca de 15 de Maio de 2003 a 15 de Dezembro de 2006. Relatório técnico. Linguateca. Dezembro de 2006. http://www.linguateca.pt/documentos/RelatorioLinguateca2003-2006.pdf. Com a colaboração (por ordem alfabética) de Alberto Simões, Ana Frankenberg-Garcia, Belinda Maia, Luís Costa, Luís Miguel Cabral, Luís Sarmento, Marcirio Chaves, Mário J. Silva, Nuno Cardoso, Paulo Gomes e Rui Vilela.

(Santos, 2006d) Diana Santos. What is natural language? Differences compared to artificial languages, and consequences for natural language processing. Palestra con-

vidada no SBLP2006 e no PROPOR'2006 . Itatiaia, RJ, Brasil. 15 de Maio de 2006. http://www.linguateca.pt/Diana/download/SantosPalestraSBLPPropor2006.pdf.

(Santos, 2007b) Diana Santos. Avaliação conjunta. Em Diana Santos, editor, *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press. Lisboa, Portugal. 20 de Março de 2007. p. 1–12.

(Santos e Bick, 2000) Diana Santos e Eckhard Bick. Providing Internet access to Portuguese corpora: the AC/DC project. Em Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis e Gregory Stainhauer, editores, *Proceedings of the Second International Conference on Language Resources and Evaluation*, LREC 2000. Atenas, Grécia. 31 de Maio a 2 de Junho de 2000. p. 205–210.

(Santos e Cardoso, 2006) Diana Santos e Nuno Cardoso. A Golden Resource for Named Entity Recognition in Portuguese. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*. Springer. Berlin/Heidelberg. 2006. p. 69–79.

(Santos e Costa, 2005) Diana Santos e Luís Costa. A Linguateca e o projecto 'Processamento Computacional do português'. *Terminómetro - Número especial nº 7 - A terminologia em Portugal e nos países de língua portuguesa em África*. p. 63–69. 2005.

(Santos e Gasperin, 2002) Diana Santos e Caroline Gasperin. Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Espanha. 29-31 de Maio de 2002. p. 597–604.

(Santos e Inácio, 2006) Diana Santos e Susana Inácio. Annotating COMPARA, a grammar-aware parallel corpus. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'2006. Génova, Itália. 22-28 de Maio de 2006. p. 1216–1221.

(Santos e Ranchhod, 1999) Diana Santos e Elisabete Ranchhod. Ambientes de processamento de corpora em português: Comparação entre dois sistemas. Em Irene Rodrigues e Paulo Quaresma, editores, *Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, PROPOR'99. Évora, Portugal. 20-21 de Setembro de 1999. p. 257–268.

(Santos e Sarmento, 2003) Diana Santos e Luís Sarmento. O projecto AC/DC: acesso a corpora/disponibilização de corpora. Em Amália Mendes e Tiago Freitas, editores, *Ac-*

tas do XVIII Encontro Nacional da Associação Portuguesa de Linguística, APL 2002. Porto, Portugal. 2-4 de Outubro de 2003. p. 705–717.

(Santos et al., 2003) Diana Santos, Luís Costa e Paulo Rocha. Cooperatively evaluating Portuguese morphology. Em Nuno J. Mamede, Jorge Baptista, Isabel Trancoso e Maria das Graças Volpe Nunes, editores, *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003*. Springer. Berlin/Heidelberg. 2003. p. 259–266.

(Santos et al., 2004) Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela e Susana Afonso. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. Em Guillermo De Ita Luna, Olac Fuentes Chávez e Mauricio Osorio Galindo, editores, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence*, IBERAMIA 2004. Puebla, México. Novembro de 2004. p. 147–154.

(Santos et al., 2006) Diana Santos, Nuno Seco, Nuno Cardoso e Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'2006. Génova, Itália. 22-28 de Maio de 2006. p. 1986–1991.

(Sarmento, 2006a) Luís Sarmento. BACO - A large database of text and co-occurrences. Em Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik e Daniel Tapias, editores, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'2006. Génova, Itália. 22-28 de Maio de 2006. p. 1787–1790.

(Sarmento, 2006b) Luís Sarmento. SIEMÊS - A Named Entity Recognizer for Portuguese Relying on Similarity Rules. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*. Springer. Berlin/Heidelberg. 2006. p. 90–99.

