UNIVERSIDADE NOVA DE LISBOA

Faculdade de Ciências e Tecnologia

Departamento de Informática

Extracção de Unigramas Relevantes

Por

João Miguel Jones Ventura

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Engenharia Informática.

Orientador: Prof. Dr. Joaquim Francisco Ferreira da Silva

Lisboa

2008

"*Ninguém é tão ignorante que não tenha algo a ensinar. Ninguém é tão sábio que não tenha algo a aprender.*"

Blaise Pascal

Dedicatória

Aos meus pais, João e Rosabela.
À minha irmã, Amarilis Ventura.
À minha namorada, Carmen Matos.

Agradecimentos

A todos os que fiz menção na dedicatória, pelo apoio e paciência que tiveram enquanto este trabalho foi realizado. Também ao Professor Paulo Quaresma do Departamento de Informática da Universidade de Évora, pelas sugestões feitas no sentido de enriquecer esta dissertação. Agradecimentos também à Professora Maria Francisca Xavier do Departamento de Linguística da Universidade Nova de Lisboa, por se ter disponibilizado para a árdua tarefa de estabelecer critérios de classificação de palavras no tocante à sua relevância. Por fim, gostaria de agradecer também ao Doutor Joaquim Francisco Ferreira da Silva, meu orientador, com quem tem sido um enorme prazer trabalhar.

Resumo

A extracção automática de Unidades Lexicais Multipalavra (ULM) a partir de *corpora* é actualmente uma área de grande aplicabilidade. Porém, os avanços na aplicação das ULMs vieram realçar uma lacuna: os conjuntos obtidos pelos extractores de ULMs são incompletos porque não incluem as unidades de uma só palavra — os Unigramas Relevantes (URs).

Com efeito, a extracção de URs é uma área ainda pouco explorada onde as abordagens actuais apresentam algumas limitações. Umas são demasiado simplistas e permissivas; outras bastante punitivas em determinadas situações. Estas limitações motivaram a criação das métricas *Score* e *SPQ*, desenvolvidas no âmbito desta dissertação.

Por outro lado, essas abordagens apenas permitem obter listas que medem a importância relativa dos unigramas. Porém, nalgumas aplicações poderá ser necessária a classificação *booleana* acerca da relevância de uma palavra, como por exemplo, obter as palavras-chave que verdadeiramente caracterizam um documento. A inexistência de qualquer abordagem capaz desta classificação, com bons resultados, motivou a criação do Método das Ilhas.

Esta dissertação propõe novas abordagens para os problemas acima mencionados e compara resultados com as abordagens existentes. Por fim, apresenta também o Método das Sílabas que, de uma forma bastante simples e, julgo, inovadora, permite melhorar substancialmente os resultados em geral.

Abstract

The automatic extraction of Multiword Lexical Units (MLU) from corpora is currently an area of great applicability. However, advances in MLU application have shown a gap: the sets obtained by MLU extractors are incomplete because they don't include single word units — the Relevant Unigrams.

Indeed, the extraction of relevant unigrams is still an unexplored area where the current approaches present some limitations. Some are too simplistic and permissive; others more complex but quite punitive in some situations. Those limitations have motivated the creation of the metrics *Score* and *SPQ*, developed during this dissertation.

On the other hand, those approaches are only capable of creating lists that measure the relative importance of each unigram. However, in some applications it may be necessary to obtain the *boolean* classification about the relevance of a word, like, for instance, to extract the keywords that truly describe a document. The non existence of such approach, with good results, has motivated the creation of the Ilhas *(Islands)* method.

This MSc's dissertation proposes new approaches for the problems above described and compares results with the current approaches. Finally, it also presents the syllable method which, on a simple and original way, is able to greatly improve the results.

Conteúdo

Lista de Figuras

Lista de Tabelas

Glossário de Termos

Bigrama Sequência de dois elementos de texto, normalmente palavras.

Cluster Termo Inglês para "Conjunto".

Corpus Colecção de textos provenientes de uma ou várias fontes distintas.

Corpora Múltiplas colecções de textos. Plural de corpus.

Multipalavra Conjunto de duas ou mais palavras.

Script Sequência de instruções a serem executadas sequencialmente.

Unidades Lexicais Multipalavra Sequências de palavras que correspondem a nomes próprios, frases idiomáticas ou colocações com categoria gramatical.

Unigramas Um elemento de texto, normalmente uma palavra.

Unipalavra Uma palavra.

Trigrama Sequência de três palavras.

Text-mining Ramo de Inteligência Artificial que estuda a extracção de informação a partir de textos.

***n*-gramas** Sequência de n elementos de texto, normalmente palavras.

Capítulo 1

Introdução

A extracção automática de palavras-chave — termo também conhecido por MWUs *(Multiword Units)* ou por Expressões Relevantes — tem-se revelado muito útil em várias aplicações, como por exemplo na caracterização dos documentos por tópicos/palavras-chave e no agrupamento e classificação de documentos. Com efeito, estes dois domínios têm sido muito pesquisados nos últimos anos: [Smadja 93], [Dagan & Church 94], [Daille 96], [Silva et al. 99a], [Silva et al. 01a], [Jone & Paynter 02], [Gael & Spela 05], [Benoit 04], entre outras publicações. No entanto, a extracção automática das palavras isoladas e relevantes, ou seja, URs (Unigramas Relevantes), tem sido descurada, provavelmente pela dificuldade que acarreta. Contudo, é fácil demonstrar que num processo de extracção de palavras-chave, ignorar unigramas como candidatos a termos relevantes empobrece o resultado final. Tome-se o seguinte exemplo:

> *"Os orçamentos deterioraram-se devido à acção dos estabilizadores automáticos e também devido a medidas orçamentais discricionárias expansionistas de alguns Estados-Membros que não dispunham de margem de manobra para o efeito. De um modo geral, e apesar das pressões orçamentais, o investimento público manteve-se ou aumentou ligeiramente, excepto na Alemanha, Grécia e Portugal."*

De acordo com este exemplo, facilmente se identificam vários termos relevantes. Mas, se por um lado, termos multipalavra como "*estabilizadores automáticos*", "*medidas orçamentais discricionárias expansionistas*", "*margem de manobra*", "*pressões orçamentais*" e "*investimento público*" fossem detectados com razoável eficiência pelos modernos extractores multipalavra, termos unipalavra como "*orçamentos*", "*Estados-Membros*", "*Alemanha*", "*Grécia*" e "*Portugal*" seriam ignorados por incapacidade desses extractores. No entanto, uma simples contagem mostra que neste caso existem tantos termos multipalavra relevantes quantos os termos unipalavra relevantes.

1.1 Motivações

Conforme foi mencionado anteriormente, a extracção automática de Unigramas Relevantes é uma área que permaneceu, até há poucos anos, descurada pelos investigadores. Estando esta área de investigação ainda na sua infância, é natural que as poucas abordagens que existem sejam ainda, em geral, ineficientes. Este facto constituiu a motivação de base para o desenvolvimento desta Dissertação de Mestrado.

1.1.1 Classificação da Relevância dos Unigramas

Tendo em vista a relevância das palavras, *Luhn* [Luhn 58] sugeriu que, através de uma simples análise à frequência de ocorrência das palavras, se poderia verificar que palavras com uma frequência de ocorrência muito alta e palavras com uma frequência de ocorrência muito baixa deveriam ser consideradas pouco ou nada relevantes, por normalmente corresponderem respectivamente a palavras muito comuns e muito raras. Infelizmente esta análise é demasiado simplista. Na verdade, os textos podem ter palavras relevantes muito frequentes, que seriam descartadas por este método, e outras irrelevantes de frequência média, que seriam erradamente seleccionadas por este método. Outro problema desta abordagem seria a escolha dos limiares de frequência se quiséssemos decidir qual a fronteira entre palavras relevantes e não relevantes.

Em [Zhou 03] e [Ortuno et al. 02] são sugeridas outras abordagens. Basicamente estes autores assumem que uma palavra relevante é normalmente o assunto principal em contextos locais, ocorrendo mais frequentemente em certas áreas de um texto e menos frequentemente noutras áreas, dando origem a grupos locais (denominados *clusters*). Esta abordagem é extremamente intuitiva, mas no entanto é bastante restritiva em determinadas situações, nomeadamente para as palavras relevantes que são muito frequentes, pois estas tendem a espalhar-se duma forma relativamente uniforme pelo documento, especialmente em documentos de grandes dimensões, não formando *clusters* (ou grupos locais) significativos. Por outro lado, as palavras muito raras, normalmente com origem em erros ortográficos ou demasiado específicas a um contexto, tendem a formar *clusters*, se bem que menos significativos, podendo ser classificadas como relevantes quando na verdade não o são.

Desta forma, uma parte desta dissertação consiste na investigação associada à concepção de um novo classificador de relevância de unigramas a partir de *corpora*, que entra em conta com as limitações das abordagens anteriores.

1.1.2 Extracção de Unigramas Relevantes

Das várias abordagens existentes actualmente, na verdade, tanto quanto saiba, não existe uma única que possa ser considerada na realidade um extractor de unigramas. De facto, todas estas abordagens apenas nos permitem obter uma lista ordenada de relevância, o que é, por sua vez, bastante diferente de inferir sobre a relevância ou não relevância de um determinado unigrama — por simplicidade, nesta dissertação denomina-se relevância *booleana* a atribuição booleana do qualificativo *relevante* ou *não relevante* a um unigrama —. Assim, tome-se como exemplo a tabela 1.1 que representa uma lista parcial de unigramas ordenados por *score* de relevância atribuído pelo critério de uma hipotética abordagem.

Tabela 1.1: Lista parcial de unigramas ordenada por *score* de relevância segundo um critério hipotético.

Palavra	Posição na Lista
comissão	6
europeia	42
Portugal	200
e	2003
ou	2145
da	2415

Olhando para esta lista, facilmente se reconhece relevância semântica nas palavras do início da lista, mas não nas que se encontram no final da lista. No entanto, este reconhecimento é feito com base no conhecimento que nós humanos temos do vocabulário de que fazem parte estas palavras. Na verdade, sem esse conhecimento, olhando para esta lista apenas se pode inferir que as palavras "Comissão ", "Europeia" e "Portugal" devem ser, provavelmente, mais relevantes que as palavras "e", "ou" e "da". Mas não seríamos capazes de decidir quais as que são relevantes e quais as que não o são.

Uma regra empírica para decidir acerca da relevância *booleana* poderia passar pelo estabelecimento dum limiar de posição (*ranking*) na relevância. Desta forma, todas as palavras que na lista estivessem em posições acima do limiar seriam consideradas relevantes enquanto as outras seriam descartadas. Infelizmente, um classificador baseado neste critério seria bastante ineficiente, essencialmente por duas razões:

1. Muitas palavras não relevantes seriam colocadas em posições acima do limiar de decisão. Em consequência disso, estas palavras seriam erradamente consideradas unigramas relevantes.

2. Muitas palavras relevantes seriam colocadas em posições abaixo do limiar de decisão, o que implicaria que fossem consideradas, erradamente, unigramas não relevantes.

Apesar disso, as listas ordenadas com limiar de relevância podem ter aplicação prática. Contudo, se se quisesse caracterizar documentos através de tópicos / palavras-chave e se para tal fossem seleccionados apenas os primeiros unigramas de alguma dessas listas, tais documentos seriam, por certo, deficientemente caracterizados.

Assim, se por hipótese nenhum limiar de decisão permitir obter bons resultados, como conseguir então separar os unigramas relevantes dos não relevantes? Uma outra parte do trabalho desenvolvido no âmbito desta dissertação consiste também na construção de um extractor de URs a partir de listas ordenadas de relevância.

1.1.3 Aplicabilidade dos Unigramas Relevantes

As potenciais aplicações dos Unigramas Relevantes constituíram outra motivação forte para a realização desta dissertação. Com efeito, a utilização de Unigramas Relevantes em conjugação com as Unidades Lexicais Multipalavra Relevantes (bigramas, trigramas, etc.) tem várias aplicações possíveis entre as quais se salientam:

- **A procura em páginas *Web*.** Sendo a *Web* actualmente um grande repositório de informação, o conhecimento dos tópicos das páginas *Web* permitiria aos motores de busca serem mais eficientes nos resultados das procuras, pesquisando essencialmente por relevância de palavra-chave ao invés do método actualmente utilizado (por ocorrência em conjunção com outras técnicas menos selectivas). A *Web* Semântica é um esforço nesse sentido, enquanto a utilização destes métodos permitiria a extracção automática de termos para a classificação de páginas ou construção automática de ontologias.

- **Agrupamento e Sumarização de documentos.** A extracção automática de tópicos descritores de documentos permitiria organizar estes últimos também automaticamente de acordo com o seu conteúdo, expresso nos tópicos descritores. Empresas e organizações públicas que lidam com extensa documentação beneficiariam bastante deste tipo de aplicações.

- **Acesso inteligente à Informação.** Se um motor de busca utilizasse as Expressões Relevantes – que incluem os URs – previamente extraídas dos documentos, poderia apresentar listas de tópicos e sub-tópicos em resposta às *queries* feitas pelos utilizadores. Desta maneira os utilizadores escolheriam apenas os assuntos específicos em que estivessem interessados, em vez de receberem, por vezes, centenas de documentos entre os quais se encontram, ou não, os documentos a que realmente querem aceder. Desta maneira, a eficiência e produtividade na procura de documentos aumentaria significativamente.

- **Outras aplicações:** enriquecimento de dicionários terminológicos, melhoria das traduções automáticas entre línguas, extracção de conceitos complexos em línguas baseadas em ideogramas (por exemplo Chinês), entre outras possíveis aplicações.

1.2 Principais Contribuições deste Trabalho

Esta dissertação, tendo resultado essencialmente num trabalho de investigação, apresenta algumas contribuições; nomeadamente:

- **Uma nova métrica de relevância de unigramas - *Score***
  Foi criada uma nova medida estatística, denominada *Score*, para atribuição de relevância a unigramas. Esta métrica, puramente estatística, baseia-se na análise da vizinhança das palavras. Por ser independente da língua, contexto e frequência, é vantajosa relativamente à generalidade das métricas e métodos correntes.

- **Outra métrica de relevância de unigramas para línguas latinas - *SPQ***
  Esta medida, *Successor-Predecessor Quotient* (*SPQ*), resulta de uma das experiências efectuadas no decurso desta dissertação. Como se verá, esta medida estatística permite obter bons resultados em línguas latinas, como é o caso do Português, Francês, Espanhol e Italiano.

