Departamento de Engenharia Informática
Universidade de Évora

Mestrado em Engenharia Informática

Classificador para língua natural

Gustavo Alexandre Teixeira Laboreiro
Orientadora: Prof. Doutora Irene Pimenta Rodrigues

Outubro de 2007

Departamento de Engenharia Informática
Universidade de Évora

Mestrado em Engenharia Informática

Classificador para língua natural

Gustavo Alexandre Teixeira Laboreiro
Orientadora: Prof. Doutora Irene Pimenta Rodrigues

Outubro de 2007

Prefácio

Este documento contém uma dissertação intitulada "Classificador para língua natural", um trabalho do aluno Gustavo Alexandre Teixeira Laboreiro[1], estudante de mestrado de Engenharia Informática na Universidade de Évora.

A orientadora deste trabalho é a Prof. Doutora Irene Pimenta Rodrigues[2], do Departamento de Informática da Universidade de Évora.

O autor do trabalho é licenciado em Engenharia Informática, pela Universidade de Évora.

Esta dissertação foi entregue em Outubro de 2007.

[1] gustavo.laboreiro@gmail.com
[2] ipr@di.uevora.pt

Agradecimentos

As minhas palavras de apreço aos meus pais e irmã, pelo apoio que me foi dado durante o tempo em que realizei este trabalho.

À Maria Cristina pelo apoio incondicional que me deu, como sempre.

À professora Irene Pimenta Rodrigues pela paciência, orientação e apoio durante todo este tempo.

Sumário

Esta dissertação apresenta um classificador para textos não anotados escritos na língua inglesa, que não necessita de treino.

Para estabelecer a relação entre palavras recorre-se à base de dados do WordNet. Cada palavra do texto é comparada com cada conceito que define os temas de catalogação. Esta comparação é efectuada tendo em consideração a estrutura hierárquica das relações definidas no WordNet. Desta forma é conservada a afinidade entre termos mais gerais ou específicos, bem como entre termos da mesma área.

O programa foi desenvolvido com o fim de integrar um sistema concorrente no TRECVID — um concurso anual que visa encorajar o avanço do desenvolvimento de aplicações na área de busca e indexação de vídeo digital.

Apesar do âmbito inicial ser específico, a aplicação revela grande potencial para ser usado em qualquer texto em inglês.

Natural language classifier

Abstract

This work presents a training-free, English language, unannotated text classifier.

WordNet's database is used as a foundation to relate words. Each word is compared to a concept that defines the classification topics. This operation takes the hierarchical nature of WordNet's relations into account. In this way, the affinity between more general and more specific terms is maintained, as well as terms in the same domain.

The program was developed to integrate a competing system at TREC-VID – an annual competition that aims to encourage research in video indexation and retrieval.

Despite the restricted initial goal, the application shows great potential to be used with any English text.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A era digital trouxe consigo o milagre da abstracção da informação. Liberta da restrição do suporte material, desapareceu o último impedimento à sua massificação e ubiquidade. Já não existe a necessidade de seleccionar o que se pretende guardar, e o que tem de desaparecer.

Para o consumidor comum, a posse e possibilidade de acesso à informação aumentou algumas ordens de magnitude, graças à facilidade e ausência de custo do armazenamento digital. Daí surgiu outro problema: como encontrar a informação que se procura.

As técnicas clássicas de pesquisa continuam a ser as mais predominantes. Este tipo de busca procura substrings (a expressão fornecida) num espaço que pode ou não ter sido indexado anteriormente. Acontece que há diversas situações comuns em que falha.

Em primeiro lugar, quando se sabe apenas superficialmente o que se quer. Por exemplo, pretende-se um livro de mistério cuja história decorra num país da América do Sul, no início do século, onde havia um regime marcial. Esta informação, embora abundante, é demasiado vaga para encontrar o livro.

Noutras alturas, desconhecem-se os termos exactos para a pesquisa. Por exemplo, procura-se por "salary" e está "wage". Procura-se "cup" e está "mug". Algumas pessoas dizem "directory", outras dizem "folder". "hard disc" e "hard drive" são usadas também de forma intermutável. Quer sejam sinónimos, quer sejam coisas semelhantes, as pessoas recordam os conceitos, não os nomes.

Quando uma palavra tem significado diferente em assuntos diferentes, as pesquisas clássicas revelam-se também ineficientes. Principalmente quando uma das áreas é bastante popular. Uma pesquisa num motor de busca por "Broken windows" quase sempre indica sugestões para formatar o disco rígido do computador. Daí as expressões conjugadas e negadas que a maioria dos sistemas de pesquisa suportam, mas que são um mero remendo e pouco utilizadas.

Por fim, quando a palavra é frequente e a lista de documentos encontrados

é grande, se o documento desejado não está entre os primeiros, há que refinar a busca. Pode ser reformulada a expressão ou ser fornecida a informação sobre o tipo de documento, tamanho, data de criação e local observado, entre outros. Esta é a situação mais comum.

Esta forma de pesquisar informação indexada funciona apenas em documentos de texto. Ficheiros de outras formas de *media* necessitam de outras abordagens, embora haja sempre a dificuldade de especificar *o que se pretende procurar.*

A pesquisa em língua natural tem o benefício de ser mais intuitiva e permitir uma maior expressividade ao utilizador. Como a linguagem verbal permite exprimir conceitos difusos e imprecisos, assim como estabelecer contextos, é vista como a interface ideal de comunicação com o mundo digital. No entanto, a tecnologia descrita nas obras de ficção científica (o HAL 9000 em "2001 odisseia no espaço", ou o computador da nave Enterprise na saga "Star Treck", por exemplo), que serve de modelo de referência para todos aqueles que cresceram com estas obras, ainda está longe de ser realidade.

Diz-se que as grandes caminhadas são compostas de muitos passos. E a cada dia, mais um é dado nessa direcção geral.

Expõe-se aqui mais uma tentativa de contribuir para esse objectivo comum.

O presente trabalho descreve um sistema de classificação de *queries* elaborado no âmbito do concurso TRECVID (que trata de reconhecimento e indexação de vídeo). Quer isto dizer que, dado o texto do pedido, identifica quais os temas abordados pelo mesmo.

O TRECVID possui uma lista de 37 conceitos que foram utilizados para anotar os vídeos no concurso de 2006. Cabe à presente aplicação a tarefa de reduzir significativamente o espaço de busca, destacando os vídeos mais propícios a conter as características procuradas, como identificadas na *query.* Quando usada para processar a tradução da transcrição áudio dos filmes, esta aplicação define também as categorias que lhes estão relacionadas, da mesma forma que o faz a qualquer outro texto.

Pode ser afirmado que esta aplicação servirá como guia aos processos seguintes. O artigo [1] descreve a totalidade do sistema concorrente, fazendo menção ao processamento da língua natural.

Apesar da origem do projecto estar relacionada com esta competição, a única influencia que adveio foi nos exemplos usados para teste. Também o tratamento dado à negação é mais adequado para textos curtos como *queries.* Ainda assim, os resultados obtidos foram todos encorajadores.

1.1 Motivação e objectivos

O TRECVID é uma competição que acompanha a série internacional de conferências em Recuperação de Informação TREC, apoiada pelo National

Institute of Standards and Technology (NIST)[2]. O seu objectivo é definido como

> "(...) to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results."

A participação efectuada na referida competição teve como ponto de partida a disponibilização, a todos os concorrentes, de um conjunto de vídeos, a transcrição do áudio traduzido de forma automática para inglês, um conjunto de *queries*, e uma lista de conceitos (acções, eventos, imagens, etc.).

O desafio colocado era eleger, os vídeos que correspondiam às *queries*, e qual o intervalo relevante.

A tarefa especificamente atribuída ao sistema descrito nesta tese é fazer corresponder as *queries* aos conceitos pré-estabelecidos. Por exemplo, da lista de *queries* de 2006:

> Find shots of multiple people in uniform and in formation

deveria, por exemplo, ser classificada com os temas "military", "policeman" e "marching". Está claro que o resultado está dependente da lista de categorias disponíveis.

O sistema de reconhecimento de vídeo está preparado para reconhecer estes conceitos, e saberá como lidar com eles na sua tarefa.

Existem duas fases principais no processamento de um texto: a análise sintáctica e a interpretação semântica (uma interpretação pragmática está fora do âmbito do projecto)[3, 4].

A análise semântica é efectuada pelo VISL (Visual Interactive Syntax Learning), através do serviço prestado pelo website. Esta componente foi escrita em Python — pela sua capacidade de prototipagem rápida, e alto nível —, e traduz o resultado do formato do VISL para Prolog de uma forma directa.

A interpretação semântica foi efectuada em Prolog, a fim de tomar partido das potencialidades fornecidas por esta linguagem por *design*: *backtracking*, expressividade lógica e alto nível da linguagem. Associado a estes benefícios, o WordNet possui uma versão da sua base de dados no formato Prolog.

O WordNet é um "léxico semântico da língua inglesa" [5]. Ou seja, estabelece relações entre palavras através dos seus significados. Esta informação permite inferir uma aproximação entre as palavras do texto e os temas disponíveis.

Continuando o exemplo acima, pretende-se estabelecer uma ligação entre as palavras do texto "uniform" e "formation", e os temas "military" e "policeman" (os únicos dois que mencionam pessoas que podem usar uniforme e se organizam em formações), e "marching", que embora possa ser visto como

uma marcha civil (de protesto, por exemplo), pode ser também uma parada (que tem pessoas de uniforme em formação).

1.2 Organização da tese

De seguida serão apresentados diversos trabalhos relacionados com o presente. Estes incluem as ferramentas já mencionadas, e outros trabalhos na mesma área.

O Capítulo 2 descreve de uma forma simples o sistema proposto, a fim de estabelecer uma ideia abrangente do todo.

No Capítulo 3 são expostas as tarefas que antecedem a interpretação semântica do texto. Estas consistem na definição dos conceitos (temas) de catalogação, e a análise sintáctica do texto, de onde se levantam as palavras mais significativas a processar.

O Capítulo 4 formaliza uma representação dos significados das palavras, tendo em conta a estrutura hierárquica que o WordNet lhes confere. Esta representação é comparada com a dos conceitos pré-definidos, resultando numa pontuação, que será usada como base na eleição dos melhores conceitos. Por fim, é apresentada a formatação XML do output do programa.

O Capítulo 5 avalia diversos testes elaborados para ilustrar o impacto de diversas decisões e abordagens durante a elaboração do sistema.

O Capítulo 6 apresenta uma lista de tarefas a avaliar melhor e implementar, a fim de melhorar o desempenho da aplicação em várias situações.

No Capítulo 7 é feita uma análise crítica de todo o trabalho.

1.3 Trabalho relacionado

1.3.1 O VISL

O VISL (Visual Interactive Syntax Learning) é um projecto do Instituto da Linguagem e Comunicação[1] da Universidade do Sul da Dinamarca[2]. Este projecto engloba o desenvolvimento de um conjunto de ferramentas que possuem fins relacionados com a investigação e educação na área da análise gramatical [6].

O analisador gramatical do VISL foi já adaptado a várias línguas. A sua abordagem de "Constraint Grammar" é baseada no ENGCG [7, 8].

O VISL não é executado localmente, tendo-se optado por usar o serviço on-line. A ligação com o Prolog é feita através de um script Python, que traduz o output de um formato no input do outro.

[1] http://www.humaniora.sdu.dk/institut.html?vis=4&lang=en
[2] http://www.sdu.dk/

1.3.2 O Prolog

O Prolog é a linguagem de programação em IA mais popular na Europa e Japão [4]. Parte do motivo prende-se com ter sido desenhada para este fim.

Como uma linguagem de alto nível, possui várias funcionalidades para permitir ao programador abstrair-se dos pormenores de implementação, e concentrar-se antes no problema em mão. Esta é uma vantagem do paradigma de programação declarativa.

A forma elegante de definir relações lógicas associada à funcionalidade de backtracking encaixa-se perfeitamente no problema actual: definir uma interpretação.

Além do que foi já dito sobre o Prolog, diversas pessoas viram também nesta linguagem uma forma simples e natural de aceder aos dados do projecto WordNet. Daí que tenha sido elaborada uma versão da sua base de dados para esta linguagem — documentado em [9] para a versão 1.7.1. Dado que esta versão continua a ser desenvolvida em paralelo com a principal, é aparente o nível de sinergia entre os sistemas.

Das diversas implementações de Prolog, optou-se pelo SWI. Esta, para além de funcionar em diversas plataformas, consegue lidar com o grande volume de dados presente no WordNet de forma bastante rápida.[3] O seu interpretador permite um ciclo de desenvolvimento mais rápido (não existe o passo de compilação), e para além disso, está disponível para diversos sistemas operativos. Por fim, possui ainda vários mecanismos que facilitam muito o desenvolvimento de aplicações [10].

1.3.3 O WordNet

O WordNet é um léxico da língua inglesa (e de outras línguas, numa escala menor) desenvolvido na Universidade de Princeton, que integra colaboração de diversas fontes, e é muito utilizado em diversas aplicações informáticas relacionadas com a análise linguística. [11, 5]

Neste sistema cada palavra possui diversos *synsets*,[4] de forma a poder ser interpretada de forma não-ambígua. Por exemplo, "playing" tem no WordNet um *synset* para indicar "the act of playing a musical instrument" e outro para "the action of taking part in a game or sport or other recreation".

Estes *synsets* são depois associados em diversas estruturas (taxonomias), no espírito mais geral sugerido em [12]:

> "A scheme that partitions a body of knowledge and defines the relationships among the pieces. It is used for classifying and undertranding the body of knowledge."

[3] O GNU Prolog não consegue sequer carregar os dados todos.

[4] Na verdade, existem também pequenas expressões e substantivos próprios com direito a significado. Neste artigo, o termo "palavra" é usado de forma a abranger também estes constituintes linguísticos.

Existem diversas relações definidas na base de dados do WordNet. A principal é a hiponímia (o cão *é um* animal), mas há também, a título exemplificativo:

- meronímias (uma matilha *é composta de* cães),
- classificação (célula *é classificada no domínio de* biologia),
- causa (mostrar *é causa para* ver),
- semelhança (agir *é semelhante a* comportar),
- atributo (bom *é atributo que indica* qualidade),
- antónimo (bom *é o oposto de* mau).

A maior parte destas relações não foi usada. Este aspecto será devidamente detalhado na Secção 4.1.

A informação no WordNet está organizada numa forma que se assemelha a uma base de dados. As relações definidas correspondem às relações no modelo Entidade-Relação, com os *synsets* correspondendo às chaves das entidades, que por sua vez encontram analogia nos significados dos *synsets*. Estes estão descritos numa relação própria.

As relações binárias, como as que foram enunciadas acima, podem ser acedidas através do Prolog como, por exemplo, `hyp(X,Y)`, em que estas variáveis unificam com dois *synsets* diferentes. Duas relações especiais são a relação `s`, que estabelece a relação entre uma palavra e um *synset*, e a relação `g`, que apresenta uma explicação, por vezes exemplificada (*gloss*).

Devido à forma como está estruturado, o WordNet foi usado como uma aproximação a uma ontologia simples. Espera-se usufruir da sua riqueza, a fim de conseguir identificar os elementos mencionados nos textos, e relacioná-los correctamente com os contextos de classificação disponíveis.

Comparação entre o WordNet e uma taxonomia Uma ideia a reter ao usar o WordNet, é que nem todos os nós têm o mesmo peso. Observando a Figura 4.1, página 36, nota-se por exemplo que a classe "Mamífero", tem uma grande significância. Todos os mamíferos passam por gestação interna, são amamentados no início de vida, possuem sangue quente, pêlo a cobrir o corpo, para além de outras características.

Mais abaixo encontra-se a classe "Mamíferos aquáticos". Esta categoria não adiciona mais aos seus nós descendentes, do que o habitat aquático.

Infelizmente, não é possível distinguir entre as categorias "ricas" e as "pobres" sem recorrer a uma fonte externa, visto que o WordNet não define qualquer sistema de propriedades para os *synsets*.

Comparando com a classificação "oficial" clássica — como o excerto visto na Figura 1.1 —, a diferença de pesos não existe: há um número fixo de

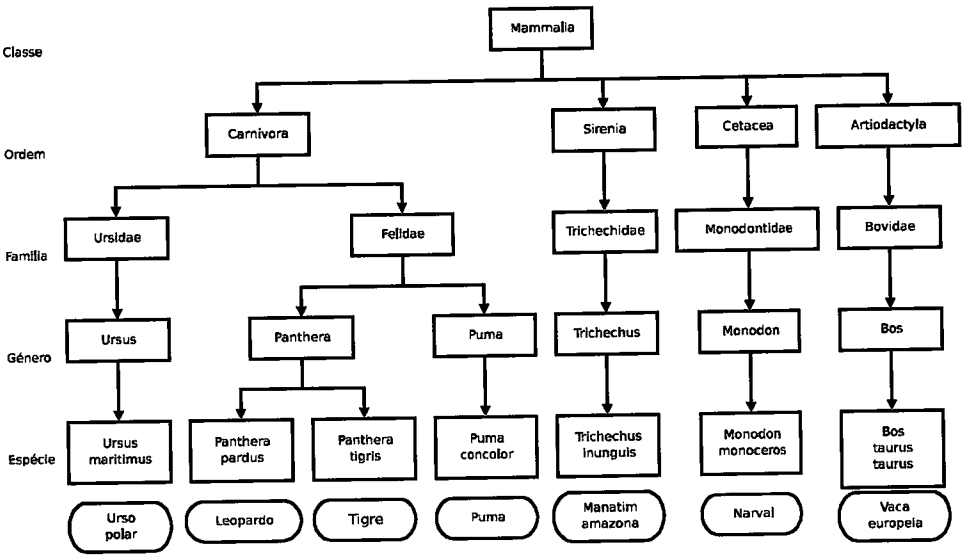

Figura 1.1: Taxonomia científica (simplificada)

categorias, e todas elas devem estar preenchidas. Sabe-se que há um determinado grau de semelhança que é partilhado por todos os indivíduos de uma classe. Esse grau vai sendo maior com a profundidade, e aproximado para todas as classes num determinado nível.

