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.