- **O extractor de Unigramas Relevantes - *Método das Ilhas***
  Esta técnica permite avaliar a relevância *booleana* de cada unigrama com base em atributos estatísticos das palavras que ocorrem na vizinhança de cada unigrama. Basicamente, um unigrama é considerado relevante se for tão ou mais relevante que todas as palavras que ocorrem na sua vizinhança imediata. Tem a vantagem de poder trabalhar com uma lista de relevância gerada por qualquer métrica, inclusive as métricas mencionadas na secção 1.1.1.

- **O método da análise das sílabas**
  Como se verá adiante, a relevância ou não das palavras está fortemente ligada à sua estrutura silábica. Com efeito, as palavras consideradas relevantes tendem a ter um determinado número de sílabas, nem muito pequeno nem muito grande. Esta abordagem, que julgo ser original, permite, como se verá, funcionar como um catalisador, melhorando substancialmente os resultados quer dos métodos aqui apresentados, quer dos métodos propostos por outros autores.

- **Independência em relação à língua e em relação às aplicações**
  O carácter estatístico e a não utilização de quaisquer filtros morfossintácticos em

todas as abordagens tornam-nas independentes quer da língua, quer do contexto. Como se verá, também não são privilegiadas quaisquer gama de frequências de ocorrência das palavras. Isto permite extrair e classificar unigramas em várias línguas e com diversas aplicações possíveis.

- **Potencialidade de aplicações**
  Como já foi descrito, são muitas as possíveis aplicações dos Unigramas Relevantes, o que constitui, a meu ver, uma contribuição muito importante.

Ainda como contribuição, no âmbito desta dissertação foi feito inicialmente um *paper*: [Ventura & Silva 07]. Na sequência desta publicação recebi um convite para participar na elaboração de um livro (*Brain, Vision and AI*), escrevendo um capítulo que veio a ter por título *Ranking and Extraction of Single Words in Text*. Agradeço ao Departamento de Informática da FCT/UNL por ter apoiado esta minha participação. Esta última publicação está disponível em *http://www.proof.i-techonline.com/bvai/13_ Ventura.pdf*.

1.3 Organização da Dissertação

Esta dissertação está dividida da seguinte forma: no capítulo 2 serão apresentadas as abordagens que constituem actualmente o "estado da arte" na classificação/avaliação da relevância de unigramas. Este capítulo está dividido em duas secções: abordagens estatísticas e abordagens não-estatísticas. No capítulo 3 serão apresentadas as duas novas métricas de relevância de unigramas. Ainda neste capítulo será apresentada a técnica baseada na análise das sílabas, que como se verá, permite melhorar os resultados de todos os métodos anteriores (inclusive os de outros autores), e será apresentado também o Método das Ilhas que permite decidir acerca da relevância *booleana* dos unigramas extraídos pelos outros métodos. No capítulo 4 serão apresentados e comentados os resultados obtidos pelos métodos aqui apresentados, comparando-os com os resultados obtidos com a implementação dos outros métodos analisados neste trabalho. Finalmente, no capítulo 5 serão apresentadas as conclusões e o trabalho futuro.

Capítulo 2

O Estado da Arte

Com o objectivo de avaliar o actual "estado da arte", neste capítulo são descritas e analisadas as características positivas e as limitações das outras abordagens de avaliação da relevância de unigramas. As abordagens são divididas em dois grupos: as que utilizam métodos estatísticos e as que utilizam outras abordagens essencialmente não-estatísticas, tais como Redes Neuronais, textos anotados e gramáticas, entre outras.

2.1 Abordagens Estatísticas

As abordagens estatísticas ao problema da extracção de unigramas relevantes são todas as que funcionam unicamente utilizando métodos estatísticos. Estas têm normalmente a vantagem de serem mais rápidas na implementação e utilização quando comparadas com as abordagens não estatísticas, no sentido em que, por não dependerem de informação simbólica/morfossintáctica específica, são por isso independentes da língua e do contexto dos textos analisados. Esta independência é obviamente um factor de grande importância. Nesta secção são analisados três dos métodos estatísticos actualmente mais conhecidos: o critério da frequência de *Luhn*, a métrica *Tf-idf* e o método de *Zhou et al.*.

2.1.1 Critério da Frequência de *Luhn*

Luhn, num dos primeiros trabalhos publicados referentes a técnicas de extracção de unigramas relevantes [Luhn 58], sugere um método de classificação de unigramas relevantes baseado na frequência de ocorrência dos termos. De acordo com o autor,

> *"... a justificação de medir a relevância de uma palavra pela frequência de ocorrência é baseada no facto de que um escritor normalmente repete determinadas palavras quando avança na sua argumentação e quando elabora determinados aspectos de um assunto..."*

Luhn justifica assim que este meio de ênfase por parte de um autor pode ser encarado como um indicador de significância. Para além disso, *Luhn* sugere ainda que quanto mais próximo certas palavras se encontrem numa frase, tanto mais significativas elas são. Salienta, por fim, que palavras com uma frequência de ocorrência muito alta podem ser descritas como muito comuns e palavras com uma frequência muito baixa podem ser descritas como muito raras, em ambos os casos pouco relevantes. Portanto, as palavras consideradas relevantes estariam entre duas fronteiras de frequência, superior e inferior, sendo que, neste intervalo, as palavras mais frequentes seriam consideradas as mais relevantes.

No entanto, daqui surgem vários problemas, um dos quais o facto de que existem palavras relevantes muito frequentes que podem, erradamente, ser classificadas como palavras comuns. Por exemplo, no decurso deste trabalho foi verificado em vários testes que das 100 palavras mais frequentes em cada *corpus*, cerca de 35% são consideradas relevantes. Se se tiver em conta que, em média, os *corpora* utilizados neste trabalho rondam as 500 000 palavras, com cerca de 24 000 distintas, facilmente depreendemos que, com o método de *Luhn*, as palavras relevantes dentre as 100 palavras mais frequentes seriam muito provavelmente descartadas. A tabela 2.1 ilustra algumas dessas palavras:

Tabela 2.1: Algumas palavras relevantes dentre as 100 mais frequentes.

Palavra	Posição na Lista	Frequência
Comissão	28	1909
Estados-Membros	38	1378
Países	41	1219
Europeu	55	874
União	92	515
Europa	99	463

O critério de *Luhn*, torna-se, neste caso, demasiado restritivo, pois como se verifica, estas palavras são bastante descritivas, ainda para mais sendo o *corpus* donde estas palavras foram retiradas baseado em textos provenientes de documentos da União Europeia.

Outro problema da abordagem de *Luhn* surge com o facto de esta encarar como relevantes as palavras na zona intermédia de frequências. Pela tabela 2.2, que mostra um conjunto aleatório de palavras retiradas das zonas intermédias de frequência, podemos verificar que tal abordagem não se mostra eficiente pois, nessa tabela, existem poucas ou nenhumas palavras descritoras dos textos do *corpus* analisado (textos da União Europeia). Aparentemente, *Luhn* resolve parcialmente este tipo de problemas recorrendo a uma lista de palavras comuns que devem, posteriormente, ser retiradas da lista de relevância final. No entanto, convém salientar que a abordagem de *Luhn* foi idealizada para textos com cerca de 700 palavras distintas (artigos científicos), tornando-se muito difícil manter uma

Tabela 2.2: Lista aleatória de palavras na zona intermédia de frequências.

Palavra	Posição na Lista	Frequência
enquanto	509	131
volume	533	124
operacionais	680	103
quer	761	91
prevê	816	84
Sapard	817	84
sendo	862	79
testes	1035	64
potenciais	1075	62
elevada	1081	61

lista de palavras comuns quando se lida com *corpora* de aproximadamente 24 000 palavras distintas e impraticável para *corpora* muito maiores.

Por fim, surge a questão dos limiares de ocorrência. Quais deverão ser? *Luhn*, no seu artigo, apenas sugere que estes devem ser procurados usando-se métodos estatísticos de modo a estabelecerem-se "limites de confiança". Em meu entender, esta justificação é bastante vaga, e de certa forma leva a crer que a experiência é a única maneira de estabelecer os limiares de frequência. Esta situação não é problemática quando se analisa o mesmo tipo de documentos, com a mesma estrutura, mas quando se analisam documentos de tipos diferentes, deixa de ser praticável.

2.1.2 *Tf-idf*

O *tf-idf* [Salton & Buckley 88] (*Term Frequency - Inverse Document Frequency*) é uma métrica de cálculo de relevância de termos bastante utilizada nas áreas de Recuperação de Informação e de *text-mining*. Essencialmente, esta técnica mede o quão importante uma determinada palavra é num determinado documento em relação a outros documentos na mesma colecção ou *corpus*. Basicamente, e de acordo com os autores, uma palavra torna-se mais importante para um determinado documento quantas mais vezes ela ocorrer nesse documento. Por outro lado, se essa palavra ocorrer noutros documentos, a sua importância decresce. Portanto, esta medida tem uma importância local. Palavras que são muito frequentes num único documento ou conjunto restrito de documentos, tendem a ser mais valorizadas em relação a palavras mais comuns que ocorrem em mais documentos, como artigos e preposições.

O procedimento formal para a implementação do *tf-idf* muda ligeiramente de aplicação para aplicação, mas a aproximação mais comum é abordada neste trabalho. Geralmente o cálculo do *tf-idf* divide-se em duas partes, calculando-se os componentes *tf* e *idf* em

separado, aplicando-se depois uma multiplicação entre as duas componentes para se obter o valor *tf-idf* final.

A componente *tf* (*term frequency*) mede simplesmente o número de vezes que uma determinada palavra ocorre num determinado documento. Esta contagem é depois normalizada para prevenir que as palavras em documentos muito extensos obtenham valores de *tf* muito elevados e, em consequência, pouco rigorosos em relação a outros documentos mais reduzidos. A equação 2.1 mede, portanto, a probabilidade de um termo i ocorrer num documento j.

$$tf_{\mathrm{i,j}} = \frac{n_{\mathrm{i,j}}}{\sum_k n_{k,j}} \ , \tag{2.1}$$

onde $n_{i,j}$ é o número de vezes que o termo i ocorre no documento j; o denominador desta equação denota o somatório da frequência de todas as palavras do documento, isto é, por outras palavras, o tamanho do documento j.

A componente *idf* mede a importância geral de um determinado termo. Basicamente consiste na contagem do número de documentos em que esse determinado termo ocorre. A equação 2.2 mede a importância geral de um termo t_i em todos os documentos de uma colecção.

$$idf_{\mathrm{i}} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \ , \tag{2.2}$$

onde $|D|$ representa o número total de documentos no *corpus* ou colecção, e $|\{d_j : t_i \in d_j\}|$ o número de documentos onde o termo t_i ocorre pelo menos uma vez, isto é, $n_{i,j} \neq 0$.

Recorrendo-se às equações 2.1 e 2.2, mede-se a importância de um termo i num documento j, a partir do cálculo do $tf\text{-}idf_{i,j}$ como na equação 2.3.

$$tf\text{-}idf_{\mathrm{i,j}} = tf_{\mathrm{i,j}} \, . \, idf_{\mathrm{i}} \ . \tag{2.3}$$

O *tf-idf* é, como se vê, uma medida bastante elegante, simples, e como tal, fácil de implementar. É também, de acordo com os seus autores e com outros que o implementaram, um método bastante eficaz. No entanto, temos de ter em conta que o objectivo deste método não é o de analisar a relevância de uma palavra num *corpus*, mas sim analisar a relevância local a um determinado documento. Isto é, não se pretende com este método caracterizar a relevância de uma determinada palavra em relação a todos os documentos, mas simplesmente descobrir em que documentos é que essa palavra pode ser considerada relevante. Deste modo e com o fim de se avaliar a prestação desta métrica, foi necessário proceder a uma alteração para generalizar o método, o que basicamente consistiu em atribuir a cada palavra dos *corpora* analisados, o máximo valor *tf-idf* encontrado em to-

dos os documentos constituintes do *corpus*. Os resultados de tal aplicação podem ser verificados no capítulo 4.

Porém, por análise das equações 2.2, 2.3 e por constatação dos resultados obtidos no capítulo 4, verifica-se que também a métrica *tf-idf* sofre de algumas desvantagens quando se analisa *corpora*. Com efeito, à semelhança do critério de frequência de *Luhn* (ver secção 2.1.1), a *tf-idf* prejudica bastante as palavras relevantes mais frequentes, pois estas tendem a existir em quase todos os documentos, e como tal, a sua importância *idf* decresce, dando origem a um baixo valor de *tf-idf*. Por outro lado, o critério *idf* também não leva em conta a probabilidade com que uma determinada palavra ocorre noutros documentos, sendo relativamente «cega» a este facto. Por exemplo, se tivermos um *corpus* constituído por três documentos, em que a palavra "Europa" ocorra num deles 100 vezes e 1 vez em cada um dos outros, o critério *idf* prejudica bastante o valor *tf-idf* desta palavra, dando-lhe um valor *idf* de $\log \frac{3}{3} = 0$ e $tf\text{-}idf = 0$, quando é notório que essa palavra é, provavelmente, muito relevante no documento onde ocorre 100 vezes. Mantendo ainda o exemplo, o facto de essa palavra ocorrer 1 ou 50 vezes nos outros documentos é irrelevante para o *tf-idf*, no entanto essa palavras tem uma importância diferente em cada caso, facto que não é captado por esta métrica.

Por fim, é também visível que a componente *idf* tende a beneficiar em excesso as palavras raras, independentemente da sua relevância, pois se por exemplo num dos documentos existir um erro ortográfico e esse erro não existir em qualquer outro documento, essa palavra terá um valor bastante elevado de *idf* e consequentemente de *tf-idf*, o que não é correcto.

2.1.3 Método de *Zhou et al.*

O método de *Zhou et al.* [Zhou 03] é uma métrica relativamente recente de cálculo da relevância de unigramas em textos. Basicamente, e na mesma linha seguida pelos métodos abordados anteriormente, os autores assumem que a relevância de determinadas palavras está essencialmente relacionada com a sua tendência para formar grupos (*clusters*). Desta forma, assumem que as palavras relevantes devem encontrar-se em determinadas áreas dos textos, quer por fazerem parte de tópicos locais ou estarem relacionadas com esses contextos locais, formando assim *clusters* nessas zonas. Por outro lado, as palavras mais comuns, logo menos relevantes, devem ocorrer aleatoriamente em todo o texto, não formando, por isso, *clusters* significativos.

Esta técnica, sendo um melhoramento e extensão da técnica criada por *Ortuño et al.* [Ortuno et al. 02], mede a relevância de uma determinada palavra de acordo com a análise da posição da cada ocorrência dessa palavra no texto. O procedimento para o

cálculo desta métrica num determinado texto consiste em se obter inicialmente uma lista da forma $\widehat{L}_w = \{-1, t_1, t_2, \ldots, t_m, n\}$, em que t_i representa a posição da i-ésima ocorrência da palavra w no texto e n representa o número total de palavras nesse texto.