No WordNet algumas categorias são escolhidas com base na classificação científica (Carnivore, Feline, Primate, Cetacean) ou na tradução desta[5] (Hoofed Mammal, Even-toes ungulate, etc.), enquanto outras são mais organizacionais, como Aquatic Mammal e Placental — pois no WordNet não há outra forma de transmitir esta informação. Seria possível incluir uma relação directa entre "Placental" e "Cetacean", dando origem a dois caminhos para este último (um mais científico, e outro mais descritivo), mas para se obter esse nível de detalhe, será melhor trabalhar com uma taxonomia especializada.

O WordNet tenta manter um equilíbrio entre o geral e o detalhado. Nunca haverá consenso no que toca ao conhecimento, e a equipa do WordNet procura agradar ao maior número de pessoas. Comumente, estas não procuram a complexidade de informação com que um perito na matéria lida.

Em última análise, apesar dos seus defeitos, o WordNet consegue fornecer mais informação (em quantidade e em qualidade) do que um algoritmo de aprendizagem, como SVMs, redes neuronais ou *clustering* consegue obter.

As principais vantagens do uso de um sistema simbólico baseado no WordNet, quando comparado com a típica abordagem de aprendizagem, são:

- Não existe a morosa fase de treino supervisionado, ou a necessidade

[5] A classificação científica de espécies criada por Linnaeus no séc. XVIII, tem por base o latim, por ser uma língua morta.

de catalogar um corpus significativo que costuma variar bastante para diferentes ambientes de uso,

- É fácil de expandir a rede para reconhecer novos elementos,
- Os erros são facilmente analisados, já que todas as decisões tomadas pelo sistema são justificáveis e facilmente verificadas,
- As definições de classificação podem seguir um caminho evolutivo baseado em testes de hipóteses, e escolher a melhor abordagem,
- Pode ser adaptado para assistir noutras tarefas relacionadas com linguística. Por exemplo, ajudar na tradução automática.

Existem também desvantagens associadas ao uso deste sistema:

- Se uma relação não está presente de forma explícita, esta não existe. Nos SVM ou clusters há um grau de probabilidade que pode ser explorado.
- O processo de expansão é moroso, pois tipicamente envolve acrescentar diversas ligações a cada *synset* novo.
- Assenta em conteúdos já criados. Para uso em contextos especializados ou outros assuntos não cobertos, será necessário criar toda a informação no WordNet.

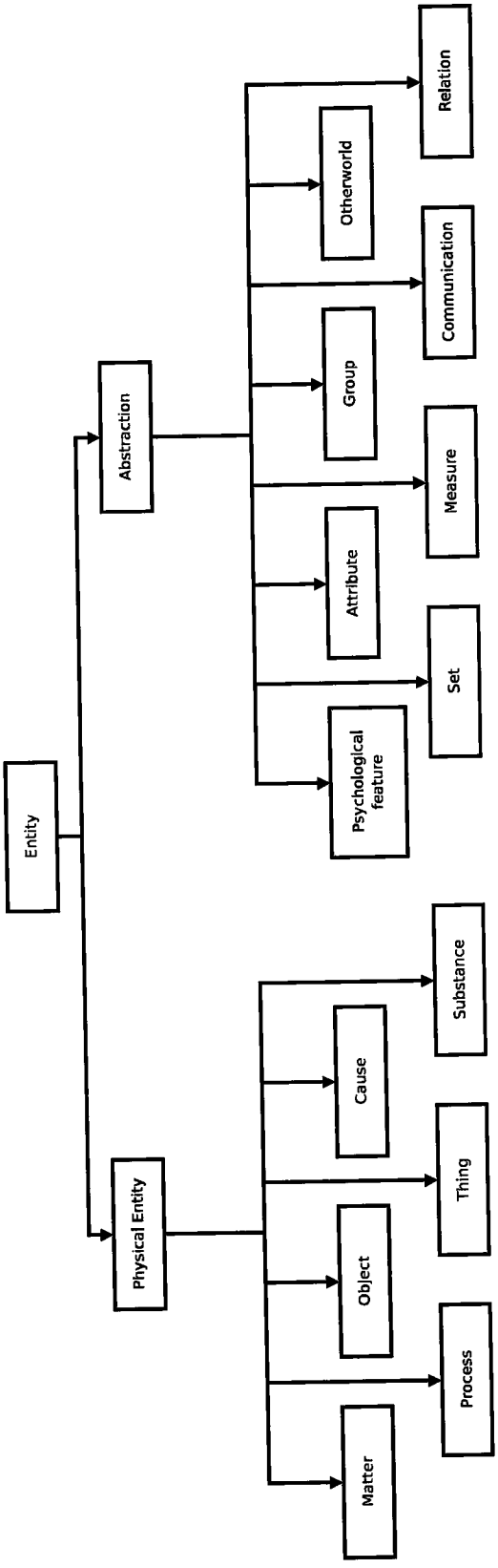

Figura 1.2: Nível de topo do WordNet

1.4 Outros sistemas de classificação

Os sistemas de classificação, como grande parte dos algoritmos em IA, podem ser divididos em dois grandes grupos: supervisionados e não supervisionados.

Destes, os algoritmos supervisionados (que necessitam de uma fase de treino a fim de poderem desempenhar as suas funções de forma adequadas) destacam-se como os mais populares nos últimos tempos. Por um lado, porque é certo que nada é estático e constante, e estes conseguem adaptar-se às novas situações sem ser necessário modificar o seu código — basta estender o treino na nova direcção. Também as necessidades de "afinação", típicas das mudanças de domínios, podem ser conseguidas da mesma forma.

De entre os algoritmos utilizados para a classificação de texto, aquele que é mais conhecido é o "naïve Bayes classifier". Isto porque foi o primeiro algoritmo eficiente para lidar com email não solicitado [13, 14]. É ainda muito usado em software como o SpamAssassin.

Actualmente as "Support Vector Machines" são a opção mais utilizada, resolvendo alguns problemas relacionados com a redes neuronais, ou seja, conseguem representar funções complexas e não lineares com uma quantidade de exemplos que não é excessiva [3]. Como foi dito em [15]:

> "Support Vector Machines have many appealing features. They are a rare example of methodology where geometric intuition, elegant mathematics, theoretical guarantees, and practical algorithms meet. They can be efficiently applied to a wide range of classification problems. They scale to massive datasets and are problem-domainin dependent. They are 'open' in nature in the sense that efficient kernel functions can be developed for each specific problem for better results. There are few model parameters to pick and there are many successful applications in the domain of bioinformatics (microarray data classification), face detection and handwriting recognition. The above simple experiment on the popular dataset gives the empirical evidence that SVMs are also well suited for text categorization."

Do ponto de vista dos algoritmos não supervisionados, existe a grande vantagem de dispensam a fase de treino. São sempre iniciados num estado pré-determinado.

Um exemplo interessante consiste num sistema (chamado SONIA) que usa técnicas de *clustering* para agrupar documentos por temas [16]. Cada um dos agrupamentos pode ser depois ampliado e ver-se reestruturado, dando origem a mais agrupamentos específicos.

1.5 Trabalhos semelhantes

Infelizmente, a tarefa de processamento textual parece ser um processo de segundo plano no TRECVID. Muitas vezes não são mencionadas, ou são descritas em menos de uma página. Daí que esta secção pode apenas dar uma ideia das direcções tomadas pelas equipas participantes.

Muitos dos grupos concorrentes deixaram o processamento do texto — parcialmente ou de forma aparentemente completa — a cargo de sistemas independentes. Exemplos disso são a Universidade Chinesa de Hong Kong [17], que usou o Lemur toolkit [18], e a AT&T [19] que usou o LinPipe [20]. Esta decisão está, claro, dependente dos aspectos em que cada equipa se quer concentrar.

A AT&T usa uma abordagem muito interessante: recorre a material divulgado/publicado na mesma altura em que os vídeos do TRECVID foram recolhidos, e usa-os para enriquecer a sua capacidade de reconhecer, por exemplo, pessoas ou locais através do seu nome [19].

A equipa da Universidade de Amesterdão recorre a dois algoritmos. O chamado "text matching" procura o conceito cuja correlação estatística seja maior relativamente ao texto da *query*. O segundo, de nome "ontology querying", procura cada palavra da *query* no WordNet. Para cada palavra encontrada, considera o *synset* mais comum, e calcula a semelhança com cada conceito. [21]

Uma equipa formada por investigadores da Universidade de Tecnologia, em Helsínquia, e um da Universidade de Dublin propõe um algoritmo que procura no WordNet sinónimos para os conceitos. Se um desses sinónimos for encontrado numa *query*, então o conceito aplica-se a ela. Se esse sinónimo surgir precedido de "not", então o conceito aplica-se de forma negativa [22].

Por fim, a equipa da Universidade de Hon Kong recorre também ao Lemur [18]. No entanto, desenvolve sobre este um algoritmo elaborado. Em primeiro lugar, usa um dicionário para detectar nomes próprios, que através do WordNet são substituídos pela sua forma mais curta. De seguida é feito um cálculo lexical e são mantidos apenas os substantivos, que são desambiguados. Por fim, uma vez mais com recurso aos sinónimos do WordNet, quatro heurísticas estão definidas para eleger os conceitos referentes à *query* [23].

Capítulo 2

Proposta

A ideia que originou este trabalho consiste no desenvolvimento de um sistema livre do requisito de treino que consiga, dado um texto não anotado em inglês, classificá-lo em relação a uma lista de conceitos. Este sistema pretende-se também genérico no âmbito de trabalho, e adaptável a várias tarefas, sejam elas a análise de queries ou de textos inteiros.

O TRECVID não coloca um limite de tempo de processamento. Os dados são entregues aos participantes, e têm um certo número de dias para entregar os resultados. Logo, esse foi um aspecto que passou para segundo plano no desenho da aplicação.

Optou-se por ter como base o WordNet, que é a principal ferramenta usada na análise simbólica da linguagem, devido à quantidade de informação que contém, à organização desta e à transversalidade dos domínios de conhecimento.

Nesta Secção serão apresentadas e justificadas as escolhas de ferramentas de trabalho usadas para elaborar o classificador. De seguida será feita uma introdução geral a toda a arquitectura do sistema, onde cada passo é definido.

A arquitectura do sistema A aplicação opera sobre um ficheiro de cada vez, que se assume conter apenas um texto[1]. Um texto pode corresponder a uma *query*, como "Find shots of multiple people in uniform and in formation", ou a textos correspondentes à tradução automática da transcrição dos vídeos para inglês (o que introduz imenso ruído no discurso escrito). Ambas as tarefas foram executadas no âmbito do TRECVID, mas um texto pode também corresponder a um documento num arquivo a ser classificado automaticamente, ao conteúdo de uma página web, a textos colocados em blogues, a notícias ou artigos de jornais, etc.. A título de exemplo, na Secção 5.15 será mostrado o resultado de correr sobre alguns textos do *corpus*

[1] Caso se deseje trabalhar com ficheiros contendo mais do que um texto, foi criada uma ferramenta para o particionar em diversos ficheiros, permitindo assim ultrapassar esta limitação.

conhecido como Reuters-21578.

Deve realçar-se que a classificação é relativa a todo o texto. Quer isto dizer que, se se procura maior granulosidade nos resultados, isto é, se os textos variam muito de tema em subsecções diferentes — e é desejado que o classificador acompanhe e reflicta essas variações — então o texto deve ser dividido em partes menores, correspondentes aos citados limites.

Cada texto passa sempre pelos mesmos processos sequenciais, indicados na Figura 2.1, até ser reduzido aos conceitos básicos disponíveis:

Definição dos conceitos Os conceitos-alvo são declarados como *synsets* do WordNet. Um conceito é um tema que pode ser usado para classificar qualquer texto. Não são mutuamente exclusivos, ou seja, mais do que um pode ser aplicado ao mesmo texto.

Análise sintáctica A ferramenta VISL efectua o processamento do texto, identificando as classes gramaticais das palavras, sujeitos da frase, etc.. O resultado deste processo é um ficheiro Prolog com toda a informação, que é posteriormente consultado pela aplicação principal.

Seleccionar palavras Nesta fase identificam-se as palavras mais relevantes do texto. Por abuso de linguagem, usa-se o termo "palavra" para definir os objectos mais significativos do documento, e que são o alvo de análise da aplicação. Um substantivo complexo ou um substantivo próprio é considerado uma única palavra, se reconhecido pelo WordNet dessa forma.

Todo o restante processamento assenta nos substantivos reconhecidos no texto. São formadas duas listas: a dos substantivos afirmados (o caso comum) e a dos negados (precedidos por "not", por exemplo). Ambos são tratadas de forma igual, sendo a única distinção feita na fase de apresentação de resultados.

Cada uma das listas passa ainda por uma fase de "agregação". Alguns substantivos não são reconhecidos correctamente pelo VISL (no passo anterior). Como tal é feita uma tentativa de recuperar essa informação perdida. Por exemplo, "Bill Clinton" surge na análise semântica como duas palavras separadas. Visto o WordNet reconhecer o nome completo, será desta forma (considerada uma palavra) que seguirá processamento.

Pontuar os conceitos Para definir a afinidade entre dois *synsets* — o do tema e o da palavra —, recorre-se aos caminhos, que são uma lista de *synsets* que descrevem o percurso levado do significado mais geral até ao mais específico numa relação do WordNet.

Obtendo o caminho do conceito e o caminho da palavra (para cada *synset* de cada um), é fácil comparar cada elemento da lista até di-

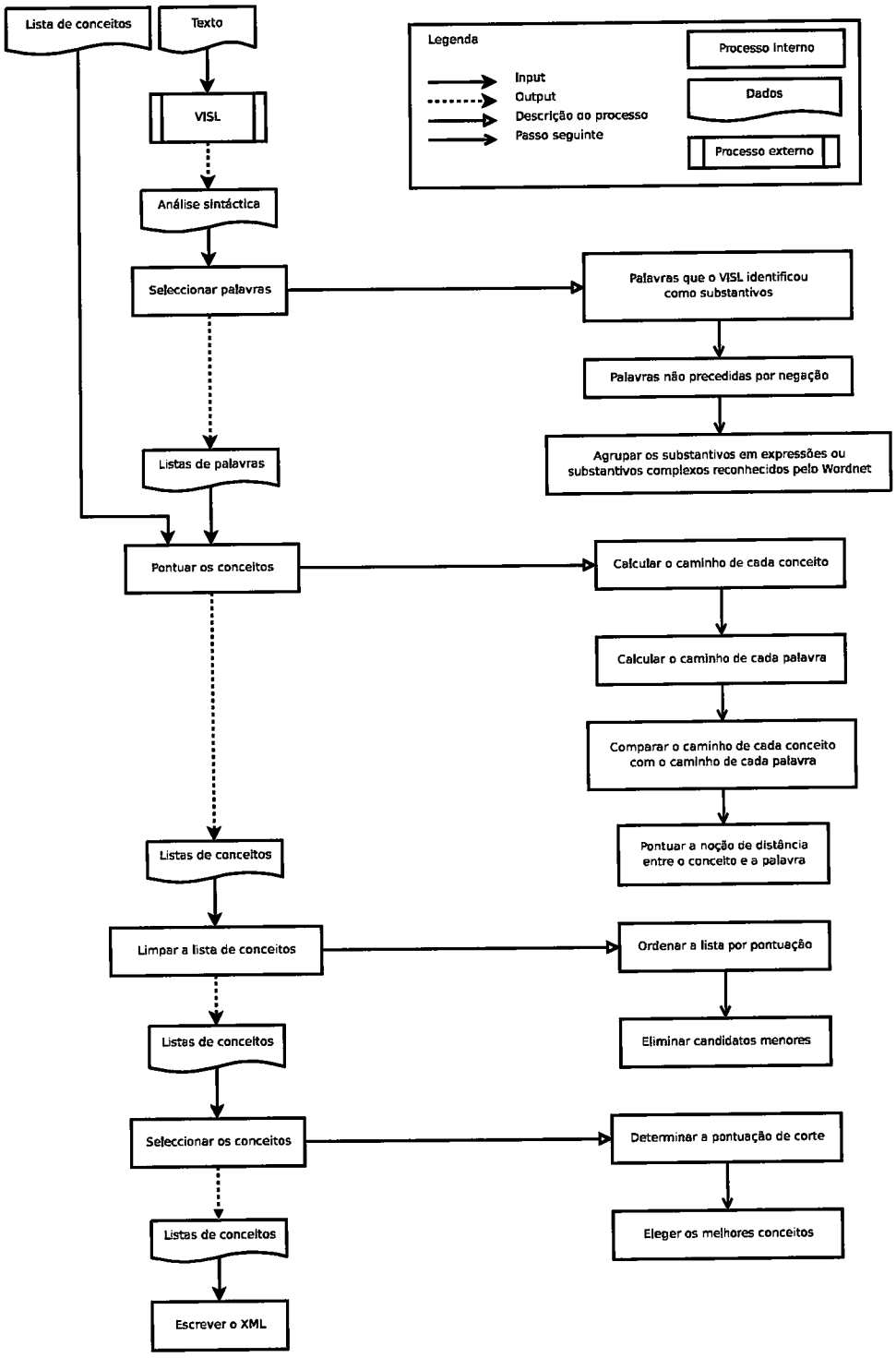

Figura 2.1: Esquema geral do processamento

vergirem, obtendo-se desse modo uma noção da distância (o total de elementos diferentes que restavam ainda nas listas).