(Sarmento e Mota, 2006) Luís Sarmento e Cristina Mota. HAREM 2.0 Proposta. Encontro do HAREM. Porto, Portugal. Apresentação. 15 de Julho de 2006. `http://www.linguateca.pt/documentos/harem_2.0.ppt`.

(Sarmento et al., 2004) Luís Sarmento, Belinda Maia e Diana Santos. The Corpógrafo - a Web-based environment for corpora research. Em Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa e Raquel Silva, editores, *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation*. Lisboa, Portugal. 26-28 de Maio de 2004. p. 449–452.

(Sarmento et al., 2006) Luís Sarmento, Ana Sofia Pinto e Luís Cabral. REPENTINO - A Wide-Scope Gazetteer for Entity Recognition in Portuguese. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*. Springer. Berlin/Heidelberg. 2006. p. 31–40.

(Savoy, 1997) Jacques Savoy. Statistical Inference in Retrieval Efectiveness Evaluation. *Information Processing and Management*. 33:495–512. 1997.

(Schank e Rieger, 1974) Roger Schank e Charles Rieger. Inference and the computer understanding of natural languages. *Artificial Intelligence*. 5(4):373–412. 1974.

(Schilder et al., 2004) Frank Schilder, Yannick Versley e Christopher Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. Em *Proceedings of the Workshop on Geographic Information Retrieval held at the 27th Annual International ACM SIGIR Conference*, GIR 2004. Sheffield, Reino Unido. 25-29 de Julho de 2004.

(Schölkopf e Smola, 2002) Bernhard Schölkopf e Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press. Cambridge, MA, EUA. 2002.

(Schröer, 2002) Ingo Schröer. A Case Study in Part-of-Speech tagging Using the ICOPOST Toolkit. Relatório técnico. Departmento de Informática da Universidade de Hamburgo. 2002.

(Seco et al., 2006) Nuno Seco, Diana Santos, Rui Vilela e Nuno Cardoso. A Complex Evaluation Architecture for HAREM. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira e Maria Carmelita Dias, editores, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*. Springer. Berlin/Heidelberg. 2006. p. 260–263.

(Sekine et al., 2002) Satoshi Sekine, Kiyoshi Sudo e Chikashi Nobata. Extended named entity hierarchy. Em Manuel González Rodrigues e Carmen Paz Suarez Araujo, editores, *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Espanha. 29-31 de Maio de 2002. p. 1818–1824.

(Sheskin, 2000) David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Springer. Nova Iorque, NY, EUA. 2ª edição. 2000.

(Silberztein, 1993) Max Silberztein. *Dictionnaires électroniques et analyse lexicale du français. Le système INTEX*. Masson. Paris, França. 1993.

(Silberztein, 2004) Max Silberztein. NooJ: A Cooperative, Object-Oriented Architecture for NLP. Em *INTEX pour la Linguistique et le traitement automatique des langues*. Presses Universitaires de Franche-Comté. Besançon, França. 2004.

(Silva, 2004) Cândida Gonçalves da Silva. Specification of the knowledge representation standard of IKF-P (E!2235). Relatório técnico. Universidade do Minho, Departamento de Informática. 2004.

(Simões e Almeida, 2002) Alberto Manuel Simões e José João Almeida. Jspell.pm - um módulo de análise morfológica para uso em Processamento de Linguagem Natural. Em Anabela Gonçalves e Clara Nunes Correia, editores, *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística*, APL 2001. Lisboa, Portugal. 2-4 de Outubro de 2002. p. 485–495.

(Smith e Crane, 2001) David A. Smith e Gregory Crane. Disambiguating Geographic Names in a Historical Digital Library. Em *Research and Advanced Technology for Digital Libraries, 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, Proceedings*. Springer. Berlin/Heidelberg. 2001. p. 127–136.

(Smith e Mann, 2003) David A. Smith e Gideon S. Mann. Bootstrapping toponym classifiers. Em *Proceedings of the HLT-NAACL 2003 Workshop on the Analysis of Geographic References*. Edmonton, Canadá. 27 de Maio a 1 de Junho de 2003.