A partir da lista $\widehat{L}_w$, recorrendo-se à equação 2.4 obtem-se $\widehat{\mu}$, que representa a separação média entre ocorrências consecutivas da palavra w no texto.

$$\widehat{\mu} = \frac{n+1}{m+1} \ . \tag{2.4}$$

O próximo passo consiste no cálculo da separação média de cada ocorrência particular da palavra w, relativamente às suas duas ocorrências vizinhos mais próximas: a anterior e a seguinte; de acordo com a equação 2.5. De certa forma, é nesta equação que se obtém informação acerca das sequências locais de $\widehat{L}_w$.

$$d(t_i) = \frac{t_{i+1} - t_{i-1}}{2}, \qquad i = 1, \ldots, m \ . \tag{2.5}$$

O passo seguinte consiste na identificação dos pontos de $\widehat{L}_w$ que fazem parte de *clusters*. Basicamente, um ponto fará parte de um *cluster* se a sua distância média $d(t_i)$ for inferior à distância média entre ocorrências da mesma palavra, representada por $\widehat{\mu}$. Portanto, um ponto t_i fará parte de um *cluster* se $d(t_i) < \widehat{\mu}$.

Desta forma, obtem-se $\delta(t_i)$ que, de acordo com a equação 2.6 identifica os pontos pertencentes a *clusters*.

$$\delta(t_i) = \begin{cases} 1, & \text{se } t_i \text{ pertencer a um cluster.} \\ 0, & \text{caso contrário.} \end{cases} \ . \tag{2.6}$$

De forma paralela, procede-se ao cálculo de $v(t_i)$ como na equação 2.7, que representa o excesso local de palavras em relação à posição t_i. Basicamente consiste em calcular a distância à media normalizada.

$$v(t_i) = \frac{\widehat{\mu} - d(t_i)}{\widehat{\mu}} \ . \tag{2.7}$$

Analisando-se a equação anterior, verifica-se que quanto menor o valor de $d(t_i)$, que representa a distância média entre as ocorrências t_{i-1} e t_{i+1}, maior o valor de $v(t_i)$. Essencialmente é o mesmo que dizer que, quanto mais próximo a ocorrência t_i estiver das ocorrências que lhe precedem e sucedem, tanto mais valorizada deverá ser essa ocorrência, pois relembra-se que o objectivo essencial deste método é valorizar a formação de *clusters*. Por fim, a pontuação dada à palavra w analisada utiliza os valores provenientes das equações 2.6 e 2.7, como se verifica na equação 2.8.

$$\Gamma(W) = \frac{1}{m} \sum_{i=1}^{m} \delta(t_i) \, . \, v(t_i) \ . \tag{2.8}$$

Desta forma, partindo da lista inicial $\widehat{L}_w = \{-1, t_1, t_2, \ldots, t_m, n\}$, em que t_i representa a i-ésima ocorrência da palavra W num texto, obtemos o valor de $\Gamma(W)$. Estando em $\delta(t_i)$ a informação acerca de t_i pertencer ou não a um *cluster*, e em $v(t_i)$ a distância à media normalizada.

Este método, sendo um método bastante engenhoso e computacionalmente eficiente de se procurar *clusters* sofre, no entanto, do mesmo problema dos métodos anteriores no que importa à penalização das palavras relevantes muito frequentes. Generalizando, este problema ocorre em todas as abordagens que assentam na premissa de que determinados assuntos têm maior representação em determinadas áreas dos textos. Não deixando de ser uma Assunção verdadeira, prejudica as palavras relevantes muito frequentes, pois estas tendem a ocorrer em todo o texto e não apenas em contextos locais. E descurar palavras relevantes muito frequentes tende a empobrecer o processo de extracção de unigramas, pois, como se já viu na secção 2.1.1, mais concretamente na tabela 2.1, as palavras relevantes muito frequentes são muito descritoras de todo o texto e não de um tópico em específico. Por outro lado, ao lidar exclusivamente com *clusters* significativos, as palavras relevantes com frequências de ocorrência muito baixas são também bastante prejudicadas por este método.

2.2 Outras Abordagens

Nesta secção são apresentadas muito sucintamente outras abordagens à problemática da extracção de Unigramas Relevantes. Geralmente os métodos não-estatísticos sofrem do problema de serem dependentes da língua ou do contexto, tornando-se, por isso, candidatos pouco interessantes à extracção de URs. No entanto, por serem muito específicas, têm normalmente na sua área de aplicação, bons resultados.

2.2.1 Linguísticas

As abordagens linguísticas permitem extrair unidades lexicais a partir de documentos, utilizando informação linguística (morfológica, sintáctica ou semântica) acerca de um determinado texto. Normalmente essa informação linguística provém ou da utilização de gramáticas ou da anotação dos textos, ambas feitas manualmente ou recorrendo a métodos automáticos de anotação. As ferramentas que utilizam abordagens linguísticas tendem a extrair termos ou combinações de termos associados a determinados *parts-of-speech*, ou

seja, partes importantes da língua tais como nomes e adjectivos, ou combinações destes para formar n-gramas.

O grande problema associado às abordagens linguísticas prende-se com o facto de na maior parte dos casos se utilizar algo externo ao próprio texto a analisar, nomeadamente gramáticas ou textos anotados. Desta forma, as abordagens linguísticas são extremamente dependentes de uma linguagem ou de contextos específicos, não sendo fácil a sua adaptação para outras linguagens ou situações muito diferentes.

Nos casos em que se utilizam gramáticas, estas são normalmente divididas em gramáticas dependentes do contexto e gramáticas independentes de contexto. Em [Taniza 01], a autora compara a utilização de gramáticas de modo a conseguir extrair sequências de nomes e adjectivos (unigramas e bigramas). Nesse trabalho são utilizadas gramaticas independentes de contexto e gramáticas probabilísticas independentes do contexto geradas automaticamente com um *parser* treinado a partir de um *corpus* Inglês anotado.

$np \rightarrow det\ noun$	[0.148146]
$np \rightarrow noun$	[0.123495]
$np \rightarrow nounp$	[0.064068]
$np \rightarrow noun\ noun$	[0.066792]
$np \rightarrow det\ adj\ noun$	[0.049849]
$np \rightarrow adj\ nounp$	[0.039799]
$np \rightarrow det\ nounp$	[0.025704]
$np \rightarrow adj\ noun$	[0.025916]
$np \rightarrow noun\ nounp$	[0.025638]
$np \rightarrow det\ noun\ noun$	[0.032025]
$np \rightarrow \#$	[0.017428]

Figura 2.1: Regras obtidas para uma gramática probabilística independente do contexto.

Após a construção das gramáticas, faz-se a análise aos textos de *input* e, no caso da utilização da gramática probabilística, o *parser* (algoritmo de procura em gramáticas) tenta encontrar qual a regra mais provável que associa o texto à gramática. Desta forma, torna-se fácil extrair os conteúdos mais importantes como adjectivos e nomes, bastando executar determinadas instruções quando, por exemplo na figura 2.1, as regras *np* forem despoletadas, visto que todas contêm nomes e/ou adjectivos. No entanto, a utilização de Gramáticas para a extracção de informação traz o problema da dependência em relação a uma língua e, como tal, a sua difícil generalização. A questão de recorrer a *corpora* anotados para a construção automática poderá servir como ponto de partida para uma maior generalização destes métodos, mas, novamente, a construção de *corpora* anotados é algo que é feito manualmente ou de forma automática muito imperfeita.

Por fim, existem as análises baseadas em textos anotados sem utilização de gramáticas. Em [Heid 99] o autor utiliza *corpora* previamente processados contendo resultados

tais como separação de palavras e frases, anotação gramatical, entre outros. A extracção de termos candidatos é depois feito recorrendo à utilização de expressões regulares sobre os caracteres (de modo a encontrar abreviações, etc.), sobre palavras (para identificar candidatos unipalavra relevantes) e sobre sequências de posições onde ocorrem nomes, adjectivos e verbos de modo a extrair termos multipalavra relevantes. No entanto isto é feito apenas para a língua Alemã, e a contemplação de cerca de 20 sufixos e 20 prefixos a analisar com expressões regulares de modo a encontrar abreviações mostra o quão dependente da língua se encontra este método.

2.2.2 Baseadas no Conhecimento

As abordagens baseadas no conhecimento são abordagens que estão normalmente associadas a ontologias (especificadas ou não), e onde a ideia é obter um modelo representativo da realidade em questão. Os exemplos mais simples de extracção de termos, neste caso, estão associados ao conhecimento da estrutura dos documentos, por exemplo, extrair palavras-chave a partir dos títulos e resumos de artigos científicos. Alguns exemplos mais exigentes, como o de *Gao & Zhao* [Gao & Zhao 03], permitem analisar se um *email* proveniente de um contacto desconhecido é uma fraude ou não.

Por utilizarem ontologias, as abordagens baseadas no conhecimento são bastante limitadas, pois a construção de uma ontologia (ou domínio de conhecimento) é algo muito específico a um determinado tema. Por exemplo, o caso da extracção de palavras-chave a partir dos títulos ou resumos de artigos científicos é algo muito específico aos artigos científicos. Na utilização de documentos sem estrutura aparente, torna-se praticamente impossível um método de extracção de termos baseado no conhecimento. Por outro lado, a própria construção das ontologias por meios automáticos é também um problema difícil de resolver actualmente o que torna estes métodos pouco generalistas e abrangentes.

2.2.3 Baseadas em Redes Neuronais

Uma rede neuronal artificial, no caso das ciências informáticas, é um modelo de programação que se assemelha, de certa forma, ao modelo neuronal biológico. Consiste num grupo de neurónios artificiais que processam a informação e passam-na para outros neurónios artificiais. A ligação entre todos os neurónios permite formar uma rede de grande poder computacional.

No âmbito do trabalho realizado nesta dissertação, as Redes Neuronais são normalmente utilizadas para responder a questões colocadas pelo utilizador. O exemplo de aplicação mais comum está associado ao processamento de documentos e comparação

com palavras de procura de um utilizador. Trata-se portanto de verificar se as palavras introduzidas por um utilizador são relevantes num determinado texto ou não.

Em [Das et al. 02], os autores sugerem um método baseado em Redes Neuronais. O funcionamento básico é o seguinte: a rede possui vários nós. Cada nó tem associado uma palavra (dos termos pesquisados por um utilizador), com o mesmo valor inicial. É dado como *input* ao modelo um artigo, e se houver concordância entre uma palavra do artigo e uma palavra da rede, o valor do nó correspondente à palavra é elevado a um nível de «energia» superior. O valor de energia está também dependente da posição da palavra no artigo. Este processo continuará até que a rede estabilize num determinado nível de energia dos seus nós. O grupo de nós activos e a energia desses nós dará o valor de relevância desse artigo associado às palavras procuradas.

Apesar deste método não partir directamente do processo de extracção de termos, utiliza uma das aplicações possíveis desta, pois permite, por exemplo, classificar os documentos com base em determinadas palavras. No entanto, este processo tem alguns problemas. Com efeito, as Redes Neuronais baseadas em *software* tendem a ser lentas devido ao cálculo do valor de retropropagação, ou seja, a actualização de todos os valores das ligações entre nós. Dessa forma, o fornecimento a uma rede neuronal de um documento com cerca de 15 000 palavras fará com que a sua análise seja demasiado lenta. Se tivermos em conta a aplicabilidade a um sistema que tenha de analisar tantos documentos como os que existem actualmente na *internet* (ou mesmo 1/4 destes), veremos que é uma tarefa quase impossível de se realizar em tempo útil.

2.2.4 Híbridas

Por fim surgem as soluções híbridas que tentam juntar o melhor de várias soluções. Por exemplo, em [Feldman et al. 06], os autores utilizam gramáticas probabilísticas independentes do contexto em conjunção com métodos estatísticos para a extracção de informação a partir de páginas *web* de modo a convertê-las em páginas da *Web* Semântica. Neste caso, as regras da gramática probabilística são criadas manualmente enquanto as probabilidades são obtidas a partir de um *corpus* anotado. Novamente nesta situação existe uma sobredependência em relação ao *corpus* utilizado, para além de um método manual de criação de regras ser algo pouco escalável.

Capítulo 3

Uma Contribuição Alternativa

Neste capítulo é proposto um conjunto de alternativas inovadoras em relação aos métodos anteriores, incluindo um extractor de unigramas. Em suma, são propostas duas novas métricas para o cálculo da relevância de unigramas, a medida *Score* e a medida *SPQ*, e um novo extractor de Unigramas Relevantes denominado Método das Ilhas. Esta dissertação propõe também um novo campo de investigação relacionado com a análise silábica das palavras, dada a sua influência na relevância dos unigramas, como iremos ver.

3.1 Acerca da Relevância

Partindo de um *corpus* composto de vários documentos, um dos objectivos deste trabalho é tentar perceber quais as palavras que são relevantes e quais não o são. No entanto, usando métodos puramente estatísticos, essa classificação não é imediata nem sequer exacta, pois apesar de a noção de relevância ser um conceito fácil de entender, normalmente não há consenso sobre a fronteira que separa a relevância da não-relevância. Por exemplo, é consensual que as palavras como "República" e "Londres" têm uma relevância significativa e palavras como "desde" e "e" não têm; mas não há consenso relativamente à relevância de palavras como "ler", "terminar" ou "próximo". Há portanto uma fronteira difusa quanto à relevância de palavras.

Dado que este assunto se reveste de uma enorme subjectividade, a Prof. Maria Francisca Xavier do Departamento de Linguística da FCSH/UNL disponibilizou-se para estabelecer um critério de relevância. Em apêndice (ver apêndice B) estão disponíveis duas listas que ilustram o critério de classificação da Prof. Maria Francisca Xavier, tanto para a língua Portuguesa como para a Inglesa. No âmbito desta dissertação, este critério foi considerado de *referência* e por isso foi com base nele que todos os testes e resultados foram obtidos.

Posteriormente, surgiu também a hipótese de associar a relevância de unigramas a uma ou mais classes morfológicas. Neste sentido, a classe dos Nomes pareceu ser a mais "próxima" da relevância. No entanto, este não é nem um critério completo nem absolutamente coerente, conforme foi confirmado pela Prof. Maria Francisca Xavier. De facto, se por um lado nomes como "lado" e "objecto", entre muitos outros, não são verdadeiramente relevantes, por outro, adjectivos como "contaminado", "assassinado" ou "corrosivo", ou mesmo advérbios como "assustadoramente" estão longe de ser irrelevantes. Assim, e de acordo com o critério ilustrado no apêndice B, no contexto deste trabalho caracterizam-se como relevantes todas e quaisquer palavras com conteúdo semântico, independentemente da classe morfológica a que pertençam.