Com base no número de elementos em comum e o número de elementos diferentes nos dois caminhos, é feita uma avaliação que resulta num índice de diferença — a pontuação. Quanto menor este número, mais semelhantes são as duas palavras.

Este processo é elaborado tendo em consideração todas as combinações (C, P), onde C é um *synset* associado a um conceito, e P é um *synset* associado a uma palavra. Daqui resulta uma lista extensa de resultados.

Limpeza da lista de conceitos A fim de se proceder à identificação dos conceitos mais significativos, a lista dos conceitos pontuados é ordenada de forma crescente da pontuação (melhores primeiros) o que simplifica a tarefa de selecção. Os elementos redundantes são depois removidos.

Por exemplo, dois conceitos iguais apontados por palavras diferentes: `(house, 0, home)` e `(house, 6, street)`. O primeiro elemento tem claramente mais interesse (pontuação inferior). Logo, se não chegar à selecção final, o segundo também não o conseguirá. Além disso, não há qualquer interesse em ter o mesmo conceito repetido no *output*. Portanto, o segundo elemento é removido da lista.

Seleccionar os melhores conceitos É definido um limite com base nos melhores desempenhos encontrados. Qualquer conceito na lista (havendo após a limpeza, uma única entrada por conceito), cuja pontuação seja superior a este limite, é automaticamente excluído.

É sempre garantido pelo menos um elemento presente — mesmo que este esteja claramente pouco relacionado com o texto.

Escrever o XML É criado um ficheiro XML onde os conceitos ainda presentes na lista são apresentados, juntamente com a informação adicional, como a pontuação e a palavra do texto que lhe estava mais próxima (responsável por ter sido eleito).

O resultado de cada conceito surge dentro de uma de 3 listas, correspondentes a 3 níveis de certeza (exacto, geral e difuso), com base na medição de distância entre os *synsets*.

A lista de afirmações e a lista de negações são escritas em separado, dentro do *tag* correspondente ao texto, sendo que este último surge apenas quando tem algum resultado a apresentar.

meronímia de parte Indica que algo compõe uma parte de um objecto ou processo.
(9097 relações)
Enxaguar faz parte de lavar.

meronímia de composição Explica os constituintes de algo.
(797 relações)
O pão é feito de farinha.

instância de Destaca um elemento como sendo um caso particular.
(8577 relações)
Lucy era um Australopithecus afarensis.

As relações formam tentativamente uma hierarquia bem definida. Destas, a principal é, sem dúvida, a dos hipónimos. No entanto, existem 346 *raízes* nesta relação. Ou seja, há 345 *synsets* que não descendem de *entity*[1] (a raiz dominante). As medições apontam para 13 263 *synsets* que não estão ligados ao "tronco" da hierarquia, contando com os "descendentes". Havendo 89 089 relações de hipónimo, quase 15% da "rede" está nestes "ramos quebrados".

Isto em nada irá afectar o sistema de momento, já que apenas 11 dos 345 "*synsets*-raiz" que se encontram fora da hierarquia principal são substantivos. Com os seus hipónimos, totalizam 27 nós "isolados" (em 74 389 substantivos). Como foi referido na Secção 3.3, apenas se vai operar sobre os substantivos.

Apesar da tentativa de particionamento do conhecimento no WordNet, não há garantia que haja apenas um caminho entre dois nós, como é o caso numa árvore. Por exemplo, "sonambulismo" descende de "andar" e de "dormir". Esta topologia causa problemas quando se tenta encontrar o caminho mais curto entre *synsets*. A dimensão do espaço de busca no caso dos hipónimos torna esta hipótese impossível usando métodos usuais, visto que não existem heurísticas para guiar na direcção certa.

4.1.1 Caminho pela raiz

Outra alternativa passa por tirar partido da "raiz única" da estrutura da relação.

Define-se como caminho o conjunto de nós visitados, deslocando-se da raiz até ao nó em questão, numa relação. Traduz-se assim um encadeamento dos elementos através dessa relação, que é transitiva. Por exemplo, um dos caminhos para cão é o seguinte:

```
[entity, physical entity, physical object, whole,
living thing, organism, animal, domestic animal, dog]
```

[1] Tecnicamente, *entity* não descende de si próprio, mas é mais fácil de passar a imagem se dito desta forma.

Capítulo 4

Pontuação dos conceitos

Após ter sido elaborada a lista das palavras, há que usá-las para chegar ao núcleo da frase. O que se pretende é uma forma de poder relacionar duas palavras: uma da frase e a outra da lista de conceitos. Dito de outra forma, procura-se um processo que aceite uma lista de palavras da frase, e uma lista de conceitos (na forma como foram definidas atrás), e que devolva uma avaliação de todos os conceitos fornecidos, a fim de se poder aferir os mais significativos.

O método desenvolvido para tal consiste nos seguintes passos:

1. Calcular o caminho para cada conceito
2. Calcular o caminho para cada palavra
3. Para cada par de caminhos $(conceito, palavra)$, calcular a proximidade
4. Associar a cada palavra o conceito que lhe está mais próximo
5. Seleccionar destes conceitos, os mais relevantes

De seguida será analisado cada um destes passos ao pormenor.

4.1 Calcular os caminhos

Como foi referido na secção 1.3.3, o WordNet mantém registo de relações entre diversos significados. Estas relações [30] variam consoante as classes de palavras envolvidas. Assim sendo, apenas as seguintes foram utilizadas:

hiponímia Uma relação de subordinação ou especialização.
(89 089 relações)
Um gato é um mamífero.

meronímia de membro Marca a pertença a um grupo de elementos.
(12 293 relações)
Um peixe é membro de um cardume.

```
Lista_sim = [scene, field, tree, sky, lake, mountain,
river, beach, ocean, grass, sunset, waterfall, animal,
people],
Lista_nao = [building, road, vehicle]
```

É de notar que tratar negações pode ser complicado numa aplicação mais geral. Por exemplo, numa frase "uma árvore sem folhas", isso poderia, em caso de indecisão, permitir que fosse interpretado como um gráfico em árvore[7]. Saber que se pretende uma árvore que possa ter folhas, mas que não as tem, ajudaria nestes casos. Terão que ser ponderados os riscos das escolhas tomadas.

3.3.5 Síntese

Em suma, para uma palavra passar à fase seguinte de processamento, deve satisfazer os seguintes requisitos:

- Ser um substantivo ou grupo de substantivos que formam um substantivo complexo
- Estar presente no WordNet
- Não ser uma palavra ignorada no início de uma frase
- Não ser precedida por uma negação

A lista das frases negadas passa também por todos estes passos, embora exigindo que cada substantivo esteja precedido por uma negação.

Exemplo Continuando com o mesmo exemplo da secção anterior ("Find shots of a sailboat in a lake with a blue sky and no clouds."), os substantivos são "shot", "sailboat", "lake", "sky" e "cloud".

Não existe qualquer substantivo complexo neste exemplo. O primeiro substantivo é eliminado por pertencer ao prefixo típico de queries do TREC-VID, e em nada de bom contribui para a classificação. Por fim, "cloud" é antecedido pela palavra "not", que é classificada como um prefixo negador, e é transferida para a lista de negações.

As lista dos substantivos afirmados resultante deste processo é `[sailboat, lake, sky]`, e a lista dos substantivos negados é `[cloud]`.

[7]Nós sem descendência num gráfico em árvore são por vezes chamados de "folhas" por alguns autores.

3.3.3 Inícios de *queries*

Quase todas as frases das queries do TRECVID começam por "Find shots of". "shot" é um substantivo, e como tal, surge sempre na lista de palavras a processar, apesar de nunca contribuir para o resultado final. "Find" pode ser também um substantivo. Acontece que ignorar sempre estas palavras nas frases poderia ser prejudicial se a frase referisse, por exemplo, tiroteios ou achados. Cortar as três primeiras palavras teria o efeito desejado, mas há sempre a hipótese de surgir outra query que tenha outro formato.

Como equilíbrio entre as duas abordagens referidas, e de modo a evitar problemas futuros, as palavras "find" e "shot" serão ignoradas se surgirem dentro das três primeiras palavras numa frase. Adicionar outra palavra é uma questão de acrescentar no ficheiro dedicado:

```
stop(Palavra).
```

3.3.4 Negativas

Ainda outra questão pode ser observada, por exemplo, na seguinte query do concurso de 2006:

> "Find shots of a natural scene - with, for example, fields, trees, sky, lake, mountain, rocks, rivers, beach, ocean, grass, sunset, waterfall, animals, or people; but no buildings, no roads, no vehicles"

A lista dos substantivos encontrados é a seguinte:

```
Lista = [scene, field, tree, sky, lake, mountain, river,
beach, ocean, grass, sunset, waterfall, animal, people,
building, road, vehicle]
```

Da forma como as palavras são eleitas, "building", "road" e "vehicle" surgem na lista de palavras final, e induzem o sistema em erro. A solução é simples e óbvia, pois cada uma destas palavras surge negada. Assim, antes de aceitar cada palavra, há que verificar se na frase esta é precedida por uma negação (a palavra "no" ou outro determinante negativo), ou outra palavra com função semelhante (por exemplo, "without").

Caso a frase surgisse como "... but no buildings, roads or vehicles", seria muito mais difícil, porque o VISL não agrupa os três substantivos na negativa. Isso significa que teria de haver uma camada sintáctica de alto nível para a tratar esse aspecto. Este grau de complexidade não foi explorado, e nenhuma negativa esteve presente na competição em 2007.

Uma lista das palavras negadas é criada, e é processada em paralelo com a lista normal (de palavras "afirmadas"), dando origem a uma secção própria no *output*.

As duas listas passam a ser:

```
Lista = [shot, 'emergency vehicle', motion, ambulance,
police, car, fire, truck]
```

Este processo continua até ser identificado o máximo de substantivos complexos, resultando em

```
Lista = [shot, 'emergency vehicle', motion, ambulance,
'police car', 'fire truck']
```

3.3.2 Abreviaturas

Verificou-se que o VISL não distingue entre os pontos finais e os pontos indicadores de abreviatura. Ambos são tratados da mesma forma, o que não é desejável. Por exemplo, "George W. Bush" não é encontrado, apesar de existir no WordNet, visto que o ponto é dissociado do 'W'. Também os pontos são muitas vezes considerados implícitos, e não são escritos.

A solução utilizada consiste em, caso o WordNet não reconheça o conjunto de palavras, adicionar um ponto a cada elemento da sublista que seja apenas uma letra maiúscula.

Esta estratégia funciona na maioria dos casos. No entanto, há diversas palavras abreviadas para mais de uma letra. Por exemplo, "Mr." e "Mrs." são casos frequentes, mas o WordNet reconhece a forma sem o ponto. "St." é uma abreviatura associada a "street" e "saint", que é também muito popular, embora no WordNet se encontre associada apenas com o último, por motivos que se compreendem. Mas observando-se o dicionário do WordNet, nota-se que ele não é totalmente consistente nesse aspecto, certamente devido a contemplar vários autores. Há "St Mihiel" e "St Patrick's Day", assim como "St. Patrick" e "St. Barbara's herb" [6]. "Jr." é outra abreviatura conhecida. Por vezes o WordNet não coloca o espaço onde deveria estar, teoricamente. Por exemplo, "O.K." e "a.s.a.p.". Embora não sejam substantivos, indiciam outro ponto frágil na abordagem *naïve*.

Uma estratégia mais robusta será descrita no trabalho futuro.

Exemplo As palavras `['George','W','Bush']` surgiram já no TRECVID 2006. A primeira sublista testada corresponde sempre à totalidade dos elementos. A concatenação dos três elementos falha.

Como existe um elemento que é uma letra maiúscula apenas, adiciona-se o ponto no fim, e tenta-se novamente. Existe uma palavra no WordNet que é "George W. Bush", e a lista é actualizada para reflectir o aproveitamento disso.

[6] Todas as situações semelhantes encontradas foram já comunicadas aos responsáveis, mas não serão resolvidas tão cedo. Muitas outras ocorrências semelhantes devem de existir.

vá substituindo por sentidos mais concretos à medida que se vão detectando resultados incorrectos nos testes.

3.1.3 Conceitos compostos

Um outro problema prende-se com os conceitos complexos. Por exemplo, "female reporter": não existe um *synset* no WordNet que cubra todo o conceito. Uma solução — que não houve tempo para implementar — seria associar estes, não a um *synset* ou conjunto de significados, mas a um ou mais conceitos.

Os conceitos de alto nível seriam destacados apenas quando todos os conceitos que o compõem forem encontrados e considerados relevantes. Uma análise mais aprofundada deste tema encontra-se na Secção 6.6.

3.2 Submissão à análise sintáctica

A análise sintáctica consiste na decomposição das frases e inferição da sua estrutura gramatical, baseada no conhecimento de uma gramática formal. [26]

Neste trabalho, a tarefa de extrair a informação gramatical sobre cada palavra recai sobre um serviço no website do VISL (Visual Interactive Syntax Learning) [6, 27]. Embora haja uma versão que funcione localmente, optou-se por usar o serviço on-line. Todo o sistema criado é independente do Sistema Operativo[4]. Recorrer à versão on-line permite poupar uma dependência, assim como (espera-se) ter sempre acesso à versão mais recente do VISL.

Escreveu-se um script em Python que, dado o nome de um ficheiro, submete o seu conteúdo ao processamento da aplicação web do VISL. O output, em HTML, é depois convertido para Prolog. A informação contida nesse ficheiro segue a forma:

```
coord(Ficheiro, N_parágrafo, N_palavra, Palavra).
```

Onde `Ficheiro` indica o nome do ficheiro processado, e os números de parágrafo e de palavra completam as coordenadas desta. O último campo armazena toda a informação que o VISL indicou sobre a palavra. Este campo não tem formato pré-definido, visto que depende do tipo de palavra, pelo que a aridade deste functor é incerta. Dada a possibilidade de se poder tentar melhorar os resultados no futuro, recorrendo a mais informação apresentada pelo VISL, optou-se por não descartar nenhuma da que nos é apresentada. Esta é representada na seguinte forma:

```
palavra(Palavra_original, Lema, Classe_gramatical,
  Informação_vária,...).
```

[4] O código Prolog é interpretado pelo SWI Prolog, e as ferramentas auxiliares são escritas na linguagem interpretada Python

optar por vários mais concretos? A primeira hipótese tem o inconveniente de poder abranger diversos *synsets* que não são desejados. A segunda implica um maior tempo de processamento [3] — importante se se pretende que o sistema funcione de forma interactiva, onde é imprescindível obter resultados rápidos — e a necessidade de actualizar a lista de conceitos sempre que se estender o WordNet com mais significados.

Em caso de dúvida, o ambiente dita sempre a resposta. Se não for provável a confusão com outro conceito, não haverá necessidade de enunciar todas as possibilidades, já que o conceito mais associável será o mesmo. Esta situação é mais comum quando o domínio dos textos é vasto. Em casos mais concretos, onde o domínio é mais específico e contido, e a margem de manobra é menor — dois significados "irmãos" podem corresponder a dois conceitos diferentes —, é recomendável ir mais ao detalhe, e recorrer a diversos significados por conceito.

Como foi referido na Secção 1.3.3, uma das vantagens de não usar um esquema que necessita de ser treinado é a de poder adaptá-lo de forma simples, rápida e fácil, já que os erros não são dispendiosos. É até recomendado que se tentem várias soluções, e ir adaptando-as na busca de um equilíbrio.

Exemplo O texto que acompanha o conceito "disaster" na lista do TREC-VID diz:

> Shots depicting the happening or aftermath of a natural disaster such as earthquake, flood, hurricane, tornado, tsunami.

Conforme já foi mencionado, há que pesar a abrangência e a precisão. Tornados e furacões são fenómenos atmosféricos. Também as tempestades, os ciclones, e o efeito de estufa estão entre os fenómenos atmosféricos classificados como desastres. Mas não o são o nascer e pôr do sol, ou a condensação. Tremores de terra, inundações e actividade vulcânica são catástrofes relacionadas com fenómenos geológicos, mas o desvio continental e a sedimentação já não.

Para alterar um conceito, basta editar um ficheiro, e a alteração tomará efeito logo que a aplicação corra. Isso significa que o refinamento dos resultados faz parte do processo de definição dos conceitos. Os erros são muito mais inconsequentes do que no caso de anotação do corpus para efeitos de treino, operação esta medida em horas ou dias.

Visto que se detectam mais rapidamente problemas relacionados com falsos positivos do que com falsos negativos — por outras palavras: nota-se mais facilmente algo que não devia estar presente do que a ausência de alguma coisa — recomenda-se que, em caso de dúvida, se opte pelo sentido mais geral (de acordo com os temas tratados e a especialização do texto), e se

[3] O aumento é linear. O número de casos vistos é dado por $C \times P$, onde C é o número de *synsets* dos conceitos e P é o número de *synsets* das palavras a analisar.

Daí que seja possível que uma frase com estas três palavras fique associada a "pessoa" e não a "animal". A fim de evitar esse tipo de surpresas, será preferível pedir ao utilizador para especificar qual ou quais *synsets* procura usar. Existe uma relação sobrejectiva que o WordNet mantém, e que serve mesmo para ajudar a desambiguar e ilustrar o uso de um *synset*. É fácil usar esta informação para guiar o utilizador na selecção dos conceitos.