(Solorio, 2005) Thamar Solorio. *Improvement of Named Entity Tagging by Machine Learning*. Tese de doutoramento. Instituto Nacional de astrofísica, óptica e electrónica, Puebla, México. 2005.

(Stitson et al., 1996) Mark O. Stitson, Jason A. E. Weston, Alex Gammerman, Volodya Vovk e Vladimir Vapnik. Theory of support vector machines. Relatório Técnico CSD-TR-96-17. Universidade de Londres, Royal Holloway. Egham, Reino Unido. Dezembro de 1996.

(Sundheim, 1995) Beth Sundheim. Overview of Results of the MUC-6 Evaluation. Em *Proceedings of the 6th Message Understanding Conference*, MUC-6. Columbia, MD, EUA. 6-8 de Novembro de 1995. p. 13–31.

(Suárez e Palomar, 2002) Armando Suárez e Manuel Palomar. A Maximum Entropy-based Word Sense Disambiguation System. Em *Proceedings of the 19th International Conference on*

Computational Linguistics, COLING 2002. Taipé, Formosa. 24 de Agosto a 1 de Setembro de 2002. p. 960–966.

(Tettamanzi, 2003) Andrea G. B. Tettamanzi. Approaches to knowledge extraction based on soft computing. Relatório Técnico IKF-I Actività R1, 15.1. Universidade de Milão. 2003.

(Téllez et al., 2005) Alberto Téllez, Manuel Montes y Gómez e Luis Villaseñor. A machine learning approach to information extraction. Em Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: Sixth International Conference*, CICLing 2005. Springer. Berlin/Heidelberg. 2005. p. 539–547.

(Tong e Koller, 2001) Simon Tong e Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*. 2:45–66. 2001.

(Toral, 2005) Antonio Toral. DRAMNERI: a free knowledge based tool to Named Entity Recognition. Em *Proceedings of the 1st Free Software Technologies Conference*. A Coruña, Espanha. 23-25 de Março de 2005. p. 27–32.

(Toral e Muñoz, 2006) Antonio Toral e Rafael Muñoz. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. Em *Proceedings of the workshop on New Text Wikis and blogs and other dynamic text sources, 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Itália. Abril de 2006.

(Vapnik, 1995) Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer. Nova Iorque, NY, EUA. 1995.

(Vatant, 2006) Bernard Vatant. The geonames ontology. 2006. `http://www.geonames.org/ontology/`.

(Voorhees e Buckley, 2002) Ellen M. Voorhees e Chris Buckley. The effect of topic set size on retrieval experiment error. Em Kalervo Järvelin, Micheline Beaulieu, Ricardo Baeza-Yates e Sung Hyon Myaeng, editores, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2002. Tampere, Finlândia. 11-15 de Agosto de 2002. p. 316–323.

(Wakao et al., 1996) Takahiro Wakao, Robert Gaizauskas e Yorick Wilks. Evaluation of an Algorithm for the Recognition and Classification of Proper Names. Em Bente Maegaard, editor, *Proceedings of the 16th International Conference of Computational Linguistics*, COLING'96. Copenhaga, Dinamarca. 5-9 de Agosto de 1996. p. 418–423.

(Will, 1993) Craig A. Will. Comparing human and machine performance for natural language information extraction: results for English microelectronics from the MUC-5 evaluation. Em *Proceedings of the 5th Message Understanding Conference*, MUC-5. Baltimore, MD, EUA. 25-25 de Agosto de 1993. p. 53–67.

(Witten e Frank, 1999) Ian H. Witten e Eibe Frank. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann. San Francisco, CA, EUA. 1999.

(Zhang e Mani, 2003) Jianping Zhang e Inderjeet Mani. kNN Approach to Unbalanced Data Distributions:. A Case Study involving Information Extraction. Em *Proceedings of The Twentieth International Conference on Machine Learning (ICML-2003), Workshop on Learning from Imbalanced Data Sets II*. Washington DC, DC, EUA. Agosto de 2003.

(Zhou e Su, 2002) GuoDong Zhou e Jian Su. Named entity recognition using an HMM--based chunk tagger. Em *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02. Filadélfia, PA, EUA. Julho de 2002. p. 473–480.

(Zipf, 1949) George K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley. Reading, MA. 1949.

Índice