3.2 A Métrica *Score*

Um dos primeiros passos para a extracção de Unigramas Relevantes consiste em obter uma lista ordenada pela potencial relevância de cada uma das palavras do *corpus*. Esta lista deverá medir a relevância relativa de cada palavra em relação às outras palavras que ocorrem no texto. Portanto, uma palavra que é classificada no topo da lista é, provavelmente, mais relevante que uma palavra classificada no fim da lista. Para este efeito, foi criada uma nova métrica em que a ideia principal é que as palavras relevantes têm, normalmente, uma apetência especial para se relacionarem com um pequeno conjunto de outras palavras. Assim, esta métrica é dividida em duas componentes, onde a primeira mede a importância de uma palavra num *corpus* baseado no estudo da relação entre essa palavra e as palavras que lhe sucedem no texto. Denominamos essa componente, o *Score por sucessor* de uma palavra w, ou seja, $Sc_{\text{suc}}(w)$.

$$Sc_{\text{suc}}(w) = \sqrt{\frac{1}{\|\mathcal{Y}\| - 1} \sum_{y_i \in \mathcal{Y}} \Big(\frac{p(w, y_i) - p(w, .)}{p(w, .)}\Big)^2} \quad . \tag{3.1}$$

Em 3.1, $\mathcal{Y}$ representa o conjunto de palavras distintas que ocorrem no *corpus*, e $\|\mathcal{Y}\|$ o tamanho do conjunto, ou seja, o número de palavras distintas no *corpus*; $p(w, y_i)$ representa a probabilidade de y_i ser um sucessor da palavra w; $p(w, .)$ representa a probabilidade média de ocorrerem sucessores de w, que é dada por:

$$p(w, .) = \frac{1}{\|\mathcal{Y}\|} \sum_{y_i \in \mathcal{Y}} p(w, y_i) \qquad p(w, y_i) = \frac{f(w, y_i)}{N} \quad , \tag{3.2}$$

onde N representa o número total de palavras no *corpus* e $f(w, y_i)$ é a frequência de ocorrência do bigrama (w, y_i) no mesmo *corpus*.

Resumindo o formalismo matemático, $Sc_{\text{suc}}(w)$ é dado por um desvio padrão «nor-

malizado» pela probabilidade média de sucessores de w. Mede, portanto, a variação da *preferência* da palavra w em ocorrer antes das restantes palavras do *corpus*. Os valores mais elevados surgem para as palavras que têm uma maior diversificação de frequências com as palavras que lhes sucedem, e os menores valores surgem para as palavras que têm menor diversificação de frequência com as palavras que lhes sucedem. Similarmente, mede-se a *preferência* que uma palavra w tem com as palavras que lhe antecedem usando a seguinte métrica designada por *Score por antecessor* ou $Sc_{\text{ant}}(w)$.

$$Sc_{\text{ant}}(w) = \sqrt{\frac{1}{\|\mathcal{Y}\| - 1} \sum_{y_i \in \mathcal{Y}} \Big(\frac{p(y_i, w) - p(., w)}{p(., w)}\Big)^2} \ , \tag{3.3}$$

onde $p(y_i, w)$ representa a probabilidade de y_i ser um antecessor da palavra w e $p(., w)$ representa a probabilidade média de antecessores de w.

Usando então as expressões (3.1) e (3.3), através da média aritmética, obtém-se uma métrica que permite classificar a relevância de uma palavra baseado nos resultados dos antecessores e sucessores dessa palavra. Essa medida é a medida *Score* e representa-se por $Sc(w)$.

$$Sc(w) = \frac{Sc_{\text{ant}}(w) + Sc_{\text{suc}}(w)}{2} \ . \tag{3.4}$$

Verifica-se pelas expressões anteriores, que a medida *Score* atribui maior valor a uma palavra quando esta tem tendência para se ligar a um conjunto restrito de palavras antecessoras e sucessores. Com efeito, uma análise cuidada à estrutura das equações 3.1 e 3.3 permitirá ao leitor concluir que esta métrica atribui um valor máximo aos unigramas que só se ligam a um único sucessor e a um único antecessor. Esta valorização da relação de exclusividade está de acordo com o que se pretende, principalmente quando o unigrama tem uma frequência igual ou superior a 2, já que quando ocorre no *corpus* apenas uma vez, isso pode dever-se a um erro ortográfico e, embora se trate de um caso particular de exclusividade, não se trata de um unigrama relevante. A métrica porém, não pode, obviamente, captar esta excepção. Assim, ao se utilizar essa métrica, deve proceder-se a uma pré-filtragem de modo a que os unigramas que ocorrem apenas uma vez sejam descartados.

A métrica *Score* não favorece qualquer gama de frequência de ocorrências. A título de exemplo, se um unigrama ocorrer com frequência 20 no *corpus*, ser-lhe-á atribuído o mesmo valor de *Score* que é atribuído a outro unigrama que ocorre apenas 2 vezes, desde que em ambos os casos os unigramas ocorram apenas com um sucessor e um antecessor. No entanto, tal como acontece na generalidade dos processamentos estatísticos, quando a frequência é mais elevada existe maior fiabilidade nos resultados. O mesmo é dizer

que, para frequências muito baixas, os resultados podem não ser conclusivos, por falta de robustez estatística.

3.3 A Medida *SPQ*

No decorrer do trabalho foi observado que as palavras consideradas relevantes têm normalmente um conjunto de características interessantes quanto ao número de sucessores e antecessores. Por exemplo, no *corpus* Português, composto de aproximadamente meio milhão de palavras (textos provenientes de documentos da União Europeia), foi constatado que a palavra "comissão", considerada relevante, ocorria cerca de 1 909 vezes no *corpus* com cerca de 41 antecessores distintos e 530 sucessores distintos. A palavra "Europa", também considerada relevante, ocorria 466 vezes no *corpus*, com 29 antecessores distintos e 171 sucessores distintos. Em ambos os casos, a maior parte dos antecessores é constituída por artigos ou proposições tais como "a", "na" e "da". De facto, as palavras de função (artigos, proposições, etc.) não mostram nenhuma relação especial com conjuntos limitados de palavras, existindo em todo o *corpus*.

Tabela 3.1: Exemplos de sequências morfossintácticas do *corpus* Português onde ocorre o unigrama "comissão".

a comissão lançou
a comissão considera
a comissão europeia
pela comissão tratada

A sequência morfossintáctica <artigo> <nome> <verbo> é muito comum no caso das línguas latinas, tais como o Português, o Espanhol, o Italiano e o Francês. Nestes casos, como normalmente existem menos artigos do que verbos, é natural que os nomes tenham mais sucessores do que antecessores. A tabela 3.1 é um exemplo do que se passa relativamente ao Português e nela pode-se constatar que a lista de artigos é reduzida ao ser comparada com a dos verbos. Por outro lado, como é sabido, a evolução da língua «produz» mais facilmente verbos novos para utilizar com a palavra "comissão" do que artigos. Seguindo este raciocínio, é proposta uma nova métrica estatística que mede a importância de uma palavra baseada na relação entre o número de antecessores e sucessores distintos. Esta métrica é designada por *SPQ* (*Successor-Predecessor Quotient* ou Quociente Sucessor-Antecessor).

$$SPQ(w) = \frac{Nsuc(w)}{Nant(w)} \ , \tag{3.5}$$

onde $Nsuc(w)$ e $Nant(w)$ representa o número de sucessores distintos da palavra w e o número de antecessores distintos de w.

Desta forma, a métrica 3.5 (SPQ) premeia as palavras que têm um maior número de sucessores e um menor número de antecessores. Isola-se assim, de certa forma, os nomes que, como é natural, possuem normalmente uma relevância maior do que artigos e verbos. No entanto, apesar da métrica ser independente da língua, verifica-se que os resultados são largamente superiores no *corpus* Português do que no *corpus* Inglês. Isto acontece porque, como se sabe, a língua Inglesa é menos restritiva na estrutura morfossintáctica <artigo> <nome> <verbo>. Sendo assim, é preferível utilizar esta métrica apenas com linguagens latinas.

3.4 A Análise Silábica

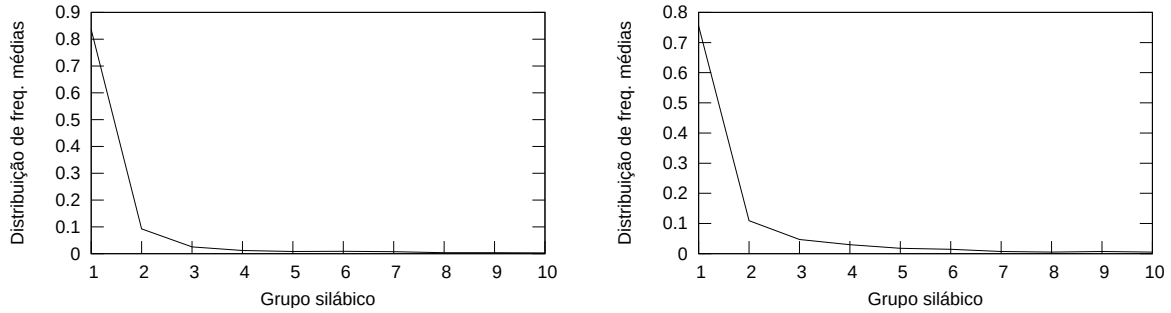
Atente-se à tabela 3.2 que é, para conveniência de leitura, uma reprodução exacta da tabela 1.1 na secção 1.1.2.

Tabela 3.2: Exemplo de uma hipotética lista de relevância.

Palavra	Posição na Lista
comissão	6
europeia	42
Portugal	200
e	2003
ou	2145
da	2415

Pela análise desta tabela, imediatamente se nota que das 6 palavras existentes, 3 são facilmente consideradas relevantes e as restantes não o são. É também fácil concluir que as palavras relevantes ("comissão", "europeia" e "Portugal") são, de facto, mais longas que as palavras não relevantes ("e", "ou" e "da"). Poder-se-ia, perante este facto, conceber uma métrica que beneficiasse as palavras maiores, uma vez que aparentemente estas tendem a ser mais relevantes. Mas no entanto, como será demonstrado posteriormente, é preferível atender ao número de ditongos vocais (ou sílabas) ao invés do comprimento das palavras. É que, por exemplo, se tomarmos em conta a probabilidade de ocorrência do artigo definido "*the*" em textos ingleses, verificamos que esta probabilidade é, aproximadamente, a mesma da sua contraparte portuguesa (o artigo "o"/"a"). No entanto, existe uma relação de 3 para 1 quanto ao número de caracteres. Por isso, uma métrica baseada no número de caracteres beneficiaria 3 vezes mais a palavra "*the*" em relação à palavra "o", quando ambas significam a mesma coisa, mas em linguagens diferentes. Em alternativa, usando

uma métrica baseada na contagem das sílabas, essa distorção não ocorreria, já que ambas as palavras têm o mesmo número (1) de sílabas.



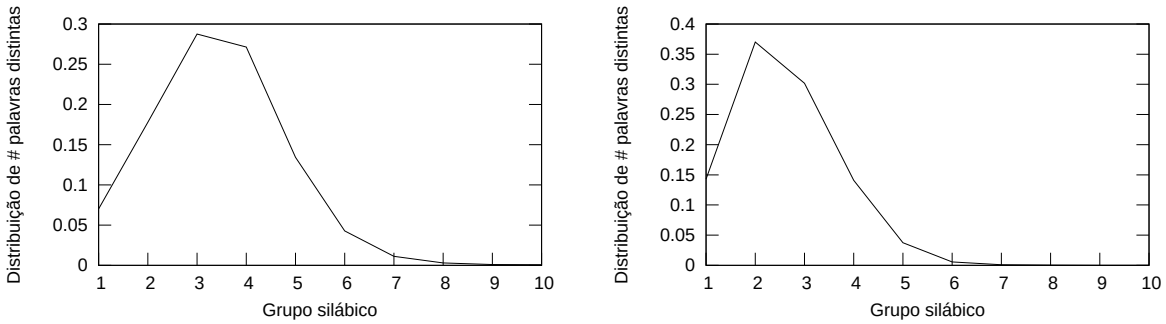
(a) *corpus* Português (b) *corpus* Inglês

Figura 3.1: Distribuição normalizada das frequências médias de ocorrência das palavras de cada grupo silábico, em dois *corpora* testados.

A figura 3.1 mostra a distribuição normalizada de frequências médias de ocorrência para cada grupo silábico, para dois *corpus* trabalhados, Português à esquerda e Inglês à direita; os valores estão normalizadas de modo a que a soma seja 1. Cada um dos gráficos representa, basicamente, a frequência de ocorrência das palavras pertencentes a cada grupo silábico. Desta forma, atentando-se a esses gráficos, depreende-se que ocorrem mais frequentemente nos textos palavras com uma sílaba, seguido das palavras de duas sílabas, etc., até às palavras de dez sílabas que são as mais raras. Em termos de valores, as palavras com uma sílaba ocorrem, em média, cerca de 8.3 vezes mais frequentemente que as palavras de duas sílabas (7.5 vezes no caso do *corpus* Inglês) e cerca de 27 vezes mais frequentemente que as palavras de três sílabas (15 vezes no caso Inglês) — $0.83/0.1 = 8.3$, $0.75/0.10 = 7.5$, $0.83/0.03 = 27$ e $0.75/0.05 = 15$.

Assim, a frequência média de ocorrência das palavras decresce com o aumento do número de sílabas. Este fenómeno está certamente relacionado com a economia dos discursos falados e escritos; é necessário que as palavras que ocorrem mais frequentemente sejam de oralidade e escrita económicas; caso contrário os discursos tornar-se-iam excessivamente longos e pesados. As palavras com 1 sílaba são normalmente artigos e outras palavras de função, tais como "e", "o", "ou" e "da". Porque estas palavras ocorrem mais frequentemente nos textos, é necessário que sejam pois simples e rápidas de pronunciar.

A figura 3.2 mostra os gráficos que representam a quantidade de palavras distintas por cada grupo silábico, para os dois *corpora* testados; os valores estão normalizados de forma a que a soma seja 1. Pelos valores mostrados nos gráficos, existem cerca de 4 vezes mais palavras distintas com 3 sílabas do que palavras distintas com 1 sílaba (2.6 vezes mais palavras distintas com 2 sílabas no caso do *corpus* Inglês). O pico no grupo de 3 sílabas (2 no caso do Inglês) permite-nos dizer que este é, de facto, o *grupo mais popular*.