Os conceitos são definidos da seguinte forma, num ficheiro:

```
% Shots of the interior of a court-room location
concept(court,103120778).
concept(court,103649459).
```

Onde os dois *synsets* correspondem a

court, courtroom: a room in which a lawcourt sits "television cameras were admitted in the courtroom"

court, lawcourt, court of law, court of justice: a tribunal that is presided over by a magistrate or by one or more judges who administer justice according to the laws

O WordNet fornece a relação `s/6`, que relaciona os *synsets* com palavras, e a relação `g/2`, que relaciona o *synset* com a descrição do seu significado. Estes dois predicados são instrumentais nesta fase.

3.1.1 Escolha lata

Por vezes existem significados próximos que se querem também usar como sendo referentes a um conceito, quer por fácil associação de ideias, em virtude de estarem intimamente relacionados, quer por serem visualmente relevantes. Por exemplo, o conceito "vegetation" descreve-se pelos *synsets* vegetação/flora ("todas as plantas numa região ou período"), planta/flora ("um organismo desprovido da capacidade de locomoção") e verdura ("folhagem verde").

No exemplo atrás, o segundo *synset* não corresponde directamente à sala do tribunal, mas ao tribunal em si. Enquanto o primeiro é algo de físico e concreto, o segundo é uma assembleia, que é um organismo social e abstracto.

Também o conceito "urban" não se revelava muito nos resultados, sendo considerado sempre irrelevante nos primeiros testes, possivelmente por ser um adjectivo. Associando este conceito também ao *synset* "city" permite obter resultados mais próximos dos desejados.

3.1.2 Abrangência

Ao definir os conceitos, é comum colocar-se uma questão sobre o nível de abstracção a manter. Será preferível escolher um *synset* mais genérico, ou

homógrafas[2] e lidar com os múltiplos sentidos das palavras. Por exemplo: "parto" pode ser o momento de dar à luz, ou uma conjugação do verbo partir, que por sua vez, pode ter o significado de dar início a uma viagem, ou de dividir algo. O uso de palavras de forma figurativa também enriquece muito a linguagem. Por exemplo: "ele é um animal" pode ser interpretado de diversas formas.

Embora não seja impossível de se conseguir fazer o mesmo de forma automática, muitas vezes os contextos não estão bem definidos no momento de enunciar os conceitos.

No caso do TRECVID, vários contextos são acompanhados de descrições simples. Por exemplo, "road" é acompanhado da simples descrição "Shots depicting a road". Outras descrições são um pouco vagas, como "Shots of outdoor locations". Poucas são completas, como "Shots depicting natural or artificial greenery, vegetation woods, etc.". A lista dos conceitos fornecida para os concursos de 2006 e 2007 encontra-se no Anexo A.

Na lista de conceitos do Columbia374 [24], ou Reuters21578 [25], esses contextos são deduzidos apenas como sendo das áreas de noticiário ou económica/financeira, respectivamente. Apesar da maioria ser bastante simples ou directo, no caso do Columbia374 a lista tinha algumas repetições, sendo um ou outro conceito menos claro.

Nestas situações (ou mesmo em caso de dúvida), é comum supor-se que o significado mais comum é o correcto. O WordNet possui um indicador que indicia, para cada palavra, qual a frequência com que é usada com determinado significado (chamado `tag_count`). Acontece que esse indicador, na versão utilizada (3.0), não está presente em 82% dos dados (essa taxa cresce para 89% quando considerados apenas os substantivos).

A dificultar ainda mais a situação está a grande diversidade de aplicações para uma palavra. Por exemplo, a palavra "face", que integrou a lista de conceitos do TRECVID em 2006 e 2007, tem no WordNet 19 *synsets*. 13 destes são substantivos (todos eles comuns). Contam-se entre estes, para além do imediato:

- a face de um penhasco,
- a face de um objecto (como uma carta de jogar),
- a face de uma cidade,
- uma face triste (aparência),
- salvar a face.

Inicialmente a lista de temas era definida como uma lista de palavras. Notou-se depois que diversas palavras têm vários significados que são inesperados. Por exemplo,"cat", "dog" e "rat" são termos chamados a pessoas.

[2]Palavras que partilham a mesma grafia, mas diferem na pronúncia

Capítulo 3

Análise do texto

Neste capítulo são enunciados os passos a percorrer antes de proceder à interpretação do texto.

Inicialmente serão abordados os conceitos. O TRECVID já possui uma lista de 37 conceitos que foram utilizados para anotar os vídeos nos concursos de 2006 e 2007. Estes temas estão já pré-definidos. Também os conceitos presentes no *corpus* Columbia374 (salvo um pequeno número, por motivos expostos) já foram igualmente definidos. Mas para usar esta aplicação noutros contextos, há que começar por definir a lista de temas. Esta tarefa levanta certas questões inesperadas, e que podem ter um grande impacto nos resultados do sistema.

De seguida é explicado o resultado da análise sintáctica feita pelo VISL, e a forma como é traduzida para Prolog.

Por fim, na secção seguinte, é explicado como esta informação é tratada, de forma a ser relacionável com os conceitos escolhidos. São apresentadas as palavras mais relevantes, e quais os obstáculos e erros de interpretação que se interpõem.

3.1 Definição dos conceitos

Frequentemente — como acontece no TRECVID — os conceitos já estão definidos. Este é o caso mais comum. Alguns sistemas avançados conseguem escolhem os conceitos usados para catalogar textos, não havendo limite pré-estabelecido. Esta possibilidade não será abordada aqui.

O propósito desta tarefa consiste em representar estes conceitos numa forma livre dos equívocos da língua natural, ou seja, defini-los como *synsets* do WordNet.

Habitualmente, o recurso ao contexto permite esclarecer ambiguidades de vocabulário, evitando confusões provocadas por palavras homónimas[1],

[1] Palavras que partilham a mesma grafia e pronúncia

Se um animal doméstico é um animal, e o cão é um animal doméstico, então cão é um animal.

```
[entity, physical entity, physical object, whole,
living thing, organism, animal, domestic animal, house cat]

[entity, physical entity, physical object, whole,
living thing, organism, plant, vascular plant, fern]
```

Cão e gato são animais domésticos. Cão e feto são organismos vivos. Assim observa-se que é possível obter informações de afinidade entre palavras sem ter de percorrer todo o grafo.

O código que calcula os caminhos é bastante simples. Encontra-se no Anexo B.

Uma palavra pode ter mais de um caminho. Isto deve-se aos seguintes factos:

- Uma palavra pode ter mais de um *synsense*. Por exemplo, "face", como foi visto na página 17;
- Um *synsense* pode ter mais de um "nó pai". Por exemplo "sleepwalking", que tem um caminho por "sleep" e outro por "walking".

Tendo um conjunto de caminhos para um conceito, e um conjunto de caminhos para uma palavra, é trivial comparar todos os pares, e determinar quais são os mais semelhantes. À partida, será aquele que detecta o maior prefixo igual.

As vantagens e desvantagens relacionadas com o uso destes caminhos são diversas. A saber:

É um algoritmo simples e rápido. Pesquisar "às cegas" num grafo implica tratar todas as "direcções" (hiponímia/hiperonímia, meronímia/holonímia, é instância de/tem instância) indiscriminadamente. Este algoritmo efectua uma busca dirigida[2] à raiz, e consegue encontrar todos os caminhos "subindo" sempre de nível.

Não garante o caminho mais curto entre duas palavras. No entanto, a associatividade entre duas palavras não está directamente relacionada com o número de nós entre si. Por exemplo, no WordNet, à distância de 2 da palavra "água" tem-se "instalação", "nutriente", "coisa", "elemento" e "composto binário", entre outros. Mas observando a três passos de especialização (ou seja, como descendentes) pode-se encontrar "água engarrafada", "canal", "maré" ou "curso de água". Tudo mais identificável com a a palavra original.

[2] A busca é efectuada em profundidade, como é típico no Prolog. Existe um ciclo no grafo, entre os verbos "restringir" e "inibir", que já foi denunciado como *bug*. No entanto, visto não se estar a trabalhar com verbos, este problema não afecta a presente solução.

substantivo	nº	descrição
sailboat		a small sailing vessel; usually with a single mast
lake	1	a body of (usually fresh) water surrounded by land
	2	any of numerous bright translucent organic pigments
	3	a purplish red pigment prepared from lac or cochineal
sky		the atmosphere and outer space as viewed from the earth
cloud	1	a group of many things in the air or on the ground; "a swarm of insects obscured the light"; "clouds of blossoms"; "it discharged a cloud of spores"
	2	a visible mass of water or ice particles suspended at a considerable altitude
	3	any collection of particles (e.g., smoke or dust) or gases that is visible
	4	out of touch with reality; "his head was in the clouds"
	5	suspicion affecting your reputation; "after that mistake he was under a cloud"
	6	a cause of worry or gloom or trouble; "the only cloud on the horizon was the possibility of dissent by the French"

Tabela 4.1: Significados das palavras

Apresenta muita informação. Talvez mais importante do que a forma como se encontra a informação, é o que fazer com ela. O resultado deste algoritmo retorna, para além do caminho entre as palavras, a "profundidade" a que se encontra cada uma (o tamanho do caminho para a atingir), e quanto em comum têm as palavras (indicada pelo início comum dos caminhos).

Permite optimizações Pré-calculando os caminhos dos conceitos (que são conhecidos desde o início). É também possível reutilizar alguns caminhos. Por exemplo, se o caminho de "água" já tiver sido determinado, será recordado caso surja para ser calculado no futuro. "maré" é hipónimo de "água". Visto o cálculo de caminhos ser feito recursivamente (como visto no código atrás), o cálculo do caminho do *synset* de "maré" é este *synset* conjunto com o caminho do *synset* de água. Ora, este foi já calculado. Isto é conhecido como "memoization". [4]

Exemplo Os significados dos substantivos das frases surgem na Tabela 4.1. "sailboat" e "sky" têm um *synset* apenas no WordNet, "lake" tem 3 e "clouds" tem 6.

Nota-se que dois dos três sentidos de "lake" são muito semelhantes.[3].

[3]Um reparo para as semelhanças entre os caminhos de dois significados de lake. Uma

Os caminhos encontrados para cada um destes significados surgem na Tabela 4.2

consulta à Wikipedia [31] para dissipar as dúvidas revela o que deve ser uma pequena lacuna na organização do WordNet. Possivelmente, o pigmento avermelhado deveria de ser chamado "carmine lake", e descender da família de pigmentos orgânicos. Ainda assim, a lacuna é pequena.

substantivo	nº	rel	caminho
sailboat		hyp	entity, physical entity, object, whole, artifact, instrumentation, transport, vehicle, craft, vessel, sailing vessel, sailboat
lake	1	hyp	entity, physical entity, thing, body of water, lake
	2	hyp	entity, abstraction, relation, constituent, substance, material, coloring material, pigment, lake
	2	hyp	entity, physical entity, matter, substance, material, coloring material, pigment, lake
	3	hyp	entity, abstraction, relation, constituent, substance, material, coloring material, pigment, lake
	3	hyp	entity, physical entity, matter, substance, material, coloring material, pigment, lake
sky		hyp	entity, physical entity, matter, fluid, gas, atmosphere, sky
		mp	earth, sky
cloud	1	mp	earth, sky, cloud
	2	hyp	entity, physical entity, process, phenomenon, natural phenomenon, physical phenomenon, atmospheric phenomenon, cloud
	3	hyp	entity, physical entity, process, phenomenon, natural phenomenon, physical phenomenon, cloud
	4	hyp	entity, abstraction, attribute, state, nonbeing, nonexistence, unreality, cloud
	5	hyp	entity, abstraction, attribute, state, hostility, suspicion, cloud
	6	hyp	entity, abstraction, attribute, state, condition, atmosphere, gloom, cloud

Tabela 4.2: Caminhos dos vários significados das palavras

O aspecto que mais se destaca é a predominância das relações de hiponímia, o que se compreende, dado que representam cerca de 3/4 do total das relações.

Observando os caminhos, começa a ser aparente como certos significados para a mesma palavra são distintos, e como os caminhos o reflectem. As nuvens como fenómenos atmosféricos são entidades físicas, enquanto nuvens como aglomeração ou metáforas são entidades abstractas. Os pigmentos conhecidos como "lake" estão presentes em ambos os ramos: de forma abstracta, como a cor, e de forma física, como a substância usada para lhe dar origem.

4.1.2 Definição de caminho comum e caminho divergente

Define-se como *caminho comum* o maior início idêntico entre dois ou mais caminhos. O final de um caminho que seja complementar ao caminho é denominado *caminho divergente.*

O caminho comum é usado para dar uma noção de *profundidade* da semelhança. Com o aumento da profundidade, presume-se a existência de maior número de propriedades partilhadas entre os sentidos das palavras comparadas.

O caminho divergente é usado para dar uma noção de distância. Assumindo o aumento da distância, deduz-se que há mais propriedades que não são partilhadas entre os sentidos das palavras. Sublinhe-se que estes dois tipos de caminho não são antónimos. Os *synsets* que pertençam ao caminho comum são de *mais alto nível*, enquanto os caminhos divergentes são relativos aos aspectos mais concretos do caminho.

Há que ter sempre presente que a profundidade (número de elementos num caminho) é sempre muito relativa. Isto porque não existe qualquer garantia ou intenção de fazer com que as relações sejam "isométricas". Quer isto dizer que a distância entre sinónimos não tem necessariamente que reflectir directamente a afinidade entre os mesmos. Pode, norma geral, ser visto como um *indicador* do mesmo, e em conjunto com outras medidas, dar uma noção *relativa* do grau de especialização do mesmo termo.

Por exemplo, observando a Figura 4.1, nota-se que o Leopardo e o Tigre têm muita coisa em comum. São ambos felinos que, por exemplo, conseguem rugir, ao contrário do Puma que não tem essa habilidade, e é classificado numa categoria diferente pelos biólogos (Figura 1.1). No entanto, todos são igualmente felinos, mesmo que Puma esteja mais afastado desta classe do que as outras duas espécies.

Da análise da Figura 1.1 (página 7), infere-se que um tigre e um leopardo têm grande afinidade, ao passo que um narval e um manatim não têm em comum senão o serem mamíferos. Na Figura 4.1 estão patentes a grande afinidade entre o leopardo e tigre (um pouco maior devido ao Placental), e a algo significativa afinidade entre o narval e o manatim, pois são ambos

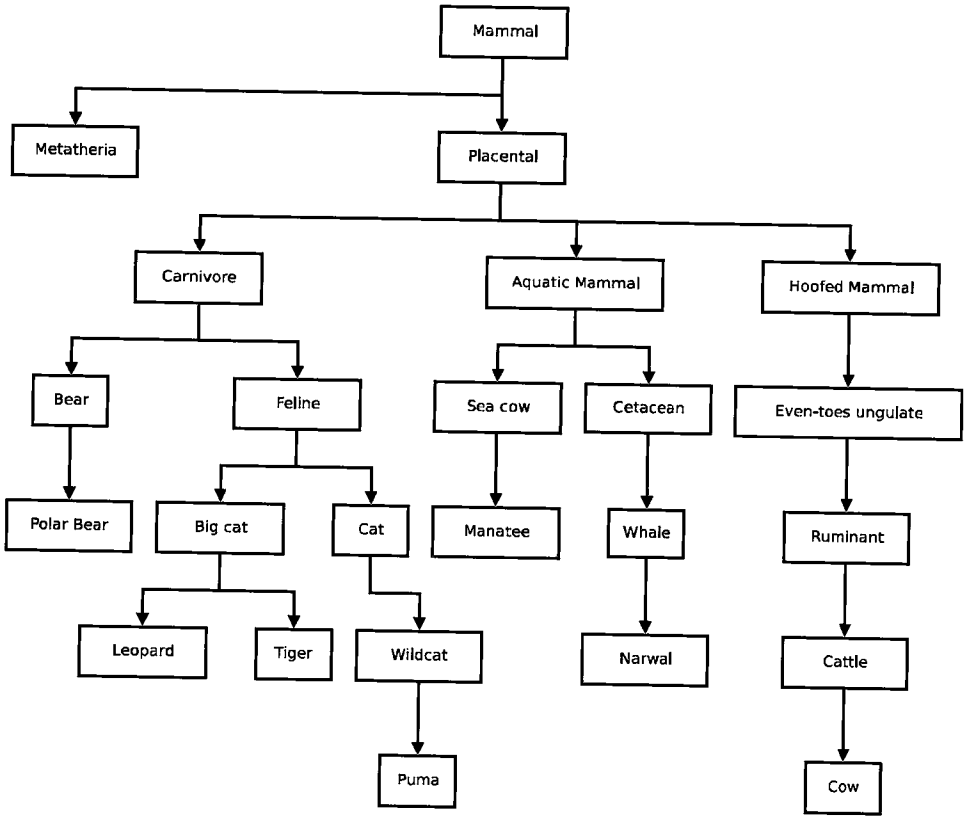

Figura 4.1: Excerto da árvore de hipónimos do WordNet

mamíferos aquáticos com placenta.

Exemplo Vejam-se as comparações entre algumas palavras e conceitos na página 38.

Da Tabela 4.3, infere-se a proximidade entre "boat" e "sailboat". Quer o barco, quer o veleiro são embarcações, sendo que um está na categoria referente a "embarcação mastreada". Ambos partilham um longo caminho comum, tendo ambos caminhos divergentes pequenos.

De seguida apresentam-se duas comparações de "lake" com "waterscape". O primeiro sentido de "lake" (referente à extensão de água), como se nota na Tabela 4.4, tem um pequeno caminho divergente, enquanto que "waterscape" (um conceito que nesta instância assume o significado de corpo de água —

> the part of the earth's surface covered with water (such as a river or lake or ocean); "they invaded our territorial waters"; "they were sitting by the water's edge"

— não possui qualquer caminho divergente. Daí poder dizer-se que "body of water" e "lake" são colineares, uma vez que o caminho de um deles é o início do caminho do outro.