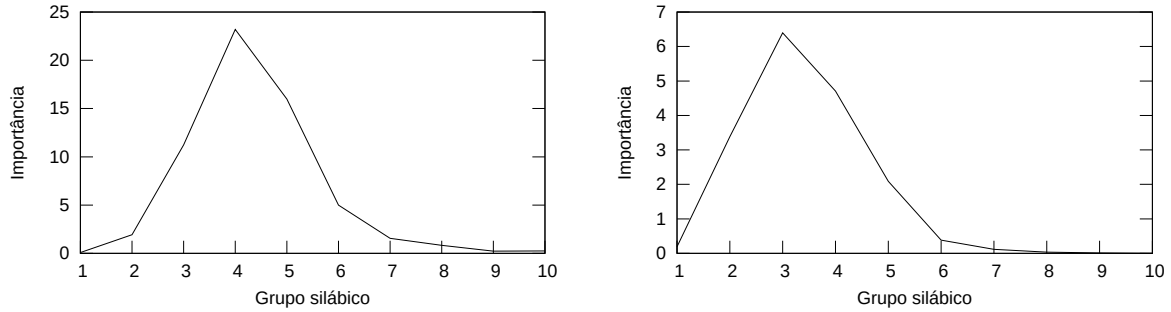
(a) *corpus* Português (b) *corpus* Inglês

Figura 3.2: Distribuição normalizada da quantidade de palavras distintas por grupo silábico, nos dois *corpus* testados.

A interpretação destas curvas está para além do domínio desta dissertação, mas, sem uma certeza absoluta, podemos arriscar a interpretação de que estas distribuições estão provavelmente relacionadas com o número de palavras distintas que se podem produzir preferencialmente com o menor número de sílabas, considerando as sequências *legais* de fonemas que se podem formar em cada língua. De facto, o número de palavras que se pode produzir com 3 sílabas é, com certeza, maior do que o número de palavras que se podem produzir com 2 sílabas. O mesmo ocorre com a quantidade de palavras distintas possíveis com 2 sílabas em relação às palavras de 1 sílaba só. No entanto repare-se que o pico no *corpus* Inglês ocorre num número de sílabas menor do que o *corpus* Português. Isto deve-se provavelmente ao facto de a língua Portuguesa ser – segundo parece – mais restritiva em relação às combinações *válidas/legais* de fonemas e, consequentemente, «esgotar mais rapidamente» o grupo das 2 sílabas, necessitando por isso de ocupar o grupo das 3 sílabas.

A figura 3.2 mostra portanto que existe um pico de palavras distintas em cada língua: no grupo de 3 sílabas no caso do Português e 2 sílabas no caso do Inglês. No entanto, seja qual for a língua analisada, no grupo de 1 sílaba pode-se encontrar sobretudo palavras de função (tais como artigos, proposições, etc.), que não têm normalmente qualquer valor semântico. Por outro lado, as palavras muito raras, que de acordo com a figura 3.2 são as palavras com um número elevado de sílabas, têm valores semânticos demasiado específicos para serem considerados, simultaneamente, relevantes e abrangentes.

Os gráficos da figura 3.3 representam, o que me pareceu poder-se considerar como a importância de cada grupo silábico para cada um dos *corpus* analisados neste trabalho. Assim, para cada grupo silábico, a sua importância é determinada pelo valor correspondente usado no gráfico correspondente da figura 3.2 (a distribuição normalizada do número de palavras distintas, segundo o número de sílabas) a dividir pelo correspondente valor usado na figura 3.1 (distribuição normalizada da frequência média de ocorrência das palavras, segundo o número de sílabas). Se a figura 3.3 fosse utilizada para avaliar a relevância



(a) *corpus* Português (b) *corpus* Inglês

Figura 3.3: Importância de cada grupo silábico, para os dois *corpus* testados.

das palavras num texto, as palavras com 4 sílabas seriam as mais relevantes no caso de um *corpus* Português, e as palavras com 3 sílabas seriam as mais relevantes num *corpus* Inglês, seguindo-se as palavras com 3 e 5 sílabas no caso Português, e 2 e 4 sílabas no caso Inglês. As palavras muito comuns, com 1 sílaba (e 2 no caso Português), seriam consideradas pouco relevantes; as palavras dos restantes grupos silábicos palavras seriam também consideradas pouco relevantes, neste caso por serem bastante raras.

Os gráficos para cada língua que aqui se mostram, não tornam este método dependente da língua, como à primeira vista pode parecer. Com efeito, para obter estas distribuições basta dispor de um *corpus* suficientemente representativo da língua em questão. As especificidades da língua não influem o método que obtém as distribuições, como se torna óbvio pelo que foi explicado atrás.

O número de sílabas é calculado com base nas sequências de vogais e consoantes segundo um algoritmo genérico, o que permite manter a não dependência da língua.

3.5 O Extractor de Unigramas - Método das Ilhas

Como já foi mencionado anteriormente, se usarmos as métricas analisadas neste trabalho, incluindo as propostas de outros autores (ver capítulo 2), apenas é possível obter listas de unigramas ordenados pela sua relevância. Assim, podemos ver se na lista uma palavra é, ou não, mais relevante que outra. No entanto, em determinadas situações é necessário saber da relevância *booleana* de uma palavra. Não se trata então de saber se uma palavra é mais relevante que outra, mas sim se essa palavra é, ou não, realmente relevante.

Numa primeira análise, a solução pode parecer simples. Uma possível abordagem poderia passar por encarar as palavras no topo de uma lista de relevância como relevantes e as do fim da lista como não-relevantes. No entanto, o problema dessa abordagem passaria por encontrar esse limiar de separação. Qual seria este limiar? As experiências

feitas no decurso desta dissertação na procura dum tal limiar não foram encorajadoras. Na verdade, elas mostraram que quando se escolhia um limiar «elevado» no sentido de obter melhor precisão na extracção de Unigramas Relevantes, muitas palavras realmente relevantes eram consideradas não-relevantes. Por outro lado, quando se experimentava um limiar «baixo» para obter boas percentagens de cobertura dos Unigramas Relevantes, acontecia a situação inversa, já que grandes percentagens de palavras não-relevantes eram erradamente consideradas relevantes.

Assim, foi necessário encontrar uma solução alternativa à escolha dum limiar de relevância.

Uma das assumções tomadas para a métrica *Score* (ver secção 3.2) é que uma palavra, para ser considerada relevante, deverá ter uma relação especial com as palavras que lhe sucedem ou antecedem (vizinhas). Na mesma linha de ideias, foi concebida uma abordagem semelhante para o problema da extracção de unigramas. Deste modo, se uma palavra for mais relevante do que as palavras que ocorrem na sua vizinhança imediata (sucessores e antecessores), esta pode ser encarada como realmente relevante. Como já foi mencionado, para verificar se uma palavra é mais relevante que outra basta verificar as suas pontuações na lista de relevância. Este método permite obter máximos locais de relevância. Esta abordagem recebeu o nome de *The Islands Method* – nome usado na publicação [Ventura & Silva 07] feita no âmbito desta dissertação –.

Após alguma investigação, verificou-se que essa nova abordagem, embora simples e com resultados prometedores, era demasiado radical, já que ainda eram descartadas bastantes palavras relevantes. Com efeito, no caso das palavras que faziam parte de bigramas (ou mesmo de n-gramas, com $n > 2$) em que ambos os constituintes eram relevantes, um deles (o de menor relevância) era descartado pela aplicação do método.

3.5.1 O Critério de Selecção

Após várias experiências, procedeu-se então a uma ligeira alteração à abordagem resultante da ideia inicial, com melhores resultados. Assim, de acordo com o Método das Ilhas:

Seja w um unigrama e $r(w)$ o seu valor de relevância dado pela métrica genérica $r(.)$;

se $r(w) \geq 0.9 \,.\, max(Med_{\text{ant}}(w), Med_{\text{suc}}(w))$

então w é um Unigrama Relevante

senão w não é um Unigrama Relevante

fimse

Em que,

$$Med_{\text{ant}}(w) = \sum_{y_i \in \{\text{antecessores de } w\}} p(y_i, w) \,.\, r(y_i) \ . \tag{3.6}$$

$$Med_{\text{suc}}(w) = \sum_{y_i \in \{\text{sucessores de } w\}} p(w, y_i) \,.\, r(y_i) \ , \tag{3.7}$$

onde $Med_{\text{ant}}(w)$ representa a média ponderada dos valores de relevância dos antecessores do unigrama w; $p(y_i, w)$ mede a probabilidade de ocorrência do bigrama (y_i, w) e $r(y_i)$ é o valor de relevância do antecessor/sucessor y_i dado por uma métrica $r(.)$ genérica. Da mesma forma, a equação 3.7 calcula a média ponderada da relevância dos sucessores de uma palavra.

Por outras palavras, de acordo com o Métodos das Ilhas, uma palavra w será considerada Unigrama Relevante *se e só se* a sua relevância fôr maior do que 0.9 da maior das relevâncias médias dentre as vizinhanças antecessora e sucessora; no entanto trata-se de uma média ponderada pela probabilidade de co-ocorrência. Basicamente mede-se o peso que cada vizinho tem enquanto antecessor/sucessor do unigrama em questão, considerando-se a probabilidade de ocorrência do bigrama contituído pelo unigrama e o vizinho. Naturalmente que, os vizinhos com maior peso são aqueles que mais influenciam esta avaliação.

Como se verá no capítulo 4, os resultados deste método são bastante encorajadores. As unigramas que se encontram relativamente «isolados» na lista de relevância em relação aos seus vizinhos são facilmente considerados Relevantes. Por outro lado, devido ao factor 0.9 introduzido no critério de selecção, grande parte dos unigramas que fazem parte de bigramas relevantes já não são descartados, aumentando assim a cobertura do método.

Capítulo 4

Resultados

Neste capítulo são apresentados os resultados referentes a todos os métodos e técnicas mencionados anteriormente. Antes de tudo serão descritos os *corpora* utilizados, assim como os critérios usados para avaliar a qualidade das listas de relevância e do extractor de unigramas, o Método das Ilhas. Em seguida serão apresentados os resultados sobre a qualidade das listas de relevância criadas pelos métodos propostos nesta dissertação (*Score* e *SPQ*), assim como pelos métodos abordados no capítulo 2, nomeadamente as técnicas *TF-IDF* e de *Zhou et al.*. Depois serão apresentados os resultados referentes ao extractor de unigramas, utilizando as listas de relevância criadas anteriormente pelos vários métodos. Por fim será analisada a aplicação do método das sílabas a todas as métricas aqui abordadas.

4.1 Os *Corpora* de Teste

Os *corpora* utilizados e já referidos ao longo desta dissertação, são composto por diversos documentos obtidos no portal para o Direito da União Europeia (http://eur-lex.europa.eu/). Neste portal é possível encontrar um repositório de diversos documentos e comunicações de interesse público no domínio da União Europeia. No âmbito desta dissertação foram usados textos quer em Inglês quer em Português, separadamente, tendo-se utilizado para a criação do *corpus* da língua Portuguesa um total de 43 documentos escritos naturalmente em Português, perfazendo um total de aproximadamente 500 000 palavras, de onde cerca de 24 000 são distintas. No caso do *corpus* para a língua Inglesa foram utilizados 40 documentos em Inglês com um total aproximado de meio milhão de palavras, das quais 18 000 são distintas. Aos dois conjuntos de documentos chamámos respectivamente *corpus* Português e *corpus* Inglês.

Os resultados apresentados ao longo deste capítulo foram obtidos a partir destes *cor-*

pora.

4.2 Critérios de Avaliação

Como já foi mencionado anteriormente (ver secção 3.1), existe uma zona difusa de relevância, onde, de certa forma, a relevância de determinadas palavras pode ser dúbia, ou pelo menos não-consensual. Desta forma importa mencionar que, nas avaliações que serão apresentadas em seguida, foram considerados os resultados de acordo com o critério de referência definido pela Prof[a]. Maria Francisca Xavier da FCSH/UNL. Este critério afectou inevitavelmente os resultados presentes neste capítulo; resultados que podem eventualmente ser considerados conservadores se não for tido em linha de conta o critério de avaliação que acabo de referir.

Nesta secção são explicados outros critérios usados de forma a realizar uma avaliação dos métodos o mais correctamente possível. Assim, serão apresentados os vários testes de qualidade efectuados e os seus propósitos, bem como a forma como foi avaliada a qualidade das listas de relevância geradas pelos vários métodos implementados. Por fim, serão abordados assuntos relacionados com a precisão e cobertura, métricas usadas na avaliação da qualidade do extractor de unigramas.

4.2.1 Conjuntos de Teste

No âmbito da avaliação da qualidade das listas de relevância, os resultados produzidos pelas métricas foram sujeitos a cinco conjuntos de teste constituídos por palavras extraídas a partir dos *corpora*, cada qual com o seu propósito específico. Para uma mais fácil visualização, a tabela 4.1 ilustra os vários conjuntos de teste com a respectiva descrição.

Tabela 4.1: Tabela de Representação dos Conjuntos de Teste de Qualidade.

Designação	Descrição do Conjunto
Teste A	100 palavras mais frequentes.
Teste B	200 palavras extraídas aleatoriamente a partir das 1.000 mais frequentes.
Teste C	300 palavras extraídas aleatoriamente a partir das 3.000 mais frequentes.
Teste D	200 palavras extraídas aleatoriamente de frequência superior a 1.
Teste E	Conjunto de todas as palavras dos conjuntos anteriores.

À primeira vista o Teste D aparenta ser suficiente para avaliar a eficácia das métricas, porque usa um conjunto de palavras extraídas aleatoriamente e de forma independente da frequência. No entanto, convém salientar que a existência dos outros testes surge pelo facto de estes adicionarem informação acerca do comportamento de cada métrica em áreas específicas de frequência de ocorrência.

Desta forma, com o Teste A pretende-se avaliar a eficácia das métricas para as palavras muito frequentes, onde, como foi mencionado anteriormente na secção 2.1.1 e ao contrário do senso comum, se encontram muitas palavras relevantes bastante ilustrativas dos tópicos existentes nos *corpora*. Com os Testes B e C pretende-se avaliar o comportamento das métricas nas zonas intermédias de frequência. Com o Teste D podemos avaliar a eficácia *global* dos métodos, ignorando-se as palavras de frequência 1, grupo que inclui a maioria dos erros ortográficos, naturalmente não-relevantes. Finalmente, no Teste E estão incluídos todos os conjuntos anteriores e é, de certa forma, um conjunto semelhante ao Teste D, mas com uma maior percentagem das palavras mais frequentes das zonas intermédias de frequência.

4.2.2 Avaliação das Listas de Relevância

A avaliação das listas de relevância é algo que pode parecer simples de realizar. Mas na verdade só o é aparentemente. Com efeito, apesar de já existirem alguns trabalhos de investigação sobre unigramas, aparentemente os autores desses trabalhos nunca realizaram uma avaliação quantitativa aos seus resultados, limitando-se a mostrar alguns exemplos soltos das posições em que ficavam determinadas palavras. Esta falta está certamente ligada à dificuldade que naturalmente há para conceber um método de avaliação da qualidade destas listas.

Assim, tendo em conta esta necessidade, para este trabalho foi criado um método de avaliação da qualidade das listas de relevância que funciona da seguinte forma: inicialmente, são identificadas e contadas manualmente as palavras relevantes de todos os conjuntos de teste (A, B, ... E). Depois, a qualidade da ordenação é calculada pelo seguinte critério: se todas as palavras consideradas relevantes surgirem nas primeiras posições da lista em avaliação (independentemente da ordem em que surgirem entre si), estaremos perante uma métrica com 100% de qualidade/eficácia. Mas, se por exemplo, existirem 30 palavras relevantes, mas apenas 25 delas surgirem nas primeiras 30 posições da lista, a eficácia será então de 25/30, o que corresponde a 83.3%.

No fundo, esta medida reflete a «percentagem» correspondente à intersecção entre o conjunto de palavras relevantes e o topo da lista de relevância gerada pela métrica em avaliação – topo de dimensão igual ao conjunto de palavras relevantes –, visto que este é o «local» de classificação onde as palavras relevantes geradas deveriam estar posicionadas, se a métrica fosse perfeita. Com efeito, para uma qualidade de 100%, as N palavras mais relevantes estariam todas colocadas antes das não-relevantes na lista já referida. No pior caso, em que a eficácia seria de 0%, todas as palavras relevantes estariam colocadas de forma errada, ou seja, nenhuma se encontraria no topo da tabela (nas primeiras N

posições).

Esta medida de avaliação das métricas tem, no entanto, uma condição importante a considerar. O que foi referido só é válido quando o número de palavras relevantes é inferior ao número de palavras não-relevantes. Caso isso não aconteça, o critério de avaliação não pode ser o mesmo, de modo a evitar avaliações distorcidas, uma vez que mesmo para o pior caso existiriam sempre palavras a ser classificadas correctamente, não necessariamente por mérito da métrica em avaliação mas eventualmente por aleatoriedade estatística resultante da proporção natural entre palavras relevantes e não-relevantes. No entanto, a solução para este problema é bastante simples: se existirem mais palavras relevantes do que não-relevantes, basta «inverter» o critério, isto é, medir a eficácia com base nas palavras não-relevantes. Assim, neste caso o critério a seguir consiste na contagem das palavras não-relevantes que surgem no fundo da lista. Tome-se o seguinte exemplo: se numa lista de 100 palavras, 90 forem relevantes e consequentemente 10 forem não-relevantes, pelo primeiro método teríamos eficácias mínimas de 88% (mesmo quando todas as palavras não-relevantes estivessem no topo da lista — $80/90 = 0.88(8)$) e máximas de 100%. Se, por outro lado, contarmos quantas palavras não-relevantes surgem no fim da lista, teremos eficácias de 0% a 100%, onde 0% surge quando nenhuma das 10 palavras não-relevantes surge nas 10 últimas posições do fim da lista (logo surgem de certeza nas primeiras 90 posições onde deveriam estar apenas as palavras relevantes), e 100% quando todas as 10 palavras não-relevantes surgem nas 10 últimas posições da lista.

4.2.3 Precisão e Abrangência

A precisão e abrangência são duas medidas estatísticas que servem para avaliar a qualidade dos resultados em domínios tais como a Recuperação de Informação, *Text Mining*, *Data Mining*, etc.. Estas medidas trabalham com informação binária. Neste trabalho, elas serão utilizadas para obter informação quantitativa acerca da qualidade do extractor de unigramas. As suas expressões são as seguintes:

$$\text{Prec} = \frac{\#(palavras\_relevantes \cap consideradas\_relevantes)}{\#consideradas\_relevantes} \ . \tag{4.1}$$

$$\text{Abrg} = \frac{\#(palavras\_relevantes \cap consideradas\_relevantes)}{\#palavras\_relevantes} \ , \tag{4.2}$$

onde: *palavras_relevantes* é o conjunto de palavras verdadeiramente relevantes (classificadas pelo método manual); *consideradas_relevantes* é o conjunto das palavras consideradas relevantes pelo extractor de unigramas; a quantidade de palavras consideradas relevantes pelo extractor e que são ao mesmo tempo realmente relevantes é representada por $\#(palavras\_relevantes \cap consideradas\_relevantes)$.

A precisão pode ser interpretada como a medida de exactidão de uma ferramenta. Basicamente, ela permite medir a proporção do número de palavras realmente relevantes, dentro do conjunto das palavras que o extractor considera relevantes. A abrangência – vulgarmente designada por *recall* na literatura anglo-saxónica – mede a proporção do número de palavras que, considerando o conjunto completo das realmente relevantes, foram detectadas como tal pelo extractor. Existem ambientes onde basta a precisão para avaliar a qualidade do sistema em avaliação. Porém, no caso da avaliação dos resultados gerados pelo extractor de unigramas são necessárias a precisão e a abrangência, já que é necessário avaliar quer quão *correcto* (precisão), quer quão *completo* (abrangência) é o extractor.

As avaliações da precisão e da abrangência são normalmente feitas manualmente. A da precisão é relativamente simples de realizar. O mesmo não se pode dizer relativamente à abrangência. Assim, atendendo ao tamanho dum *corpus* de trabalho de dimensão média, não é praticável fazer uma extracção manual de todos os unigramas relevantes que existem nele, de modo a posteriormente apurar quantas destas palavras foram detectadas pelo extractor. Sendo assim, é necessário colher aleatoriamente uma conjunto de palavras, a fim de obter uma amostra representativa. Esta amostra deve, pois, ser suficientemente grande para ser representativa do *corpus* e suficientemente pequena para que seja fácil extrair manualmente os unigramas relevantes que ela contém. Assim, para avaliação da abrangência foi usada uma amostra aleatória de cada *corpus*, com uma dimensão de 800 palavras distintas.

4.3 Resultados das Listas de Relevância

As tabelas 4.2 e 4.3 mostram os resultados dos vários testes de qualidade aplicados ao métodos propostos no capítulo 3, nomeadamente a medida *Score* (*Sc*) e a medida *SPQ*, e aos métodos propostos por outros autores, especificados nas secções 2.1.2 (*Tf-idf*) e 2.1.3 (método de *Zhou et al.*). As tabelas correspondem à utilização dos *corpora* Português e Inglês, respectivamente.

Tabela 4.2: Qualidade da lista de ordenação de unigramas para o *corpus* Português; valores em percentagem.

Teste	*Sc*	*SPQ*	*Tf-idf*	*Zhou et al.*
A	57.1	67.9	50.0	28.6
B	63.6	66.2	55.8	59.7
C	49.4	51.8	44.7	48.2
D	34.1	27.3	50.0	22.7
E	59.5	58.2	53.6	48.1

Tabela 4.3: Qualidade da lista de ordenação de unigramas para o *corpus* Inglês; valores em percentagem.

Teste	*Sc*	*SPQ*	*Tf-idf*	*Zhou et al.*
A	60.0	62.9	62.9	62.9
B	47.4	52.6	62.9	64.1
C	48.4	53.2	59.7	58.9
D	46.8	45.4	66.7	55.3
E	48.4	58.0	65.4	60.4

De acordo com a tabela 4.2 que representa os resultados da qualidade de ordenação para o *corpus* Português, podemos ler que os métodos propostos nesta dissertação têm, na sua generalidade, resultados superiores aos métodos *tf-idf* e de *Zhou et al.* De notar que para o Teste A, que analisa as palavras relevantes dentre as 100 palavras mais frequentes, os métodos dos outros autores são bastante ineficientes porque, como foi mencionado nas respectivas secções do capítulo 2, estes tendem a prejudicar as palavras relevantes mais frequentes. A métrica *Score* torna-se ineficiente a partir do teste C e quase todos os métodos sofrem uma queda bastante abrupta no teste D que, relembro, é o teste que entra em linha de conta com todas as palavras de frequência superior a 1. Este teste, note-se, é um teste que inclui palavras muito pouco frequentes, sendo por isso um teste bastante difícil para métricas estatísticas. De notar que o método *SPQ* se mantém com os melhores valores para quase todos os testes no corpus Português, excepto o teste D.

No entanto, de acordo com a tabela 4.3 que representa os resultados da qualidade de ordenação para o *corpus* Inglês, nota-se que praticamente todos os métodos têm um aumento ligeiro nos seus valores. Ao contrário do caso Português, os método *Tf-idf* e de *Zhou et al.* são menos ineficientes para as palavras muito frequentes. A medida *Score*, se bem que independente da língua, tem, neste caso, valores médios inferiores aos dos restantes métodos.

4.4 Resultados do Método das Ilhas

As tabelas 4.4 e 4.5 mostram os resultados de precisão e abrangência de acordo com o Método das Ilhas proposto na secção 3.5. As tabelas 4.6 e 4.7 mostram alguns exemplos da classificação, onde o "X" significa que a palavra é considerada relevante pelo critério de referência (o critério de relevância já referido e ilustrado no apêndice B) e os "1" significam que a palavra foi considerada relevante tendo como base a ordenação de relevância do método da respectiva coluna. Relembra-se que o Método das Ilhas classifica os unigramas com base na pontuação dada por uma lista de relevância de unigramas calculada por um dos métodos implementados. Portanto, nas colunas das tabelas pode-se consultar qual o

método de origem das listas de relevância.

Tabela 4.4: Precisão e abrangência para o Método das Ilhas para o *corpus* Português; valores em percentagem.

	Sc	*SPQ*	*Tf-idf*	*Zhou et al.*
Precisão	82.0	80.8	88.8	78.4
Abrangência	86.5	60.1	57.3	76.6

Tabela 4.5: Precisão e abrangência para o Método das Ilhas para o *corpus* Inglês; valores em percentagem.

	Sc	*SPQ*	*Tf-idf*	*Zhou et al.*
Precisão	62.2	66.0	75.4	68.1
Abrangência	76.4	49.1	47.1	75.3

Para ambos os *corpora*, podemos verificar que todos os métodos têm valores bastante bons. Esta variação relativamente baixa da precisão em função das métricas é certamente um indicador de como faz sentido o algoritmo que sustenta o Método das Ilhas. Por outro lado, verifica-se que em termos de abrangência, as métricas *SPQ* e *Tf-idf* apresentam os piores resultados em ambos os *corpora* testados.

Pela tabela 4.5 podemos ver como a precisão e a abrangência são informativas: a métrica *tf-idf*, apesar de ser a que tem um valor de precisão maior (75.4%), tem no entanto um valor muito baixo de abrangência. Significa isto que apesar do método classificar bem as palavras relevantes, elas são poucas em relação ao total de palavras relevantes. Isto, contudo, deve-se não ao Método das Ilhas mas sim à lista de ordenação de relevância criada pelo *tf-idf*, já que para outras métricas foram obtidos valores de abrangência significativamente mais elevados (76.4% e 75.3% respectivamente para *Sc* e *Zhou et al.*).

Tabela 4.6: Exemplos de resultados de classificação — Caso Português

Palavra	Class.	*Sc*	*SPQ*	Tf-idf	Zhou et al.
'das'					
'na'					
'empresas'	X	1	1	1	1
'nas'					
'avaliação'	X	1	1	1	1
'legislação'	X	1	1	1	1
'investigação'	X	1	1	1	1
'segurança'	X	1	1	1	1

Como foi mencionado e por razões de comodidade de leitura, as tabelas 4.6 e 4.7 mostram apenas alguns exemplos de resultados de classificação do Método das Ilhas. No

Tabela 4.7: Exemplos de resultados de classificação — Caso Inglês

Palavra	Class.	*Sc*	*SPQ*	Tf-idf	Zhou et al.
'that'					
'other'					
'directives'	X	1	1	1	1
'ireland'	X	1	1	1	1
'transport'	X		1	1	1
'30'				1	
'21'					

entanto, em apêndice (C) são apresentadas duas listas de 200 palavras aleatoriamente seleccionadas para cada língua, classificadas quanto à sua relevância booleana por este método, permitindo assim uma melhor compreensão dos resultados globais obtidos. As tabelas 4.6 e 4.7 mostram pois que o Método das Ilhas consegue, neste exemplo, separar as palavras relevantes das não relevantes em quase todas as métricas de relevância utilizadas, excepto no caso da palavra "30" que foi considerada relevante usando a métrica de ordenação *Tf-idf*, e no caso da palavra "transport" cuja relevância foi ignorada quando foi usada a métrica *Sc*.

4.5 Aplicação do Método das Sílabas

O Método das Sílabas, proposto na secção 3.4, além de poder ser utilizado como uma métrica autónoma, pode também ser aplicado em conjugação com os outros métodos propostos (nesta dissertação e por outros autores). Neste caso, a aplicação do Método das Sílabas a outros métodos/métricas passa por multiplicar, para uma determinada palavra w, o seu valor obtido por uma qualquer métrica estatística pela importância que o grupo silábico da palavra w adquire de acordo com os gráficos da figura 3.3 na secção 3.5. Nas tabelas 4.8 e 4.9 podemos ler a qualidade dos resultados relativamente ao Método das Sílabas de forma isolada e quando aplicado aos métodos anteriormente analisados (em comparação com os métodos isolados). Nas tabelas 4.10 e 4.11 podemos consultar os resultados de precisão e abrangência após a extracção de unigramas com o Método das Ilhas utilizando as métricas em conjugação com o Método das Sílabas.

Como se pode constatar pelas tabelas anteriores, o Método das Sílabas, para além de ser um bom candidato a métrica isolada para ordenação da relevância de unigramas, quando conjugado com as outras métricas, melhora os resultados destas, como se pode verificar para praticamente todos os casos. Um dos casos mais flagrantes é, por exemplo a subida de 28.6% para 89.3% de qualidade no Teste A do *corpus* Português, para o método de *Zhou et al.*. Inclusivante, os resultados para o Teste D, que em ambos os *corpora* eram

Tabela 4.8: Qualidade da lista de ordenação de unigramas para o *corpus* Português. Influência da aplicação do Método das Sílabas; valores em percentagem.

Método	Teste A	Teste B	Teste C	Teste D	Teste E
Método Sílabas Isolado	78.6	76.6	57.6	81.8	73.0
Sc	57.1	63.6	49.4	34.1	59.5
Sc & Sílabas	82.1	77.9	63.5	84.1	77.6
SPQ	67.9	66.2	51.8	27.3	58.2
SPQ & Sílabas	85.7	77.9	61.2	79.5	75.1
Tf-idf	50.0	55.8	44.7	50.0	53.6
Tf-idf & Sílabas	82.1	75.3	65.9	77.3	75.9
Zhou et al.	28.6	59.7	48.2	22.7	48.1
Zhou et al. & Sílabas	89.3	80.5	65.9	79.5	77.2

Tabela 4.9: Qualidade da lista de ordenação de unigramas para o *corpus* Inglês. Influência da aplicação do Método das Sílabas; valores em percentagem.