Por fim, repete-se o mesmo exemplo, mas usando outro *synset* de "lake". A Tabela 4.5 mostra que os caminhos divergentes representam uma parte já significativa dos caminhos de cada um dos significados. Seria ainda pior se se tentasse o mesmo com um *synset* abstracto. Assim se ilustra que os caminhos podem ser usados para aferir se dois sentidos se referem a coisas semelhantes ou pouco relacionáveis.

palavra	nº	caminho comum	caminho divergente
sailboat		entity, physical entity, object, whole, artifact, instrumentation, transport, vehicle, craft, vessel	sailing vessel, sailboat
boat		entity, physical entity, object, whole, artifact, instrumentation, transport, vehicle, craft, vessel	boat

Tabela 4.3: Comparação dos caminhos "sailboat" e "boat"

palavra	nº	caminho comum	caminho divergente
lake	1	entity, physical entity, thing, body of water	lake
waterscape		entity, physical entity, thing, body of water	

Tabela 4.4: Comparação dos caminhos "lake" (lago) e "waterscape"

palavra	nº	caminho comum	caminho divergente
lake	2	entity, physical entity	matter, substance, material, coloring material, pigment, lake
waterscape		entity, physical entity	thing, body of water

Tabela 4.5: Comparação dos caminhos "lake" (pigmento) e "waterscape"

4.2 Função de pontuação e sua evolução

Tudo o que se relaciona com a informática necessita de ser quantificado. Os termos difusos como "próximo", "afim" e "semelhante" necessitam de ser transformados em conceitos mensuráveis, a fim de poder ter qualquer utilidade no presente âmbito. Pretende-se, agora, mostrar a forma como tal foi conseguido.

Como foi indicado na Secção 4.1, a primeira abordagem consistia na distância entre nós no grafo (número mínimo de saltos para ir de um nó até ao outro). Quer isso dizer que a semelhança entre um tigre e um leopardo é análoga (em envergadura) àquela que existe entre um primata e um carnívoro. Isto está claramente incorrecto.

O erro reside na não utilização da informação do caminho comum. Cada nó que é partilhado entre dois caminhos significa mais um conjunto (mais ou menos significativo) de propriedades idênticas entre os dois *synsets*.

A profundidade é, depois, relacionada com os caminhos divergentes, que representam a distância entre os nós. Esta é medida como o total de elementos nestes caminhos.

A Figura 4.2 é uma representação simplificada de situações comuns que se encontram no WordNet. Nesta, as letras A e B identificam dois *synsets* que se pretendem comparar. Em ordem crescente de afinidade entre os termos encontram-se:

i. A e B são adjacentes (relação pai-filho). Esta é a segunda melhor situação, logo a seguir a $A = B$ (nesse caso o valor de I é óptimo). Ser colinear é também bom, visto que se pode considerar que apenas um dos *synsets* se afasta e diverge. B é, desta forma, uma situação mais específica de A;

ii. A e B distam 2 (irmãos). Esta situação é facilmente aceitável. Pode ser interpretada como sendo elementos próximos;

iii. A e B divergem de forma assimétrica. A é um conceito de nível superior a B. A penalização desta situação vai depender mais do tamanho do caminho divergente de A, já que se considera este como uma medida do desvio da situação em que seriam colineares — seria mais fácil obter a colinearidade modificando A do que B;

iv. A e B estão bastante afastados. Possivelmente não estarão relacionados.

Relembrando o que foi dito sobre o caminho comum, observando as relações da Figura 4.2, e comparando a sua aplicação às Figuras 4.1 e 1.2, é imediato que faz toda a diferença se A for "Leopard" ou se for "Object". Quanto mais alto o nível (quanto menor o caminho comum), mais grosseiro é o erro.

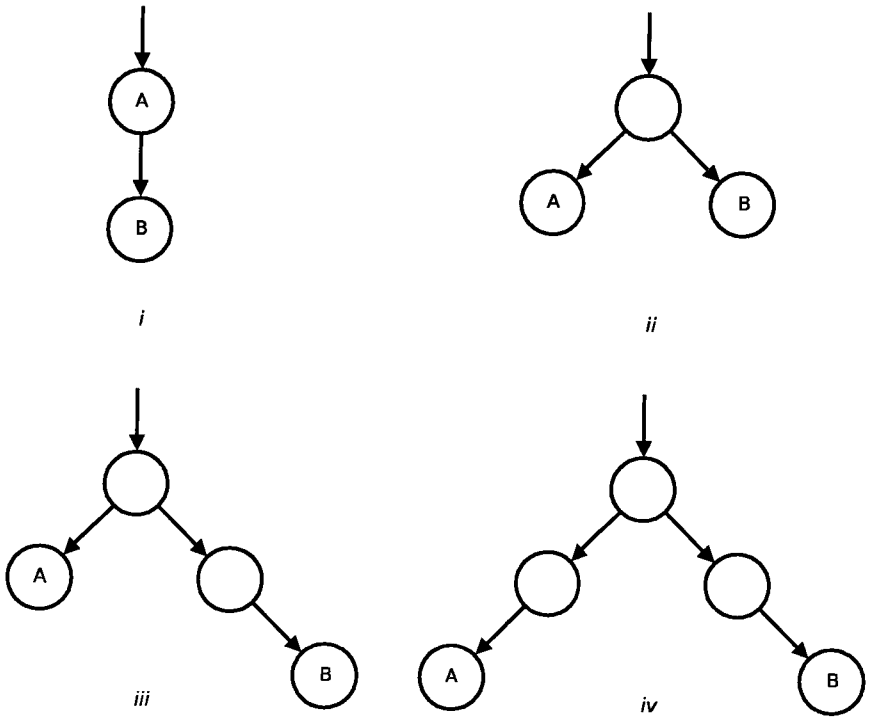

Figura 4.2: Quatro situações básicas

Primeira abordagem Foi com base nestas considerações que se começou a esboçar a forma de comparar dois *synsets* a partir dos seus caminhos.

Sendo A e B os caminhos disjuntos de dois *synsets*, e M o número de elementos no caminho comum, foram elaboradas algumas abordagens para determinar I, o índice de diferença. Este valor é menor para significados semelhantes, e portanto, procura-se o menor I para o conjunto de combinações de *synsets* de conceitos e *synsets* de palavras.

Procura-se uma fórmula que beneficie este índice se M for alto, e o prejudique se A e/ou B forem altos.

Começou-se com a equação

$$I = (A + B)^2 - M + min(A, B) \tag{4.1}$$

que aplica os princípios acima descritos. Optou-se por usar uma penalização quadrática nos caminhos divergentes. Tal efeito faz realçar mais o benefício da profundidade[32]. O mínimo é introduzido para distinguir a situação *avô-neto* $((0 + 2)^2 + M = 4 + M)$ da situação *irmãos* $((1 + 1)^2 + M = 4 + M)$. Penalizando ambas as situações com base no menor dos caminhos divergentes, dá-se preferência à colinearidade.

Recorrendo à Figura 4.1, nota-se que é preferível definir um leopardo a partir de um felino, onde apenas se acrescentam propriedades, do que a

partir de um tigre, onde será necessário removê-las primeiro, para obter um "Big cat".

Acontece que não se pode tratar os conceitos e as palavras analisadas de forma intermutável. Olhando a Figura 4.1, assume-se que existe o conceito "Felino". Se um texto menciona "gato", então é certo que cai dentro do âmbito do conceito, e que esse classificador será válido. Se a palavra encontrada fosse "mamífero", o contexto manteria a sua relevância, embora em menor grau. Se a palavra fosse "cetáceo", a relevância seria ainda mais reduzida.

Deste exemplo retira-se que, nas situações em que os *synsets* são colineares, faz diferença se o conceito é mais genérico ou específico que o *synset* em causa. Daí a seguinte equação foi esboçada:

Sejam C e P caminhos dos *synsets* de um conceito e uma palavra; e c, p e m o tamanho dos seus caminhos divergentes e caminho comum.

$$I = \begin{cases} -100 & \text{se } C = P \\ -50 & \text{se } C \subset P \\ 8 \min(c,p)^2 + \max(c,p) - m & \text{c. c.} \end{cases} \tag{4.2}$$

Dizer que $C \subset P$ é o mesmo que dizer que C é início de P, visto que "Entidade" não surge como subordinado em qualquer relação.

Observam-se dois casos especiais nesta função. Tal deve-se ao reconhecimento da importância dos dois casos particulares, já antes apresentados. O valor -50 poderia ser substituído por uma função específica para lidar com casos colineares, e fornecer valores dependentes do tamanho do caminho divergente da palavra. Na presente situação, esta constante foi usada porque os conceitos são muito diferentes, e nunca partilham porções significativas dos seus caminhos. De outra forma, seria conveniente ter um mecanismo que indicasse qual dos conceitos de nível superior estaria mais próximo. Daí a distinção de pontuação na hipotética função.

De resto, esta função I partilha os mesmos princípios que tinham sido definidos anteriormente. A constante 8 no último ramo da função não tem qualquer valor especial, servindo apenas para conferir um peso superior ao caminho divergente mais curto, como foi justificado anteriormente.

4.3 Comparar os caminhos

Para cada par ordenado (*synset de conceito*, *synset de palavra*), produz-se a seguinte estrutura:

- O nome do conceito,
- O *synset* associado ao conceito,
- A palavra que mais contribuiu para a eleição do conceito,
- O *synset* desta palavra,

substantivo	palavra	nº	m	c	p	pontuação
boat	sailboat		10	1	2	$(8 \cdot 1^2 + 2 - 10)$ 0
waterscape	lake	1	4	0	1	-50
waterscape	lake	2	2	2	6	$(8 \cdot 2^2 + 6 - 2)$ 36

Tabela 4.6: Pontuação de alguns caminhos

- A pontuação associada ao par dos dois *synsets*,
- Os caminhos dos *synsets*.

Da lista produzida para cada palavra (vários synsets), mantém-se apenas aquela cujo *synset* tem maior afinidade com o conceito em causa. De lembrar que uma palavra poderá contribuir para mais de um conceito. Os conceitos são também filtrados, permitindo que persista apenas uma referência de cada um. O resultado reflecte, para cada conceito (pois todos estão presentes), qual a palavra que mais contribuiu para a sua relevância, e com que pontuação.

Este processo é executado para a lista de palavras afirmadas e para a lista de palavras negadas.

Exemplo Na Tabela 4.6 encontram-se os resultados das pontuações referentes aos três exemplos apresentados nas Tabelas 4.3, 4.4 e 4.5, usando os símbolos definidos na Equação 4.2.

A Equação 4.2 consegue traduzir numericamente a noção de proximidade entre *synsets* de forma bastante satisfatória. É imediata a diferença entre as duas interpretações de "lake" apresentadas.

4.4 Eleição dos melhores conceitos

Após a pontuação de todos os pares ordenados $(Conceito, Palavra)$, resta eleger os melhores. Isto não é uma tarefa de difícil execução. A dificuldade encontra-se na definição de "melhores".

É simples dizer que "os melhores são aqueles considerados superiores em relação ao resto", mas isso não ajuda a definir *quantos* eleger para este conjunto restrito, já que saber *quais* incluir é tarefa fácil.

A primeira abordagem recaiu sobre a ideia de que frases mais longas têm maior probabilidade de caírem sobre mais conceitos do que frases pequenas. Daí surgiu a seguinte equação:

$$M = \text{round}(P/2) \tag{4.3}$$

em que M representa o número de conceitos a eleger, e P é o número de substantivos analisados. Esta equação define que é eleito um conceito por

cada duas palavras (arredondando para cima). É sempre garantido pelo menos um conceito, por forma a salvaguardar um resultado no caso de frases simples.

Infelizmente, esta solução não resulta em todas as situações. Certas palavras invocam mais de um conceito (por exemplo, "polícia" implica os conceitos "polícia" e "pessoa").

Surge então a necessidade de definir o número de conceitos com base na adequação ou mérito destes. Claro que esta é uma noção relativa. Por exemplo,"Puma" poderá relacionar-se com o conceito "Ser vivo", mas se houver também um conceito "Gato", o anterior perderá muita relevância.

Assim, foram elaborados os seguintes critérios para a eleição dos melhores conceitos:

1. Todos os conceitos iguais ou ascendentes directos de uma palavra são seleccionados (ou seja, possuem uma pontuação inferior ou igual a -50).
   Estes são sempre considerados bons candidatos;

2. À exclusão dos elementos seleccionados pelo ponto anterior, é determinado o conceito com pontuação mais baixa. Esta determina o valor de corte I;

3. Se $I < 0$, então todos os conceitos com pontuação inferior a $I/2$ são seleccionados;

4. Se $I = 0$, então todos os conceitos com pontuação igual a zero são seleccionados;

5. Se $I > 0$, é verificado se já foi seleccionado algum conceito no primeiro ponto (selecção directa).
   Se assim for, pode-se ser mais exigente e seleccionar os elementos com pontuação (positiva, claro) inferior a 2 (possivelmente nenhum).
   Caso contrário, nenhum dos conceitos é considerado como "adequado". Mas, a fim de garantir um resultado, são seleccionados os conceitos com a pontuação mínima de entre os presentes (I).

Não é executada qualquer tentativa de resolver conflitos entre os resultados da lista de afirmações e da lista de negações, ou seja, se um conceito surgir em ambas, não será removido de nenhuma.

No TRECVID é raro surgirem frases negadas, e para o bom processamento de textos mais correntes, o tratamento da negação deverá ser melhor delineado e mais avançado.

Mas visto que ambas as listas apresentam todos os resultados pontuados, no caso em que o mesmo conceito surge repetido, será fácil ignorar ou eliminar o pior, preferir o falso positivo ao falso negativo, ou optar por outra solução (isto é, dar preferência à afirmação).


```
<DOC>
  <Frase id="Nome do ficheiro">
    <Sim Conteudo="Lista de palavras afirmadas">
      <Lista_exactos>
        <Conceito>
          <Nome>Nome do conceito</Nome>
          <Pontuacao>Pontuação atribuída</Pontuacao>
          <Origem Significado="Significado">Palavra</Origem>
        </Conceito>
        <Conceito>
          ...
        </Conceito>
      </Lista_exactos>
      <Lista_genericos>
        <Conceito>
          ...
        </Conceito>
      </Lista_genericos>
      <Lista_difusos>
        <Conceito>
          ...
        </Conceito>
      </Lista_difusos>
    </Sim>
  </Frase>
    <Nao Conteudo="Lista de palavras negadas">
      ...
    </Nao>
</DOC>
```

Os *tags* `Conceito` surgem o número de vezes necessário dentro dos *tags* `Lista_exactos`, `Lista_genericos` e `Lista_difusos`. Estes três correspondem, por ordem decrescente de segurança, aos ramos da Equação 4.2.

O *tag* `Origem` é usado para efeitos de *debugging*, para confirmar qual a fonte de tal pontuação, o que devido à riqueza da língua inglesa, nem sempre é imediato.

Por fim, o *tag* `Nao` repete a mesma estrutura do *tag* `Sim`. Se não houver palavras negadas (o caso mais comum), esta tag é omitida, de forma a simplificar o *output*.

Exemplo Do exemplo que tem vindo a ser usado de forma ilustrativa, apesar de não terem sido apresentados todos os resultados, a melhor pontuação superior a -50 é 0.

Daí que os conceitos seleccionados sejam: "sky" (com -100 pontos), "waterscape" (-50 pontos) — ambas de forma automática — e "boat" (0 pontos) da lista das palavras afirmadas.

A lista de palavras negadas tem apenas dois elementos seleccionados: "disaster" e "sky".

Como referido atrás, a presente implementação recusa-se a tomar qualquer decisão sobre a exclusão de termos de qualquer uma das listas nesta situação, deixando que o conceito "sky" permaneça em ambas as frentes.

A referência a "disaster" a partir da palavra "cloud" parece um pouco inesperada. A relação céu–nuvem é bastante clara, e pode ser observada na Tabela 4.2, na relação de meronímia: as nuvens fazem parte do céu. Mas a relação com desastres convém ser vista.

Como referido na Secção 3.1.2, optou-se por usar o sentido "atmospheric phenomenon" no conceito "disaster". Sendo nuvens um descendente (directo) de fenómenos atmosféricos — como surge indicado na Tabela 4.2 — o conceito a que esse sentido pertence é imediatamente seleccionado por ser uma noção mais genérica que aquela em causa.

4.5 Apresentação dos resultados

Os resultados são apresentados num ficheiro XML. O seu formato (a título ilustrativo) é

Exemplo O output final é o seguinte:

```
<DOC>
  <Frase id="ex.txt»
    <Sim Conteudo="[sailboat, lake, sky]»
      <Lista_exactos>
        <Conceito>
          <Nome>sky</Nome>
          <Pontuacao>-100</Pontuacao>
          <Origem>sky</Origem>
        </Conceito>
      </Lista_exactos>
      <Lista_genericos>
        <Conceito>
          <Nome>waterscape</Nome>
          <Pontuacao>-50</Pontuacao>
          <Origem>lake</Origem>
        </Conceito>
      </Lista_genericos>
      <Lista_difusos>
        <Conceito>
          <Nome>boat</Nome>
          <Pontuacao>0</Pontuacao>
          <Origem>sailboat</Origem>
        </Conceito>
      </Lista_difusos>
    </Sim>
      <Nao Conteudo="[cloud, cloud]»
        <Lista_genericos>
          <Conceito>
            <Nome>disaster</Nome>
            <Pontuacao>-50</Pontuacao>
            <Origem>cloud</Origem>
          </Conceito>
          <Conceito>
            <Nome>sky</Nome>
            <Pontuacao>-50</Pontuacao>
            <Origem>cloud</Origem>
          </Conceito>
        </Lista_genericos>
    </Nao>
  </Frase>
</DOC>
```

Capítulo 5

Análise do desempenho

Atenta a inexistência de qualquer grupo de resultados definido como "correcto" para os resultados do TRECVID, cabe a cada equipa de desenvolvimento a responsabilidade de definir quais os resultados esperados para a sua aplicação. Acontece que é com base nesses objectivos muito individuais que os projectos são orientados, inviabilizando qualquer possibilidade de comparação de resultados de forma directa. Muitas das vezes as escolhas do sistema não podem ser consideradas *más*, e consegue-se encontrar com facilidade uma linha lógica que a justificaria.

5.1 Metodologia

O sistema de teste é um Intel Pentium 4 a 2.8GHz, com 1GB de RAM, correndo o Sistema Operativo Debian GNU/Linux.

Foi elaborado um sript que corre os testes sobre as *queries* do TRECVID 2006. Cada teste é corrido 5 vezes. São excluídos os valores extremos, e é calculada a média dos tempos restantes.

Os tempos apresentados não incluem a análise sintáctica, que é elaborada via Web, e como tal depende de condições impossíveis de reproduzir.

A unidade de tempo utilizada é o segundo (s).

Os testes foram corridos usando as *queries* do TRECVID 2006.

5.2 Determinar o overhead

Executar o programa sem pedir para analisar qualquer texto leva 4,154s a correr.

Isto inclui o tempo de carregamento do interpretador SWI Prolog, e carregar a aplicação e os dados do WordNet.

Logo, este valor indica o tempo total de "arranque" da aplicação, e é o tempo mínimo possível para qualquer um dos conjuntos de testes executados.

5.3 Informação sobre os testes do TRECVID

As *queries* do TRECVID 2006 estão apresentadas na Tabela 5.3.

Na Tabela 5.3 constam os substantivos tirados de cada texto, junto com os conceitos eleitos por uma pessoa, contra as quais os resultados dos testes são medidos.

Como dito atrás, estas escolhas são subjectivas, e pessoas diferentes fariam escolhas diferentes. Pode ser debatida a presença de alguns conceitos nesta lista, como "policeman" na frase 177 (é comum surgirem nas coberturas jornalísticas, mesmo que hajam apenas 3 ou 4 na proximidade), ou "us_flag" nas frases 178, 181 e 194 (nos discursos políticos surgem as cores da bandeira), ou "outdoors" na frase 183 (é muito possivelmente uma cena no exterior). Estes termos não conseguem surgir apenas por análise semântica. À partida, não irão ser reconhecidos nestes testes. Mas dado que se deveriam efectuar os testes da perspectiva da etapa seguinte de processamento (aquilo que seria preferível ter no input), eles estão presentes.

Texto	conteúdo
173	Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)
174	Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible
175	Find shots with one or more people leaving or entering a vehicle
176	Find shots with one or more soldiers, police, or guards escorting a prisoner
177	Find shots of a daytime demonstration or protest with at least part of one building visible
178	Find shots of US Vice President Dick Cheney
179	Find shots of Saddam Hussein with at least one other person's face at least partially visible
180	Find shots of multiple people in uniform and in formation
181	Find shots of US President George W. Bush, Jr. walking
182	Find shots of one or more soldiers or police with one or more weapons and military vehicles
183	Find shots of water with one or more boats or ships
184	Find shots of one or more people seated at a computer with display visible
185	Find shots of one or more people reading a newspaper
186	Find shots of a natural scene - with, for example, fields, trees, sky, lake, mountain, rocks, rivers, beach, ocean, grass, sunset, waterfall, animals, or people; but no buildings, no roads, no vehicles
187	Find shots of one or more helicopters in flight
188	Find shots of something burning with flames visible
189	Find shots of a group including least four people dressed in suits, seated, and with at least one flag
190	Find shots of at least one person and at least 10 books
191	Find shots containing at least one adult person and at least one child
192	Find shots of a greeting by at least one kiss on the cheek
193	Find shots of one or more smokestacks, chimneys, or cooling towers with smoke or vapor coming out
194	Find shots of Condoleeza Rice
195	Find shots of one or more soccer goalposts
196	Find shots of scenes with snow

Tabela 5.1: Lista das *queries* TRECVID 2006

Texto	Lista de palavras (substantivos encontrados)	Conceitos "correctos"
173	emergency, vehicle, motion, ambulance, police, car, truck	car, truck, road, boat
174	view, building, story, top, story	building, urban, outdoors
175	people, vehicle	person, car, bus, truck, boat, airplane
176	soldier, police, prisoner	policeman, prisoner, military
177	daytime, demonstration, protest, part, building	crowd, building, marching, person, policeman
178	vice, President	face, person, us_flag
179	Saddam, Hussein, person, face	face, person, military
180	people, uniform, formation	military, policeman, marching
181	President, George, W, Bush, Jr	face, person, walking, us_flag
182	soldier, police, weapon, vehicle	military, policeman, car
183	water, boat, ship	waterscape, boat, outdoors
184	people, computer, display	person, screen
185	people, newspaper	person
186	scene, field, tree, sky, lake, mountain, river, beach, ocean, grass, sunset, waterfall, animal, people	outdoors, animal, vegetation, mountain, sky, waterscape
187	helicopter, flight	airplane, sky
188	flame	explosion_fire
189	group, people, suit, flag	meeting, us_flag
190	person, book	person, office
191	person, child	person
192	greeting, kiss, cheek	person, face
193	smokestack, chimney, tower, smoke, vapor	sky, building
194	rice	person, face, us_flag
195	soccer, goalpost	sport, vegetation, crowd
196	scene, snow	snow, outdoors

Tabela 5.2: Respostas desejadas para as *queries* TRECVID 2006

5.4 O código normal

Este teste mostra o desempenho da versão corrente do analisador. Todos os testes seguintes são modificações deste código, a fim de destacar apenas a funcionalidade em questão.

O presente teste demorou em média 700.226s a correr.

Os valores, presentes na Tabela 5.3, em geral, podem ser considerados bons, exceptuando três frases que resultaram em zeros.

A frase 180 falha pelos seguintes motivos: a relação entre "people" e "person" é um pouco problemática. O primeiro é um conceito abstracto (um conjunto de pessoas), enquanto que o segundo é concreto. "Uniform" está associado a roupa. Só com alguma dificuldade consegue ser feita a ligação entre "military uniform" e "military" (pode ser feita através do domínio, mas essa relação não está a ser usada actualmente). Por fim, "formation" é associado a "formação rochosa", e daí a montanha.

A frase 194 é mais uma associação visual. "Vapor" fez a ligação com barcos a vapor, o que é compreensível. Mas não se deverá ignorar a ligação entre "chaminé" e "torre" com edifício.

Por fim, com "rice" não é possível fazer algo, já que o nome não é reconhecido devidamente pelo VISL, e como tal é impossível processá-lo da forma correcta..

Em contrapartida, há também 3 respostas perfeitas.

5.5 Sem detectar abreviaturas

Os resultados estão presentes na Tabela 5.4.

O tempo médio de execução é 745.825s.

Os nomes George e Bush são detectados correctamente como nomes próprios. Não é necessário identificar a pessoa, portanto.

A única diferença ocorre no tempo de processamento, que tem de lidar com mais uma palavra.

5.6 1 conceito por cada 2 palavras

O tempo médio de execução foi 722.534s.

Os resultados estão na Tabela 5.5.

A abrangência é reduzida, pois selecciona-se apenas um conceito ou dois por frase. Isso significa menos conceitos eleitos, o que se reflecte na precisão.

5.7 Sem detectar nomes compostos

A duração média de execução foi 788.918s.

Os resultados estão na Tabela 5.6.

Mais uma vez, os conceitos foram bem escolhidos, — apenas diverge na adição de "building" na primeira *query*.

Nota-se que o aumento do número de palavras penaliza o tempo de execução.

5.8 Ignorando negações

O tempo médio de execução foi 695.840s

Os resultados podem ser vistos na Tabela 5.7.

Há uma redução na precisão (devido à frase 186), e nota-se um ligeiro aumento da velocidade. Tal é compreensível, pois deixa de ter de verificar se cada substantivo que lê está negado.

5.9 Usando a função de pontuação anterior

O tempo médio de execução foi 712.666s.

Os resultados constam da Tabela 5.8

Ocorre um ligeiro aumento da precisão, e redução da abrangência.

5.10 Sem usar caminhos pré-calculados

O tempo médio de execução foi 713.658.

O resultado da execução está presente na Tabela 5.9.

Observa-se um pequeno aumento do tempo de execução, devido ao cálculo dos caminhos dos conceitos na primeira execução.

5.11 Sem ignorar palavras inúteis no início da *query*

A execução demorou em média 1080.709.

Os resultados deste teste podem ser vistos na Tabela 5.10.

Nota-se o aumento da abrangência (mais palavras seleccionadas), e redução da precisão. O tempo que demorou o exercício também aumentou bastante, pois mais palavras merecem atenção.

5.12 Execução sem recordar caminhos

O tempo médio de execução foi de 711.842s.

Os resultados podem ser lidos na Tabela 5.11.

Não tem caminhos pré-calculados nem guarda os caminhos. Leva pouco mais tempo relativamente à ausência de pré-cálculo, visto que há menos repetição de palavras.

5.13 Sem usar hipónimos

O tempo médio de execução foi 690.822s.

Os resultados podem ser vistos na Tabela 5.12.

Houve algum ganho de tempo, mas os resultados são melhores que o esperado. Desta forma se vê que os hipónimos não são indispensáveis.

5.14 Usando apenas hipónimos

O tempo médio de execução: foi 499.579, o que é bastante rápido.

Os resultados são apresentados na Tabela 5.13.

Obtém-se um excelente ganho em tempo, o que foi inesperado, visto que os hipónimos são mais que as meronímias. Resultados semelhantes ao teste sem hiponímias.

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking, boat, disaster, airplane	0.750	0.500	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, face, desert	0.833	0.556	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, face	0.500	0.500	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.572	0.581	0.252

Tabela 5.3: Resultados da execução normal

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking, boat, disaster, airplane	0.750	0.500	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, face, desert	0.833	0.556	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, face	0.500	0.500	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.572	0.581	0.252

Tabela 5.4: Resultados da execução sem detectar abreviaturas

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking	0.500	0.667	0.125
174	building, boat, crowd	0.333	0.333	0.333
175	boat	0.167	1.000	0.000
176	prisoner, animal	0.333	0.500	0.000
177	marching, building, snow	0.400	0.667	0.000
178	person	0.333	1.000	0.000
179	person, face	0.667	1.000	0.167
180	mountain, person	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military	0.333	0.500	0.167
183	boat, waterscape	0.667	1.000	0.167
184	screen	0.500	1.000	0.000
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation	0.833	0.714	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation	0.000	0.000	0.000
190	person	0.500	1.000	0.500
191	person	1.000	1.000	1.000
192	animal, face	0.500	0.500	0.000
193	boat, building, disaster	0.500	0.333	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow	0.500	1.000	0.500
Média		0.465	0.717	0.241

Tabela 5.5: Resultados da execução com a estratégia de corte anterior

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, building, walking, boat, disaster, airplane	0.750	0.429	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	person, screen	1.000	1.000	0.500
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, face, desert	0.833	0.556	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, face	0.500	0.500	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.572	0.578	0.262

Tabela 5.6: Resultados da execução sem detectar nomes compostos

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking, boat, disaster, airplane	0.750	0.500	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, road, mountain, building, disaster, person, waterscape, vegetation, boat, truck, car, airplane	0.833	0.385	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, face	0.500	0.500	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.572	0.574	0.252

Tabela 5.7: Resultados da execução sem lidar com palavras negadas

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking	0.500	0.667	0.125
174	building, boat, crowd, meeting	0.333	0.250	0.333
175	person	0.167	1.000	0.167
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner	0.667	0.500	0.167
180	person	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, policeman, boat	0.667	0.400	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, desert	0.833	0.625	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner	0.500	0.500	0.500
191	person, prisoner	1.000	0.500	1.000
192	animal, face	0.500	0.500	0.000
193	boat, building	0.500	0.500	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.548	0.627	0.259

Tabela 5.8: Resultados da execução usando a função de pontuação anterior

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking, boat, disaster, airplane	0.750	0.500	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, face, desert	0.833	0.556	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, face	0.500	0.500	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.572	0.581	0.252

Tabela 5.9: Resultados da execução sem caminhos pré-calculados

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, person, walking, boat, disaster, airplane	0.750	0.429	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, person, snow, outdoor	0.600	0.600	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	person, mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, person, snow	0.667	0.500	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, face, desert	0.833	0.556	0.000
187	person, airplane	0.500	0.500	0.250
188	explosion_fire, person	1.000	0.500	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, person, face	1.000	0.667	0.250
193	boat, person	0.000	0.000	0.000
194	person, vegetation	0.333	0.500	0.333
195	person, sport	0.333	0.500	0.167
196	snow, disaster, person	0.500	0.333	0.500
Média		0.615	0.533	0.259

Tabela 5.10: Resultados da execução sem remover as palavras sem sentido no início da *query*

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking, boat, disaster, airplane	0.750	0.500	0.125
174	building, boat	0.333	0.500	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person, policeman	1.000	0.600	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, animal, prisoner, military, policeman	1.000	0.500	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, face, desert	0.833	0.556	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	animal, face	0.500	0.500	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.572	0.581	0.252

Tabela 5.11: Resultados da execução sem recordar caminhos

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck	0.250	1.000	0.000
174	building, boat	0.333	0.500	0.333
175	person	0.167	1.000	0.167
176	prisoner, policeman	0.667	1.000	0.167
177	marching, building	0.400	1.000	0.000
178	marching, boat, urban, mountain, desert, building	0.000	0.000	0.000
179	person, face, animal	0.667	0.667	0.167
180	person	0.000	0.000	0.000
181	marching, boat, urban, mountain, desert, building	0.000	0.000	0.000
182	policeman	0.333	1.000	0.000
183	boat, waterscape, snow	0.667	0.667	0.167
184	screen, person	1.000	1.000	0.250
185	person	1.000	1.000	1.000
186	animal, sky, mountain, person	0.500	0.750	0.000
187	chart, marching, walking, boat, truck, bus, car, airplane, screen, animal, person, urban, snow, road, mountain, vegetation, desert, building	0.500	0.056	0.000
188	explosion_fire	1.000	1.000	1.000
189	person	0.000	0.000	0.000
190	person	0.500	1.000	0.500
191	person	1.000	1.000	1.000
192	animal, face	0.500	0.500	0.000
193	chart, marching, walking, boat, truck, bus, car, airplane, screen, animal, person, urban, snow, road, mountain, vegetation, desert, building	0.500	0.056	0.000
194	animal, waterscape, road	0.000	0.000	0.000
195	chart, walking, boat, truck, bus, car, airplane, screen, animal, person, urban, snow, road, mountain, vegetation, desert, building	0.33 3	0.059	0.000
196	snow	0.500	1.000	0.500
Média		0.451	0.594	0.219

Tabela 5.12: Resultados da execução sem hipónimos

Query	Conceitos	Abrangência	Precisão	Precisão média
173	truck, car, walking, boat, disaster, airplane	0.750	0.500	0.125
174	building	0.333	1.000	0.333
175	boat, truck, car, airplane	0.667	1.000	0.000
176	prisoner, animal, military, person	0.667	0.500	0.000
177	marching, building, snow, outdoor	0.400	0.500	0.000
178	person	0.333	1.000	0.000
179	person, face, prisoner, military, policeman	1.000	0.600	0.167
180	mountain	0.000	0.000	0.000
181	person	0.250	1.000	0.000
182	animal, military, person, boat, truck, car, airplane	0.667	0.286	0.167
183	boat, waterscape	0.667	1.000	0.167
184	screen	0.500	1.000	0.000
185	marching, vegetation	0.000	0.000	0.000
186	animal, sky, mountain, disaster, person, waterscape, vegetation, desert	0.833	0.625	0.000
187	airplane	0.500	1.000	0.500
188	explosion_fire	1.000	1.000	1.000
189	person, vegetation, flag_us	0.500	0.333	0.000
190	person, prisoner, military, policeman	0.500	0.250	0.500
191	person, prisoner, military, policeman	1.000	0.250	1.000
192	face	0.500	1.000	0.000
193	boat	0.000	0.000	0.000
194	vegetation	0.000	0.000	0.000
195	sport	0.333	1.000	0.333
196	snow, disaster	0.500	0.500	0.500
Média		0.496	0.598	0.200

Tabela 5.13: Resultados da execução só usando hipónimos

5.15 Teste Reuters-21578

Elaborou-se também, a nível de exemplo, um pequeno teste fora do ambiente do TRECVID.

Foram vistos os primeiros 100 documentos presentes no corpus Reuters-21578. Destes, foram escolhidos os textos que estavam associados as temas que eram substantivos traduzíveis facilmente para o WordNet. Destes textos, foram seleccionados 12 textos anotados com temas traduzíveis directamente para o WordNet. Os testes demoraram em média 11331.305s (pouco mais de 3 horas). Os resultados estão apresentados na Tabela 5.14.

O conjunto dos conceitos presentes no sistema correspondem aos temas dos textos seleccionados. Isto é: "cocoa", "grain", "wheat", "copper", "housing", "sugar", "trade", "reserve", "ship" e "grain".