Método	Teste A	Teste B	Teste C	Teste D	Teste E
Método Sílabas Isolado	74.3	61.5	59.7	68.1	64.8
Sc	60.0	47.4	48.4	46.8	48.4
Sc & Sílabas	82.9	61.5	64.5	68.1	68.7
SPQ	62.9	52.6	53.2	45.4	58.0
SPQ & Sílabas	74.3	61.5	66.1	68.8	69.5
Tf-idf	62.9	62.8	59.7	66.7	65.4
Tf-idf & Sílabas	77.1	69.2	71.0	73.8	72.0
Zhou et al.	62.9	64.1	58.9	55.3	60.4
Zhou et al. & Sílabas	85.7	65.4	68.5	70.9	70.1

bastante baixos, atingem 81.8% no caso Português e 73.8% no caso Inglês.

Tabela 4.10: Precisão e abrangência para o Método das Ilhas para o *corpus* Português. Influência da aplicação do Método das Sílabas; valores em percentagem.

Método	Precisão	Abrangência
Método Sílabas Isolado	88.9	78.9
Sc	82.0	86.5
Sc & Sílabas	89.3	75.9
SPQ	80.8	60.1
SPQ & Sílabas	91.0	69.4
Tf-idf	88.8	57.3
Tf-idf & Sílabas	93.2	63.7
Zhou et al.	78.4	76.6
Zhou et al. & Sílabas	90.8	77.2

Nas tabelas 4.10 e 4.11 podemos verificar que o Método das Sílabas quando utilizado isoladamente para fornecer a lista ao Método das Ilhas gera, por si só, resultados muito

Tabela 4.11: Precisão e abrangência para o Método das Ilhas para o *corpus* Inglês. Influência da aplicação do Método das Sílabas; valores em percentagem.

Método	Precisão	Abrangência
Método Sílabas Isolado	68.6	80.8
Sc	62.2	76.4
Sc & Sílabas	69.8	76.0
SPQ	66.0	49.1
SPQ & Sílabas	71.8	64.5
Tf-idf	75.4	47.1
Tf-idf & Sílabas	81.1	54.0
Zhou et al.	68.1	75.3
Zhou et al. & Sílabas	71.7	75.3

bons. Note-se que em todos os métodos a precisão sobe, chegando a passar a barreira dos 90% no caso do Português e a barreira dos 80% no caso Inglês. Acerca da abrangência podemos ver que esta sobe na maior parte dos casos quando as métricas são conjugadas com o Método das Sílabas, estando em média nos 75% no caso Português e ligeiramente menos no caso Inglês.

Podemos portanto constatar que também os resultados da relevância booleana dos unigramas são melhorados pelo Método das Sílabas. As listas de resultados para o Método das Ilhas podem ser encontradas no apêndice C, como já foi referido.

4.6 Comentários Gerais sobre os Resultados

Pretende-se com esta secção tecer algumas considerações sobre os métodos propostos nesta dissertação, salientado sobretudo o Método das Ilhas, dado este ser, essencialmente, um dos maiores objectivos deste trabalho e a partir do qual se obtêm os resultados finais.

O método das Ilhas é, como foi mencionado anteriormente, um método que utiliza listas de ordenação de relevância de modo a poder decidir que palavras são relevantes e que palavras não o são. Essencialmente e de uma forma muito simples, segundo o critério das Ilhas, uma palavra é relevante se essa palavra for mais relevante que as palavras que ocorrem na sua vizinhança imediata. Desta forma, na maior parte das situações o método das Ilhas é capaz de identificar correctamente as palavras relevantes nos *corpora* utilizados, com valores médios de Precisão e Cobertura na ordem dos 75%-80% (atingindo valores ainda mais elevados nalgumas situações). No entanto, pelas suas próprias características e pelo facto de utilizar listas de relevância externas ao próprio método em si, existem algumas situações em que determinados erros ocorrem.

De forma a se poder analisar que tipos de erros ocorrem, foi realizada uma triagem de

modo a detectar potenciais casos extremos. Por exemplo, uma palavra relevante que seja considerada relevante pelo método das Ilhas em 8 de 9 métricas (falhando apenas uma), não constitui um caso extremo, pois neste caso apenas uma das 9 métricas não considerou tal palavra mais relevante que outra qualquer na sua vizinhança, erro que deve certamente estar associado a características da própria métrica que falhou. No entanto, ocorrem outras situações tais como palavras relevantes que foram classificadas como irrelevantes na maior parte das métricas e palavras irrelevantes que foram classificadas como relevantes pela maioria das métricas de ordenação. Sendo assim, decidiu-se no âmbito desta secção adoptar o seguinte critério de triagem: foram analisadas as palavras que, sendo relevantes, foram classificadas como tal por 3 ou menos métricas de relevância; e foram também analisadas as palavras que, sendo irrelevantes, foram classificadas como relevantes por 4 ou mais métricas de relevância, num total de 9. A origem deste critério tem simplesmente a ver com o facto de se pretender analisar apenas os casos mais graves na classificação, justificando os erros nos casos menos graves como particularidades das métricas.

A tabela 4.12 mostra alguns dos casos extremos analisado para o *corpus* português.

Tabela 4.12: Lista de exemplos de erros de classificação

Palavra	Classificação	Sílabas	*Sc*	*Sc* & Sil.	*SPQ*	*SPQ* & Sil.	*Tf-idf*	*Tf-idf* & Sil.	Zhou	Zhou & Sil.
'pequenas'		1	1	1			1	1	1	1
'diferentes'		1		1	1	1		1		1
'qual'			1		1		1		1	
'países'	X	1							1	1
'cães'	X		1		1				1	
'civil'	X								1	
'comunidades'	X		1	1						

De acordo com a tabela 4.12, as palavras "pequenas", "diferentes" e "qual" são palavras irrelevantes que foram consideradas relevantes pela maioria das métricas de relevância. Por outro lado, as palavras "países", "cães", "civil" e "comunidades" são palavras relevantes que foram classificadas como irrelevantes pela maioria dos métodos.

Para ilustrar melhor o que sucede no caso das palavras irrelevantes classificadas como relevantes, a tabela seguinte (tabela 4.13) mostra os sucessores e antecessores mais importantes da palavra "qual" e respectivas pontuações de acordo com cada métrica de relevância.

Como se pode verificar, na tabela 4.13 a palavra "qual", que é, no contexto deste trabalho, uma palavra irrelevante, possui na sua vizinhança imediata apenas palavras

Tabela 4.13: Lista de Sucessores e Antecessores da palavra "qual"

Palavra	Sílabas	*Sc*	*Sc* & Sil.	*SPQ*	*SPQ* & Sil.	*Tf-idf*	*Tf-idf* & Sil.	Zhou	Zhou & Sil.
'qual'	0.083	44.384	3.723	2	0.167	0.001	0.003	0.436	0.036
'o'	0.083	17.439	1.462	0.365	0.031	0	0	0.341	0.029
'do'	0.083	12.875	1.079	0.391	0.032	0	0	0.340	0.029
'a'	0.083	14.447	1.211	0.593	0.049	0	0	0.328	0.028
'na'	0.083	14.705	1.233	0.347	0.029	0	0	0.367	0.031
'no'	0.083	16.903	1.417	0.200	0.016	0	0	0.340	0.028
'se'	0.083	44.142	3.702	1.362	0.114	0	0	0.387	0.032
'é'	0.083	17.838	1.496	0.629	0.052	0	0	0.399	0.033
'os'	0.083	22.402	1.879	0.422	0.035	0	0	0.377	0.031
'as'	0.083	19.416	1.628	0.389	0.032	0	0	0.374	0.031

irrelevantes. No entanto, a palavra "qual" possui praticamente em todos os métodos uma pontuação superior a todas as palavras da sua vizinhança, o que de acordo com o método das Ilhas significa que a palavra deve ser considerada relevante, apesar da sua pontuação não ser muito superior às pontuações dos unigramas que constituem a sua vizinhança. Este é, portanto, um dos erros que sucedem no Método das Ilhas: palavras irrelevantes que estão rodeadas por outras palavras irrelevantes com pontuação inferior são classificadas como relevantes. São o que no âmbito do Método das Ilhas se poderia designar por «Ilhéus», palavras irrelevantes que se destacam pouco da vizinhança, também ela irrelevante.

Por outro lado, a tabela 4.14 apresenta os antecessores e sucessores mais importantes da palavra "países", considerada relevante. A palavra "países", apesar de ser considerada relevante, surge no entanto associada a outras palavras com maior pontuação. "Países baixos", "países candidatos" e "países membros" são alguns dos bigramas possíveis com a palavra "países". Significa portanto que outro erro de classificação está associado à ocorrência de bigramas (ou generalizando, n-gramas), onde uma das partes do bigrama é bastante mais cotada do que a outra. Apesar do Método das Ilhas atribuir 90% ao peso que a pontuação da vizinhança tem para o cálculo que decide sobre a relevância booleana duma palavra, em determinados casos (como o da palavra "países") a pontuação das palavras vizinhas é demasiado superior, levando a que a redução para 90% não seja suficiente.

A formação de «Ilhéus» e a questão dos n-gramas são os dois tipos de erros mais comuns de classificação, e ambos estão essencialmente ligados a características próprias

Tabela 4.14: Lista de Sucessores e Antecessores da palavra "países"

Palavra	Sílabas	*Sc*	*Sc* & Sil.	*SPQ*	*SPQ* & Sil.	*Tf-idf*	*Tf-idf* & Sil.	Zhou	Zhou & Sil.
'países'	11.23	43.49	488.68	1.46	16.48	0.00	0.01	0.50	5.64
'os'	0.08	22.40	1.87	0.42	0.03	0	0	0.37	0.03
'dos'	0.08	14.84	1.24	0.40	0.03	0	0	0.37	0.03
'baixos'	1.92	80.78	155.54	5.12	9.86	0.00	0.00	0.49	0.94
'candidatos'	23.19	84.71	1964.94	10.16	235.83	0.00	0.08	0.67	15.77
'terceiros'	11.23	73.48	825.58	6.77	76.14	0.00	0.01	0.47	5.35
'não-membros'	11.23	93.65	1052.16	8	89.87	0.01	0.07	0.89	10.02
'membros'	1.92	74.59	143.62	1.31	2.52	0.00	0.00	0.50	0.97

das métricas de ordenação que lhes dão origem. Uma das soluções poderia passar por reduzir o peso que a vizinhança tem sobre a palavra analisada, baixando de 90% para, por exemplo, 80% ou 70%. Testes realizados nestas condições mostram, no entanto, que neste caso, apesar da Cobertura subir, a Precisão tende a baixar, se bem que não na mesma proporção, mas nem sempre beneficiando o método. Por outro lado, a generalização do Método das Ilhas sugere uma outra possível solução. Na sua forma actual, o Método das Ilhas apenas analisa a vizinhança imediata das palavras. Uma possível melhoria poderia passar por alargar a influência da vizinhança, por exemplo, à 2ª ordem, ou seja, analisando-se as influências da vizinhança à esquerda dos antecessores e da vizinhança à direita dos sucessores.

Capítulo 5

Conclusões

A extracção automática de elementos textuais relevantes (palavras isoladas e multipalavras relevantes) é actualmente uma área de grande aplicação. Entre essas áreas estão a classificação e agrupamento de documentos, indexação de documentos, recuperação de informação (*Information Retrieval*) entre outras áreas do *Text Mining*. Contudo, apesar da extracção automática de expressões multipalavra relevantes ter sido, nos últimos tempos, uma área sujeita a bastante inovação e desenvolvimento, a extracção de unigramas relevantes, talvez por envolver grande complexidade, tem sido uma área descurada pelos investigadores. Como foi demonstrado, descurar unipalavras no processo de extracção de palavras-chave em documentos torna o resultado pobre e pouco realista. Os motores de busca actuais, por exemplo, beneficiariam muito se utilizassem extractores unipalavra e multiplavra, em vez de devolverem resultados baseados simplesmente na ocorrência de termos, como é feito actualmente na maior parte das situações.

Uma das grandes limitações no processo de extracção de unigramas está associado à enorme subjectividade que existe na classificação da relevância de unigramas. Por outro lado, associar à relevância apenas determinadas classes morfológicas como Nomes é, como se demonstrou, incompleto e de pouca precisão pelo facto de existirem outras classes morfológicas que contribuem com palavras relevantes e por existirem nomes que são de conteúdo irrelevante. Com efeito, foram também efectuadas experiências no decorrer deste trabalho utilizando como referência o critério que estabelece que o conjunto das palavras relevantes é o conjunto dos Nomes: os resultados obtidos situaram-se na ordem dos 20% a 30% abaixo dos resultados actuais obtidos pelo critério das palavras de conteúdo que referi ao longo da dissertação e que serviu de referência.

As poucas abordagens existentes no campo dos unigramas têm, como se viu, alguns problemas. Entre eles está por exemplo o facto de prejudicarem gravemente as palavras relevantes frequentes, quando estas são descritoras dos assuntos comuns à grande parte

dos documentos constituintes de um *corpus*. Por outro lado, o facto de estas abordagens apenas criarem listas ordenadas de relevância de unigramas e não decidirem sobre a relevância booleana dos unigramas, constitui uma lacuna importante. Basicamente nenhuma abordagem proposta anteriormente permite extrair, realmente, Unigramas Relevantes.

Quanto ao problema da valorização das palavras relevantes frequentes, foram propostas nesta dissertação duas novas métricas de classificação de unigramas, que são, ao mesmo tempo, independentes da língua, da frequência de ocorrência das palavras e do contexto geral dos documentos analisados. A medida *Score*, que utiliza uma abordagem baseada na co-ocorrência da palavra com a sua vizinhança, permite melhorar bastante os resultados para as palavras relevantes frequentes e mantém-se aceitável para as restantes situações. Por outro lado, também foi criada a medida *SPQ* que, como se viu, tem bons valores de qualidade em línguas latinas.