Texto	Tema Reuters	Conceitos	Abrangência	Precisão	Precisão média
01	cocoa	trade, cocoa, soybean	1.000	0.333	0.000
05	grain	sorghum, conr, corn, reserve, wheat, grain, rye, oat	1.000	0.125	0.000
19	wheat	wheat, grain, rye, oat, conr, corn, sorghum	1.000	0.143	1.000
22	copper	copper	1.000	1.000	1.000
29	housing	housing, reserve	1.000	0.500	1.000
42	housing	coffee, oilseed, grain	0.000	0.000	0.000
46	sugar	sugar	1.000	1.000	1.000
47	trade	trade, reserve	1.000	0.500	1.000
48	reserve	soybean, rye, corn, ship, housing	0.000	0.000	0.000
49	ship	reserve, ship	1.000	0.500	0.000
57	grain	conr, corn, grain, soybean	1.000	0.250	0.000
59	ship	ship	1.000	1.000	1.000
Média			0.833	0.446	0.500

Tabela 5.14: Resultados do teste sobre textos do Reuters-21578

Este é o primeiro teste do classificador fora do ambiente TRECVID. O propósito da sua presença é dar apenas uma ideia do seu desempenho em textos correntes, já que a selecção dos textos não foi totalmente imparcial.

Dois aspectos saltam à vista: o corpus da Reuters/Carnegie Group é muito ecléctico nos seus temas. No TRECVID as *queries* descrevem uma cena, que se procura traduzir o mais fielmente possível nos conceitos. Isto por certo está também relacionado com a diferença nos tamanhos dos corpus.

O segundo ponto é que o classificador aceita com muita facilidade conceitos como sendo excelentes. Muitas palavras são referidas no texto sem que esse assunto esteja presente nos temas.

O problemas expostos já têm soluções previstas (pontos 6.2, 6.3, 6.5 e 6.6) que poderão revelar-se benéficas no futuro, onde um melhor (e mais justo) teste no corpus Reuters-21578 no futuro.

Capítulo 6

Trabalho futuro

Várias ideias surgiram durante a elaboração do trabalho que, apesar de em muito contribuir para a sua melhoria e crescimento, não tiveram hipótese de ser exploradas por falta de tempo.

Algumas são mais simples e imediatas, outras carecem de análise mais detalhada, para testar a sua viabilidade.

Segue-se uma enumeração das mais significativas.

6.1 Trabalhar por parágrafos.

Como foi referido logo na Secção 2, o presente sistema apenas opera sobre ficheiros completos. Acontece que essa limitação é significativa em alguns casos, mesmo sendo trivial partir o texto para análise.

Por exemplo, um livro de memórias é normalmente uma aglomeração de pensamentos que se cruzam, e não se alinham por capítulos. Um catalogador automático de correio electrónico iria deparar-se com a mesma situação. Ainda outro exemplo são artigos de opinião que comentam diversos eventos contemporâneos, sejam eles sociais, políticos, ou económicos.

Um parágrafo é suposto abordar uma ideia apenas. É, portanto, o melhor candidato a unidade de processamento. O limite do campo de submissão de texto no website do VISL é outro factor que apoia esta divisão, pois aceita apenas algumas dezenas de KBs. Além disso, este analisador sintáctico não distingue entre o uso de pontos para terminar frases ou assinalar abreviaturas, o que torna a análise por frases impossível.

Classificar parágrafos é mais difícil do que classificar textos, porque os indícios mais significativos encontram-se um pouco distribuídos. Por exemplo, um artigo desportivo pode mencionar "bola" e "relvado", que com um grau de certeza significativo, pode reduzir a classificação, não só ao contexto de desporto, como até a meia-dúzia de desportos. Mas se no parágrafo sob análise é descrito quem ficou no "banco", é fácil apontar como provável o contexto financeiro.

Um texto não pode ser visto só parágrafo-a-parágrafo, da mesma forma que uma frase não pode ser vista palavra-a-palavra. Há que saber observar a articulação entre eles.

Algumas ideias a abordar:

- Um discurso não costuma saltitar entre coisas diferentes. E quando o faz, costuma usar alguns padrões linguísticos para anunciá-lo. "Agora outro assunto", "Já me esquecia", "Mudando o tema da conversa", ...

  Não o fazer confunde o leitor, que não encontra a linha condutora do discurso, e deixa-o distrair-se facilmente da mensagem que se pretende transmitir.

- Os textos são muitas vezes agrupados por temas. Quando abordam mais que um assunto, não alternam entre eles. Sabendo isto, é mais fácil detectar falsos indícios espontâneos que não têm continuação.

  No exemplo supra, entre "bola" e "relvado" surge "banco". Com toda a probabilidade, o sentido de "banco" não será o de instituição bancária. Mesmo que não se consiga estabelecer a relação com "desporto" (abordado no ponto 6.3), pode, ao menos, ser marcada como "outlier", e ser revista mais tarde.

- Nem sempre o conceito mais próximo é o melhor. Um conceito mais constante pode ser mais adequado. Esta ideia é abordada de forma mais detalhada no ponto 6.2.

- Como indicar que o livro começou como uma história de guerra, e terminou como um romance? É necessário repensar o esquema XML para indicar subsecções.

6.2 Perseverança de temas

Assuma-se que no exemplo de desporto já mencionado, 50 palavras apontam para um tema, ainda que não de forma dominante ("campo", "equipamento", "desafio", "treino", ...), e uma única palavra coincide com um conceito, que não volta a ser mencionado (como é o caso do "banco" apontar para conceitos financeiros, ou mencionar "leões" ou "águias" e surgir o conceito "animal"). No final, esse mero acaso pode surgir na classificação final do texto, e o conceito mencionado constantemente pode ficar excluído. Isto porque, na forma presente, considera-se o conceito *mais próximo.*

Este sistema foi crescendo a partir das *queries* do TRECVID, que são frases curtas e directas, como a maior parte das perguntas. Ainda precisa de crescer e adaptar-se a textos maiores e menos bem delineados. Assim, pode dar-se o devido valor a palavras que surgem repetidas várias vezes (que

poderiam ser, mas não são ignoradas no presente), já que insistem no mesmo tema.

Reconhecer o mérito de um tema que é "rondado", e nunca é visto como mencionado directamente, é um sinal de maturidade que se procura para este programa. Da mesma forma, outro objectivo é evitar que um "golpe de sorte" coloque outro conceito no topo de pontuação, sem outro indício para a sua eleição. Possivelmente essa palavra foi até mal interpretada. Esse aspecto cairá no âmbito do ponto 6.3.

6.3 Determinar contextos de frases

Se uma palavra tiver cinco sentidos, como determinar qual o desejado pelo autor? De momento, é aquele que mais convém: o que estiver mais próximo de um conceito. Mas uma escolha mais adequada seria aquele que mais se enquadra no resto da frase, parágrafo ou texto. Veja-se o exemplo:

> The artist used carmine lake for the background.

Como visto na Tabela 4.1, "lake" pode ser interpretado de várias formas. A associação com a cor, e "artist" deveria de ser suficiente para persuadir o sistema a tomar este significado. Nesta situação, o significado de "lake" pode e deve ser ajustado.

Espera-se que ao enveredar por este tipo de selecção de significados, se encontrem menos classificações surpreendentes. Este é também o primeiro passo para conseguir definir conceitos por palavra, e não por *synset*, como falado no ponto 6.4..

Para conseguir este objectivo, há que analisar todas as combinações de *synsets* de palavras, e avaliá-las com todos os *synsets* de conceitos. Isto é,

$$\sum_{i=1}^{n} S(P_i) \sum_{i=1}^{m} S(C_i) \tag{6.1}$$

torna-se

$$\prod_{i=1}^{n} S(P_i) \prod_{i=1}^{m} S(C_i) \tag{6.2}$$

onde P e C correspondem ao conjunto de palavras e de conceitos, n e m são o número de elementos nesses conjuntos, e S é uma função que indica o número de *synsets* associados a cada palavra ou conceito. Isto traduz-se num grande aumento de tempo de processamento.

Uma forma de tentar reduzir o número de possibilidades a analisar consiste em aproveitar o backtracking do Prolog, e ir cortando soluções assim que se nota que é inferior à melhor encontrada até ao momento.

É fácil modificar a frase usada como exemplo para uma situação em que essa adaptação não deva ocorrer. Daí que seja conveniente efectuar medições,

a fim de verificar se haverão ganhos. Possivelmente, limitar a "coerção" a palavras adjacentes ou agrupadas por uma análise semântica mais avançada, iria também reduzir o tempo de processamento.

6.4 Definir conceitos por palavras

Embora seja uma actividade a desempenhar raramente, a modificação da lista de conceitos usados para classificar textos é morosa. Se fosse possível defini-los apenas como palavras, a usabilidade da aplicação aumentaria, já que é uma actividade mais usual.

Este requisito pressupõe a existência de utilizadores, o que implicaria um estado de desenvolvimento bastante mais avançado que o actual.

6.5 Penalizar associações

O TRECVID define claramente na lista de conceitos que uma pessoa não conta como animal. No WordNet o contrário é indicado. Existem duas soluções:

- Remover as relação do WordNet, ou
- Definir uma emenda.

A possibilidade de penalizar soluções permite um aumento da flexibilidade da aplicação, visto que reutiliza a mesma base (WordNet) para vários utilizadores, e permite configurar pesos atribuídos a diversas associações (podendo até beneficiar algumas).

Penalizar uma relação incrementando a pontuação com um valor suficientemente alto será praticamente o mesmo que remover a relação do WordNet.

Uma possibilidade será usar um algoritmo de aprendizagem para ir ajustando os valores de penalização. Isso poderia dar origem a pesos diferentes às relações.

6.6 Conceitos compostos

Como foi dito na Secção 3.1.3, os conceitos compostos servem o propósito de referir uma situação mais específica que aquela que é representável por um único *synset* no WordNet.

"daytime outdoor", "female reporter", "head and shoulder", "indoor sports venue", "male news subject", "military building" são todos formados por combinações de dois ou três *synsets*, e é até fácil de fazer.

A solução mais simples consiste em definir conceitos compostos (como "female-reporter"), que teriam associados a si uma lista de conceitos (neste

caso, "female" e "reporter"). Os conceitos ditos normais estariam presentes também nesta forma, com apenas um elemento na sua lista.

Antes de apresentar os resultados ao utilizador, a lista com os conceitos seria analisada, em busca dos conceitos compostos. À medida que cada um destes conceitos vai sendo encontrado, são removidos os conceitos simples da lista de conceitos que o compõem, e acrescenta-se na lista de conceitos compostos o respectivo elemento (com a pontuação média ou mínima dos conceitos encontrados). No final do processo, os conceitos que só por si não conseguem satisfazer os requisitos de qualquer conceito composto são descartados. Será a lista dos conceitos complexos que será transformada no output.

Isto coloca uma questão quanto a eventuais conflitos. Poderá um conceito simples pertencer a mais de um conceito? Se sim, então não se devem remover da lista de conceitos ao serem convertidos em conceitos complexos. Caso contrário, a ordem pela qual são especificados será relevante.

6.7 Melhorar a aglomeração de palavras

Na Secção 3.3.1, mencionou-se que os substantivos complexos são formados por dois ou mais substantivos. Esta definição deveria ser expandida, e criada a palavra complexa, pois há muitos nomes que não são formados só por substantivos — embora a maioria das pessoas esteja disposta a aceitar as preposições sem questionar.

Nomes como "Bank of America", "Department of Justice" ou "Bill of Rights" não são ainda reconhecidos, pois "of" não é um substantivo.

Uma solução simples consiste em admitir preposições na lista de substantivos, se forem a única palavra entre dois substantivos. Estas preposições serão removidas após a fase de aglomeração, em que se faz uma passagem para seleccionar apenas os substantivos reconhecidas pelo WordNet, tal como estão (sem abreviaturas).

6.8 Refazer as negações

Quando se apresentou a forma de tratar negações, na Secção 3.3.4, referiu-se que as palavras que as indicavam actuavam sobre o substantivo seguinte. Esta forma de negação não abrange substantivos complexos. Para o conseguir há que modificar algumas formas de representação, a fim de conseguir lidar com a complexidade acrescida.

Por outro lado, nem sempre a negação precede o nome. O caso "A whale is not a fish." necessita apenas de uma simples adaptação para ser reconhecido. "A bat doesn't have feathers.", "These people are not good neighbors." ou "I had no good or bad experience." já são mais complexos. Este tipo de análise

sintáctica de alto nível deverá ser planeada de modo a lidar com outros casos a serem tratados a esse nível, antes da análise de conceitos.

Ainda por outro lado, como foi também já referido, num discurso, nem sempre a negação implica a inexistência de algo: "I'm like no man." ilustra essa situação. Noutras alturas, a palavra negada é relevante para o contexto: "A glass with no wine." desambigua "glass" como copo, e não como vidro.

No que toca a *queries*, que são curtas e directas, pode-se dizer que a negativa é intencional — isto é, está lá para ser respeitada, já que as perguntas costumam ser tão curtas quanto possíveis.

Num texto, as negações servem para esclarecer dúvidas: coisas que poderiam ser, que o receptor da mensagem poderia pensar que tivessem sido, mas não é verdade. Logo, se há essa forte possibilidade, é relevante, e ajuda a estabelecer o contexto.

Estas ideias carecem ainda de testes práticos, a fim de confirmar a sua validade.

Dada a complexidade do assunto, alterações mais do que simples serão efectuadas apenas se o seu desempenho não for preciso o suficiente.

6.9 Lidar com abreviaturas

As abreviaturas foram explicadas na Secção 3.3.2. Encontram-se dois problemas com o seu uso:

O VISL trata todos os pontos como finais Não é possível ter a certeza se um ponto é ponto final ou se indica uma abreviatura, salvo testando para saber se a abreviatura é reconhecida.

De momento, o ponto é aplicado apenas a palavras de uma letra, mas como foi dito, há abreviaturas de duas ou mais letras que devem ser reconhecidas.

A actual estratégia pode não parecer muito significativa no geral, mas após pesá-la com as situações que resolve — não são muitas — merece ser revista. Em especial porque, ao estender a cobertura de palavras abreviadas de apenas uma letra para (por exemplo) 3, o impacto no tempo de processamento irá também aumentar.

A melhor solução poderá passar por fazer um levantamento da base de dados do WordNet, e recolher todas as palavras que surgem com um ponto. Esta lista será guardada num ficheiro, para evitar ter de refazer a busca.

Sempre que uma palavra não consta no WordNet, é procurada nesta (curta) lista na altura de selecção de palavras. Tal ocorrerá, não na fase de aglomeração, como é feito actualmente, mas na fase de selecção dos substantivos, garantindo que esta informação é procurada apenas uma vez para cada palavra. Caso seja encontrada, é adicionada à lista de palavras com o ponto adicionado. Se acontecer que não venha a fazer parte de uma palavra

composta, será descartada da lista. Isto enquadra-se bem no processo de inclusão de preposições, descrito no ponto 6.7.

O WordNet não é consistente nas abreviaturas Por exemplo, a abreviação de "United States" parece difícil de resolver completamente. Existe no WordNet "U. S. Army" e "US Navy". Claramente, é necessário estabelecer consistência.

Não há outra solução senão modificar manualmente todas as situações irregulares, e submeter a alteração para ser integrada no projecto, a fim de ser resolvido na versão seguinte e posteriores. Caso contrário, esta solução (assim como outras) não irá cobrir todos os casos.

Procurar um-a-um será uma tarefa morosa. O mais indicado será efectuar pesquisas por:

- Duas letras maiúsculas seguidas,
- Duas letras maiúsculas separadas por um espaço,
- Uma letra maiúscula isolada (por exemplo, no fim — "Dr. J", "Mr. T", "Melanie C"),[1]
- Pontos sem espaços a seguir, e
- Abreviaturas conhecidas, sem estarem terminadas por um ponto.

Os casos encontrados serão observados, e modificados de forma adequada e consistente. Há que ter em atenção o caso particular de siglas, que são uma forma de abreviatura que não usa pontos.

Visto que o problema não se limita ao WordNet, deverá ser criado um mecanismo que traduza esta variabilidade gráfica em algo reconhecido durante a selecção de substantivos do texto.

6.10 Nomes próprios que são também comuns

As dificuldades com o nome "Condoleezza Rice", apresentado na Secção 3.3.1 revelam um problema interessante.

Testar todas as palavras em maiúscula e minúsculas é muito moroso. Fazer a lista dos nomes comuns que são também nomes próprios no WordNet (análogo ao sugerido no ponto 6.9) é uma possibilidade, mas não muito elegante.

Na Secção 3.2 é apresentado o output do VISL. Pode notar-se que é mostrada a palavra original, e que a mesma informação é transferida para o Prolog. Logo, mesmo que o nome seja "United Nations" — que o VISL

[1] Devido à grafia predominante nestas situações, pode optar-se por ignorá-los

traduziria para "United nation" — é possível recuperar quer a grafia, quer o número.

A palavra original deve ser introduzida na lista de palavras em vez do lema se se verificarem todas as seguintes condições:

1. A palavra é um substantivo.

2. A palavra original foi escrita com maiúscula, e o lema apresentado pelo VISL não é.

3. A palavra não inicia a frase.[2]

4. A palavra é precedida por um ou mais substantivos próprios, devidamente reconhecidos.[3]

Desta forma, deverá ser possível reconhecer a grande maioria dos casos em falta. Em inglês, no caso de um nome composto, apenas a última palavra pode ser passada para o plural (ao contrário do português, onde a regra é mais complexa). No caso de nomes próprios, são os apelidos que são mais propensos a confusões.

O VISL, quando não reconhece uma palavra, classifica-a como substantivo, e não a altera.

6.11 Conflitos nas listas de conceitos afirmados e negados

Os conflitos entre as listas de palavras afirmadas e negadas foram apresentados na Secção 3.3.4, sendo uma análise dos problemas relacionados apenas com a negativa feita no ponto 6.8 (Refazer as negações).

A resolução proposta para o caso das *queries* será, antes de escrever os resultados, remover cada conceito conflituoso da lista em que possui a pontuação inferior. Em caso de empate, deve ser removido da lista de palavras negadas.

Existem dois motivos para dar preferência à lista das palavras afirmadas: primeiro, porque se tende a dar mais atenção a esta lista, e como tal, se este comportamento não for adequado, é mais provável que seja mudado. Segundo, porque o sistema de de reconhecimento visual necessita do máximo de informação possível. Caso testes forneçam a indicação de que esta escolha está incorrecta, pode ser alterada com facilidade.

[2]Dado que o VISL considera todos os pontos como pontos finais, pode haver problema com nomes precedidos por abreviaturas, como "U. S. S. Tropedo".

[3]Ou possivelmente pelo menos uma abreviatura.

6.12 Caminhos mistos

Quando se explicou a forma como os caminhos são definidos, e como são calculados, foi dito que seguiam apenas *uma relação.* Seria interessante aferir os ganhos obtidos por permitir o cruzamento de relações.

Possivelmente serão promissores, visto que as relação de meronímia passariam a beneficiar da força dos hiperónimos. Mas estas vantagens levam ao rápido crescimento do número de caminhos para cada *synset.* Em testes preliminares, 125MB de *Global stack* (o máximo) eram insuficientes para calcular muitos dos caminhos. Além disso, o tempo de processamento aumentou muito.

Será necessário uma outra abordagem à geração de caminhos, já que não podem ser todos guardados. Possivelmente um algoritmo de "gera e compara com o melhor até ao momento", ou outro mais agressivo de "interrompe este caminho, porque por aqui os resultados serão piores que outro já visto".

6.13 Traduzir para substantivos colectivos

Um dos conceitos definidos no TRECVID 2006 foi "crowd". No WordNet, tal significa apenas um agrupamento, sem mencionar que é de pessoas, o que implica que a distância entre "person" e "crowd" é grande.

Assim, se surgir uma frase "Find shots with many persons", dificilmente este conceito seria eleito (com a palavra "people" já será reconhecido).

"Many" é um adjectivo, assim como "several" e outros quantificadores. Logo, esta ideia vai depender do ponto 6.15, que se refere ao uso de palavras de outras classes gramaticais.

6.14 Recorrer à WWW

Quando confrontado com uma palavra desconhecida, o sistema deve procurar ajuda para a entender, antes de desistir e ignorá-la.

É possível que o erro se deva a um problema ortográfico no texto. Mas antes de enveredar por esse campo, há que considerar a hipótese de alargar a base de conhecimento, e fazer um esforço para reconhecer a palavra.

Actualmente, existem várias obras de referência de dimensão significativa na internet: dicionários, enciclopédias, biografias, livros, e muito mais. Tudo acessível através da *World Wide Web.*

Apesar de ser possível tentar encontrar a palavra num número fixo de recursos, e desenvolver *parsers* para extrair a informação, é preferível encontrar um serviço que procure em mais de um sítio, e retorne um resumo.

Uma pesquisa no Google iniciada por "define:" executa uma busca numa série de websites, e apresenta um excerto do texto que define o que foi escrito na *query.* Normalmente, o excerto corresponde às primeiras linhas, que

costumam dar uma ideia geral.

As vantagens de fazer uma pesquisa deste tipo prendem-se com o número de *websites* consultados pelo Google, o resumo da informação, e estar tudo junto, no mesmo formato. Desta forma, só é preciso fazer *parsing* a uma página.

Após receber o resultado da *query*, e compilado o texto com a informação útil, este texto pode ser submetido ao VISL para análise sintáctica, e dar início a mais um ciclo do sistema, a fim de obter os conceitos associados. Os melhores conceitos são escolhidos, e associados (mantendo a pontuação) à palavra original.

Uma área em que este sistema ajudará bastante é nos nomes próprios. O WordNet é um pouco pobre nesse campo, assim como nos nomes de marcas e empresas.

Por exemplo, de acordo com o que foi descrito no ponto 6.10 (Nomes próprios que são também comuns), existe a forte suspeita de que "Condoleezza Rice" é um nome próprio. No entanto, não consta no WordNet. É feita a busca na WWW.

O resultado inclui um excerto da versão inglesa da Wikipedia. Diz:

> Condoleezza Rice (born November 14 1954) is the 66th United States Secretary of State, and the second in the administration of President George W. Bush to hold the office. She succeeded Colin Powell on January 26, 2005, after his resignation.

Este texto é processado pelo VISL, passa para o Prolog, e é dado início ao seu processamento.

São identificadas as seguintes palavras:

```
[rice, November, State, secretary, State, administration,
President George W. Bush, office, Powell, January,
resignation]
```

que fazem sobressair os conceitos presentes na Tabela 6.14.

Conceito	Pontuação	Palavra
office	-100	office
person	-50	secretary
vegetation	-50	rice
studio	-8	office

Tabela 6.1: Conceitos para "Condoleezza Rice"

Estes conceitos (que seriam todos seleccionados) são depois associados a "Condoleezza Rice", e o resto da frase continua o seu processamento.

Apesar de não ser perfeito (a "resposta correcta" seria apenas "person"), consegue resultados muito melhores do que ignorar o que se desconhece.

6.15 Usar mais do que os substantivos

Apesar de haver ainda muito que explorar apenas com os substantivos, eventualmente será necessário expandir a aplicação para abranger outras relações.

Os adjectivos e os verbos ajudam a limitar os significados de uma palavra. Será este o passo inicial antes de tentar usar estas classes para influenciar directamente o resultado de alguma forma (como serem comparados com conceitos).

A função dos adjectivos é caracterizar os substantivos. Visto que nem todos os adjectivos podem ser relacionados com todos os significados possíveis de um substantivo,[4] o número de escolhas pode ser reduzido. Por exemplo, ao dizer "empty glass", fala-se de copo e não de vidro (o material). Espera-se que muitas das vezes o recurso aos adjectivos torne possível fazer a escolha entre significados concretos e abstractos.

Os verbos podem também ajudar a reduzir ambiguidades no texto. Por exemplo, "Do it right" ou "Turn right", mas na língua inglesa operam sempre sobre um sujeito da frase (exemplo: "He runs"), ou mais (exemplo: "He greets the coleague and his friend"). Esse sujeito pode estar implícito na forma oculta ou indeterminada. Dado que é difícil descobrir quais os sujeitos a que o verbo se refere em frases mais elaboradas, enquanto que os adjectivos precedem os substantivos que caracterizam, serão estes últimos que irão ser adaptados ao classificador primeiro.

6.16 Multi-threading

Até aqui tem-se falado mais do desempenho em termos de resultados do que do tempo de processamento. Um classificador de *queries* é uma ferramenta interactiva muito útil (no TRECVID não existe interactividade, o tempo de execução é passado para segundo plano).

Após a análise de toda uma biblioteca, os livros podem ser seleccionados comparando os temas da query com os temas do conteúdo do livro (ou do seu resumo, se tal houver).

Por exemplo: "Find scary short stories", "Find romances set in Paris", "Find epic fantasy books with no elves" ou "Find books on Kepler's laws".

Para ser útil, o classificador deve ser rápido, apesar dos bons resultados serem o mais importante.

Felizmente, a abordagem utilizada é altamente paralelizável. Como as palavras são analisadas uma a uma, e de forma independente, estas podem ser distribuídas por várias *threads*. Cada *thread* irá pedir uma outra palavra para analisar após ter pontuado todos os caminhos da palavra anterior, e determinado a pontuação para cada conceito.

[4] Assume-se que este classificador não será usado em poesia.

A base de dados do WordNet é usada apenas para consulta (logo, partilhável, o que é agradável já que representa muitos dados em memória), e existem apenas duas estruturas de dados significativas — a lista de palavras e a lista de conceitos.

A fase inicial, onde se define a lista de palavras a analisar, não beneficia de paralelismo. Logo não é necessário regular o aceso a ela.

As listas de conceitos irão requerer semáforos para evitar problemas de concorrência na escrita.

É conveniente ter as threads sempre ocupadas. Quer isso dizer que, para evitar o *downtime* das tarefas de selecção de palavras e de *output*, poderá optar-se por uma *thread* especializada, que garantia uma lista de palavras sempre à frente. O *output* poderá ser guardado em memória até ao fim do processamento.

Pensando em maior escala, pode imaginar-se o mesmo processo multiplicado por várias máquinas, havendo um coordenador central que vai distribuindo as listas de palavras, e reconstruindo o resultado na ordem correcta.

Há que lembrar, no entanto, que várias máquinas *multicore* não são bons substitutos para algoritmos eficientes, e os actuais beneficiariam de algumas abordagens já apresentadas, a fim de reduzir o número de caminhos a comparar.

Capítulo 7

Conclusão

Terminado o trabalho, há que fazer a sua avaliação, e verificar se os objectivos inicialmente traçados foram cumpridos.

A primeira ilação que se pode retirar, é que o sistem desempenha a sua função de classificação de forma correcta, ainda que lenta; sublinhe-se, no entanto, que foi desde logo claramente assumido este aspecto como secundário. A abrangência e precisão perto dos 60% é um nível razoável, e seria superior se o grau de exigência não fosse tão alto. O teste simples da Reuters apresentou também bons resultados, com uma abrangência superior a 80% e 45% de precisão.

A análise sintáctica efectuada pelo VISL, embora apresente imperfeições, estas são contornáveis. A final, a análise considera-se como satisfatória.

O WordNet apresenta algumas inconsistências e falhas na informação, nalguns pontos desactualizada, noutros omissa. Atendendo à globalidade da sua informação, seria injusto apontar alguns casos particulares a uma aglomeração de conhecimento destas dimensões.

Trabalhar com o WordNet relembra o que muitos apontam como a origem do sucesso do Google: deixar as máquinas executarem a tarefa em que são melhores (indexar grandes volumes de dados), e os humanos fazerem o que eles fazem bem (relacionar informação).

Aqui usa-se o conhecimento e trabalho de peritos em linguística, o que faz toda a diferença, já que o sistema não começa do zero. No entanto, seria interessante complementar, um dia, esta informação com a capacidade de aprendizagem.

No geral, observando os resultados, que se afiguram bastante satisfatórios, e consultando o rol de ideias presente na secção dedicada ao trabalho futuro, é caso para ter esperança num feliz desempenho caso se verifique uma participação no TRECVID 2008. Quanto a uma aplicação fora deste âmbito, ainda não é possível fazer projecções, visto quase não haver testes elaborados. Salienta-se ainda que, no presente estado, o tempo de execução é muito elevado, apesar de existirem já soluções planeadas.

Bibliografia

[1] CALISTRU, C. et al. Inesc, porto at trecvid 2007: Automatic and interactive video search. 2007.

[2] SMEATON, A. F.; OVER, P.; KRAAIJ, W. Evaluation campaigns and trecvid. In: *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval.* New York, NY, USA: ACM Press, 2006. p. 321–330. ISBN 1-59593-495-2.

[3] RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach.* 2nd edition. ed. [S.l.]: Prentice-Hall, Englewood Cliffs, NJ, 2003. 482,749,750,790-795 p.

[4] NORVIG, P. *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp.* [S.l.]: Morgan Kaufmann Publishers, 1992. ix,655,656 p.

[5] WIKIPEDIA: WordNet. `http://en.wikipedia.org/wiki/Wordnet` em 2007-10-20.

[6] VISL: website. `http://visl.sdu.dk/` em 2007-09-15.

[7] VISL: References and credits. `http://visl.sdu.dk/visl/en/credits.html`.

[8] KARLSSON, F. et al. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text.* [S.l.: s.n.], 1995.

[9] WITZIG, S. Accessing wordnet from prolog. 2003.

[10] WIELEMAKER, J. An overview of the SWI-Prolog programming environment. In: MESNARD, F.; SEREBENIK, A. (Ed.). *Proceedings of the 13th International Workshop on Logic Programming Environments.* Heverlee, Belgium: Katholieke Universiteit Leuven, 2003. CW 371.

[11] WORDNET: Website oficial. `http://wordnet.princeton.edu/` em 2007-09-20.

[12] IEEE. Standard taxonomy for software engineering standards (ANSI). *The Institute of Electrical and Electronics Engineers Inc.*, 1986.

[13] PANTEL, P.; LIN, D. Spamcop — a spam classification & organization program. In: *Proceedings of AAAI-98 Workshop on Learning for Text Categorization.* [S.l.]: AAI Press, 1998.

[14] SAHAMI, M. et al. A bayesian approach to filtering junk e-mail. In: *Proceedings of AAAI-98 Workshop on Learning for Text Categorization.* [S.l.]: AAI Press, 1998.

[15] SAHAY, S. Support vector machines and document classification.

[16] SAHAMI, M.; YUSUFALI, S.; BALDONADO, M. Q. W. Real-time full-text clustering of networked documents.

[17] HOI, S. C. H.; WONG, L. L. S.; LYU, A. Chinese University of Hong Kong at TRECVID 2006: Shot boundary detection and video search. 2006.

[18] THE Lemur Toolkit for Language Modeling and Information Retrieval. `http://www.lemurproject.org/`.

[19] LIU, Z. et al. AT&T research at TRECVID 2006. 2006.

[20] LINGPIPE. `http://www.alias-i.com/lingpipe/`.

[21] SMEULDERS, A. et al. Semanic video search. 2006.

[22] SJöBERG, M. et al. Picsom experiments in trecvid 2006. 2006.

[23] JIANG, Y.-G. et al. Modeling local interest points for smantic detection and video search at trecvid 2006. 2006.

[24] COLUMBIA374. `http://www.ee.columbia.edu/ln/dvmm/columbia374/` em 2007-10-14.

[25] REUTERS21578. `http://www.daviddlewis.com/resources/testcollections/reuters21578/` em 2007-10-14.

[26] WIKIPEDIA: Parsing. `http://en.wikipedia.org/wiki/Parsing` em 2007-10-16.

[27] VISL: aplicação web. Setembro. `http://beta.visl.sdu.dk/visl/en/parsing/automatic/parse.php` em 2007-08-30.

[28] WIKIPEDIA: Lema. `http://pt.wikipedia.org/wiki/Lema_%28ling%C3%BC%C3%ADstica%29` em 2007-09-29.

[29] Cunha, C.; Cintra, L. F. L. *Nova gramática do português contemporâneo.* 6ª. ed. [S.l.]: Edições João Sá da Costa, 1989.

[30] DUVIDAS da língua portuguesa. `http://ciberduvidas.sapo.pt/pergunta.php?id=9773`, `http://ciberduvidas.sapo.pt/pergunta.php?id=13044` em 2007-10-02.

[31] WIKIPEDIA: Lake pigment. `http://en.wikipedia.org/wiki/Lake_pigment` em 2007-10-18.

[32] SECO, N.; VEALE, T.; HAYES, J. An intrinsic information content metric for semantic similarity in wordnet. IOS Press, v. 110, 2004.

[33] MAHESH, K. *Text Retrieval Quality: A Primer.* `http://www.oracle.com/technology/products/text/htdocs/imt_quality.htm` em 2007-10-24.

Apêndice A

Lista de conceitos

sport	Shots depicting any sport in action
entertainment	DROPPED (nao vai ser usado)
weather	Shots depicting any weather related news or bulletin
court	Shots of the interior of a court-room location
office	Shots of the interior of an office setting
meeting	Shots of a Meeting taking place indoors
studio	Shots of the studio setting including anchors, interviews and all events that happen in a news room
outdoor	Shots of Outdoor locations
building	Shots of an exterior of a building
desert	Shots with the desert in the background
vegetation	Shots depicting natural or artificial greenery, vegetation woods, etc.
mountain	Shots depicting a mountain or mountain range with the slopes visible
road	Shots depicting a road
sky	Shots depicting sky
snow	Shots depicting snow
urban	Shots depicting an urban or suburban setting
crowd	Shots depicting a crowd
waterscape	Shots depicting a waterscape or waterfront
waterfront	Shots depicting a waterscape or waterfront
face	Shots depicting a face
person	Shots depicting a person (the face may or may not be visible)
government leader	DROPPED (nao vai ser usado)
corporate leader	DROPPED(nao vai ser usado)
military	Shots depicting the military personnel
police	Shots depicting law enforcement or private security agency personnel
security	Shots depicting law enforcement or private security agency personnel
prisoner	Shots depicting a captive person, e.g., imprisoned, behind bars, in jail or in handcuffs, etc.
animal	Shots depicting an animal, not counting a human as an animal

american flag	Shots depicting a US flag
computer screen	Shots depicting a television or computer screen
television	Shots depicting a television or computer screen
airplaine	Shots of an airplane
car	Shots of a car
bus	Shots of a bus
truck	Shots of a truck
boat	Shots of a boat or ship
walking	Shots depicting a person walking or running
running	Shots depicting a person walking or running
geological phenomen	Shots depicting the happening or aftermath of a natural disaster such as earthquake, flood, hurricane, tornado, tsunami
boat	Shots of a boat or ship
ship	Shots of a boat or ship
people	Shots depicting many people marching as in a parade or a protest
marching	Shots depicting many people marching as in a parade or a protest
explosion	Shots of an explosion or a fire
fire	Shots of an explosion or a fire
map	Shots depicting regional territory graphically as a geographical or political map
chart	Shots depicting any graphics that is artificially generated such as bar graphs, line charts, etc. (maps should not be included)

Apêndice B

Calcular um caminho

```
% Calcula um caminho, em pseudo-prolog
caminho_inv(Função, Nó, [Nó|Caminho]):-
   Função(Nó, Pai),
   caminho(Função, Nó, Caminho).

caminho_inv(Função, Nó, [Nó]):-
   \+ Função(Nó, _Pai),
   Função(_Filho, Nó),
   !.

caminho(No, Função, Caminho):-
   caminho_inv(Função, Nó, Caminho_inverso):-
   reverse(Caminho_inverso, Caminho).
```