Em relação à extracção de Unigramas Relevantes, foi proposto nesta dissertação o Método das Ilhas que permite extrair, com bastante sucesso, Unigramas Relevantes a partir de listas ordenadas de relevância. Essas listas podem ser criadas por qualquer métrica vocacionada para este efeito, o que torna o método independente da métrica utilizada. Por outro lado, os bons valores de precisão e abrangência atingidos pelo Método das Ilhas em praticamente todas as listas utilizadas (geradas por métricas diferentes), permite concluir a robustez do critério usado: este algoritmo elege como relevantes as palavras cujo valor de relevância se salienta em relação à sua vizinhança imediata, constituindo por assim dizer máximos locais de relevância. A outra alternativa de extracção de unigramas, como se mostrou na secção 3.5, passaria por tentar encontrar limiares de separação booleana das listas de ordenação da relevância, o que acarretaria problemas bem mais complicados. Os limiares simplesmente não garantem bons níveis de precisão e abrangência, e por vezes necessitam de ser ajustados ao tamanho do *corpus*. Alguns dos erros associados ao Método das Ilhas estão associados a palavras relevantes que fazem parte de n-gramas e de palavras irrelevantes que se destacam apenas ligeiramente de outras palavras irrelevantes. Estes resultados sugerem a análise a uma vizinhança mais alargada, contemplando-se, se bem que com cada vez menor influência, os segundos (e terceiros, etc.) vizinhos.

Por fim, foi proposta também uma alternativa às actuais abordagens de avaliação da relevância de unigramas: com base na análise silábica, foi desenvolvido o Método das Sílabas. Como foi demonstrado na secção 3.4 e validado pelos seus resultados na secção 4.5, este método, apesar de poder funcionar como uma métrica isolada para a avaliação da relevância de unigramas, quando conjugado com qualquer dos métodos aqui propostos (inclusive os de outros autores) permite melhorar substancialmente a qualidade

dos resultados destes.

Os resultados obtidos neste trabalho são bastante animadores, contudo sugerem algumas melhorias e convidam à investigação na sequência do trabalho aqui apresentado. Em termos futuros, há a considerar o interesse em aumentar a qualidade das métricas de criação de listas de relevância de unigramas; dever-se-á também considerar a melhoria da precisão e abrangência do Método das Ilhas considerando o já mencionado alargamento da vizinhança; também existe a necessidade de associar palavras no singular e plural (por exemplo "comunidades" e "comunidade") e sinónimos (utilizando *thesaurus* ou ontologias), e, sobretudo, realizar mais investigação na área das sílabas, área bastante promissora.

Apêndice A

Considerações sobre o Protótipo

Neste apêndice pretende-se descrever, de grosso modo, o funcionamento do protótipo criado para testar as várias abordagens analisadas nesta dissertação. A linguagem de programação utilizada foi a *MatLab*. Sendo de utilização bastante intuitiva, esta é uma linguagem interactiva direccionada essencialmente para o cálculo numérico, que integra análise numérica, cálculo matricial, processamento de sinais, e construção de gráficos num ambiente interactivo.

O protótipo criado não é, em si, uma aplicação do tipo «produto comercial» mas antes um conjunto de *scripts* ou sequências de instruções executadas na linha de comandos do *Matlab*. Portanto, este apêndice, em vez de apresentar a aplicação, descreve brevemente o funcionamento geral dos *scripts* mais importantes, de acordo com a sua sequência de aplicação com vista à obtenção dum resultado final.

A sequência de passos é, basicamente, a seguinte:

- Criação dos *corpora*,
- Tratamento de espaços e pontuações,
- Contagem dos elementos,
- Cálculo das métricas,
- Implementação do Método das Ilhas,
- Apresentação de resultados.

A.1 Criação dos *corpora*

Os *corpora* utilizados neste trabalho são compostos por diversos documentos obtidos a partir do portal para o Direito da União Europeia (http://europa.eu.int/eur-lex/),

onde é possível encontrar-se um repositório de diversos documentos e comunicações de interesse público no domínio. Foram utilizados textos em Português para a criação do *corpus* Português e textos em Inglês para a criação do *corpus* Inglês. Os documentos podem ser transferidos em formato PDF (Portable Document Format) ou em DOC (Formato do Microsoft Word). Foi então necessário fazer a conversão destes formatos para texto corrido (TXT). Para esta tarefa utilizou-se o programa ConvertDoc (http://www.softinterface.com/Convert-Doc/Convert-Doc.htm), que para além de fazer conversões desses tipos de ficheiros, permite automatizar esse processo de uma forma bastante eficiente.

A.2 Tratamento de espaços e pontuações

A separação do texto em palavras, ou mais correctamente, em unigramas, é feita com algum tratamento adicional, para além de uma simples separação em palavras. Este processo é executado sobre os *corpora* em formato de texto corrido, já no ambiente de trabalho do *MatLab* de forma totalmente automatizada e no início do processo de extracção dos unigramas relevantes.

Verificou-se que era benéfico retirar os sinais de pontuação ao texto original. Manter estes sinais não trazia, aparentemente, nada de benéfico ao processo de extracção de unigramas. A maior parte dos sinais de pontuação era interpretado como fazendo parte de palavras, uma vez que são normalmente colocados junto à palavra que os antecede. Outros eram considerados como palavras separadas dentro do texto.

Portanto, numa primeira fase, o corpus é percorrido de forma a serem removidos os sinais de pontuação, tais como: ( ) , ; : . ! ? “. Todos estes caracteres são substituídos por espaços que são posteriormente ignorados quando as palavras são separadas pelos caracteres de espaços, tabulações e mudanças de linha aquando do processamento do texto.

Após a separação em unigramas é executado um último passo no tratamento dos caracteres. Este passo consiste na conversão de todos os caracteres dos unigramas para maiúsculas, pois esse é um aspecto irrelevante, e pretende-se que unigramas iguais, mesmo que escritos de forma diferente em termos de maiúsculas ou minúsculas, sejam considerados a mesma palavra. Por exemplo, os unigramas “palavra”, “PALAVRA” e “Palavra” são identificados como sendo o mesmo unigrama representado por “PALAVRA”.

A.3 Contagem dos elementos

A contagem dos elementos constitui a fase da extracção de unigramas mais demorada, demorando cerca de 4 minutos para um *corpus* de cerca de 500 000 palavras. O texto é percorrido sendo as diversas palavras do texto indexadas numa tabela de *hash*, mecanismo este que garante uma boa eficiência no processo. Uma tabela de *hash* é uma estrutura de dados que associa chaves a valores. A função primária de uma tabela de *hash* é a procura: dando-se uma chave, que neste caso concreto é a soma dos caracteres ASCII da palavra módulo 569, devolve a posição na tabela onde essa palavra se encontra juntamente com a lista de vizinhos e ocorrências.

A cada palavra percorrida, verifica-se a sua existência na tabela de *hash*, actualizando-se a sua frequência de ocorrência no texto e as listas de sucessores e antecessores dessa palavra incluindo a frequência com que estes ocorrem também.

No caso do método *Tf-idf*, a contagem dos elementos é feita de outro modo, pois este método trabalha não com um macro-documento, mas sim com conjuntos de documentos. Neste caso, os *corpora* utilizados possuem *tags* (ou marcações) com a descrição do inicio de cada documento, o que permite, de certa forma, simular a existência de vários documentos dentro de um único — na verdade os *corpora* utilizados são constituídos por vários documentos individuais e as marcações surgem no início de cada documento individual —. A contagem para o *Tf-idf* é feita anotando, numa tabela de *hash*, para cada palavra w do *corpus* a sua lista de ocorrências nos vários documentos, da forma $[n_1, n_2, \ldots n_m]$, onde n_i representa a frequência de ocorrência da palavra w no documento i, e m o total de documentos de que os *corpora* são constituídos — o número de documentos do *corpus* Português é diferente do Inglês —.

A contagem para o método de Zhou et al. também difere dos anteriores, pois é baseado nas distâncias entre as várias ocorrências de cada palavra. Dessa forma, a contagem é feita recorrendo-se novamente a uma tabela de *hash* em que para cada palavra w é guardada uma lista da forma $[t_1, t_2, \ldots t_m]$, onde t_i representa a posição da i-ésima ocorrência da palavra w, e t_m a posição da última ocorrência da palavra w nos *corpora*.

O uso da tabela de *hash* deve-se pois ao facto deste modelo de tabela ter uma eficiência enorme na função de procura. A utilização da função de *hash*, soma dos caracteres ASCII da palavra módulo 569 garante uma boa distribuição das palavras pela tabela, aumentando significativamente a *performance* geral da procura e inserção.

A.4 Cálculo das métricas

No caso das medidas *Score* e *SPQ*, os valores são calculados para cada entrada na tabela de *hash* que contém todos os dados necessários para o efeito: a frequência do unigrama e as respectivas listas de sucessores e antecessores. É feito o cálculo, de acordo com as expressões 3.4 (no caso da medida *Score*) e 3.5 (no caso da medida *SPQ*), enquanto os valores são transportados para uma tabela final ordenada pelos valores, que é basicamente uma lista de ordenação de relevância. Devido à estrutura da tabela de *hash* na fase de contagem dos unigramas, o cálculo é feito de forma rápida, tendo facilmente acesso a todos os valores necessários para o fazer.

No caso do Método das Sílabas (isolado), o cálculo é feito baseado apenas no número de sílabas de cada palavra, de acordo com a sua importância dada pela figura 3.3 (obviamente em código MatLab). A lista resultante é depois ordenada pela pontuação.

No caso do *Tf-idf*, é construída uma lista ordenada de relevância de unigramas baseada no valor máximo de *tf-idf* para cada palavra distinta dos *corpora*. O cálculo do valor de *tf-idf* está explicado na secção 2.1.2.

De acordo com o método de *Zhou et al.*, o cálculo deste é feito baseado na separação das ocorrências de cada palavra nos *corpora*. Como na fase de contagem é criada uma lista com a posição de cada ocorrência da cada palavra distinta, o cálculo de pontuação é feito de forma rápida. Gera-se a partir daqui uma lista ordenada que serve de lista de ordenação de relevância de unigramas para este método. Mais detalhes acerca do método de contagem desta abordagem podem ser consultados na secção 2.1.3.

A.5 Implementação do Método das Ilhas

O Método das Ilhas, descrito na secção 3.5, utiliza uma qualquer lista de ordenação de relevâncias como as obtidas no passo anterior do protótipo. Utilizando a tabela de *hash* criada para a contagem de vizinhos, tem-se todos os elementos necessários para aplicar o critério das *Ilhas* (ver secção 3.5 para maiores detalhes). Obtém-se, após este passo, a lista ou as listas com indicação das palavras que o método considerou relevantes.

A.6 Apresentação de resultados

Finalmente, após a contagem de unigramas e respectivos cálculos, tem-se, nesta fase, vários tipos de listas. Listas de relevância de unigramas dadas pelas métricas em duas línguas (Inglês e Português), e listas de palavras consideradas relevantes pelo Método das Ilhas separadas por língua e pela métrica que lhes deu origem. Para realizar os vários

testes de qualidade (ver secção 4.2.1), procede-se à avaliação da capacidade das métricas em ordenar correctamente os unigramas. Para tal utiliza-se o método já mencionado na secção 4.2.2, que consiste basicamente em contar quantas palavras relevantes (classificadas manualmente) se encontram nas posições superiores da lista de ordenação. Para realizar os cálculos de precisão e abrangência, de modo a avaliar o Método das Ilhas, procede-se como foi explicado na secção 4.2.3, o que consiste essencialmente em contar quantas palavras classificadas manualmente como relevantes foram consideradas relevantes pelo Método das Ilhas. Por fim, a todas as listas de relevância aplica-se o Método das Sílabas, que, como foi explicado, consiste na multiplicação do valor que uma palavra tem numa lista de relevância pela sua importância baseada no número de sílabas, voltando-se a repetir os passos para avaliar todos os resultados.

Apêndice B

Ilustração do Critério de Relevância

Através de duas listas aleatórias de 200 palavras cada, pretende-se ilustrar o critério de relevância definido pela Professora Maria Francisca Xavier, critério utilizado para os testes e resultados obtidos nesta dissertação. Cada lista corresponde a uma das línguas. Nelas, à esquerda encontram-se as palavras e à direita a sua classificação com o seguinte significado: "X" indica que a palavra em questão é considerada relevante; "?" significa que a palavra pode ou não ser relevante, dependendo do contexto.

B.1 Exemplos de aplicação do critério de relevância — Caso do Português

Tabela B.1: Lista aleatória de 200 palavras — Caso Português

Palavra	Class.	Palavra	Class.	Palavra	Class.
'0'		'outro'		'z'	
'das'		'qual'		'rural'	x
'na'		'relativas'		'utilizadores'	x
'2'		'durante'		'respeita'	x
'à'		'orçamento'	x	'23'	
'4'		'autoridades'	x	'meses'	x
'mais'		'competências'	x	'função'	x
'7'		'contra'		'últimos'	
'países'	x	'l'		'operações'	x
'aos'		'requisito'	x	'criar'	x

'the'		'1998'	x	'certos'	
'empresas'	x	'valor'	x	'existentes'	x
'outros'		'grande'		'máximo'	x
'foi'		'5%'	?	'existentes'	x
'of'		'orçamental'	x	'alimentares'	x
'nível'		'promover'	x	'seguimento'	x
'sua'		'fed'	x	'comerciais'	x
'ue'	x	'métodos'	x	'previsto'	x
'nas'		'coordenação'	x	'construção'	x
'também'		'anexo'	x	'civil'	x
'todos'		'específicos'	x	'inocuidade'	x
'avaliação'	x	'24'		'era'	
'relatório'	x	'conformidade'	x	'regulamentação'	x
'regulamento'	x	'embora'		'cidadãos'	x
'política'	x	'curso'	x	'documento'	x
'd'		'19'		'novembro'	x
'está'		'22'		'primavera'	x
'ainda'		'*'		'profissional'	x
'resultados'	x	'estatísticas'	x	'with'	
'parte'		'concorrência'	x	'terceiros'	x
'legislação'	x	'análise'	x	'n°'	
'apoio'	x	'taxas'	x	'500'	
'relativamente'		'assistência'	x	'2000/01'	x
'novas'		'julho'	?	'carácter'	x
'maior'		'25'		'pouco'	
'investigação'	x	'avaliações'	x	'27'	
'segurança'	x	'livro'	x	'phare'	x
'12'		'humanos'	x	'económicos'	x
'saúde'	x	'peixes'	x	'irl'	x
'geral'		'funcionamento'	x	'despesa'	x
'progressos'	x	'elaboração'	x	'metade'	
'artigo'	x	'eurostat'	x	'comunidades'	x
'm'		'21'		'desafios'	x
'11'		'resposta'	x	'utilizar'	x
'consumidores'	x	'privado'	x	'fonte'	x
'conta'	x	'riscos'	x	'concessão'	x
'lei'	x	'cães'	x	'região'	x

'novos'		'luxemburgo'	x	'resolução'	x
'principais'		'origem'	x	'experiência'	x
'especial'		'espacial'	x	'nl'	x
'será'		'coluna'	x	'decision'	x
'candidatos'	x	'há'		'televisão'	x
'três'		'decisões'	x	'intercâmbio'	x
'estes'		'planos'	x	'information'	x
'pequenas'		'comparação'	x	'discriminação'	x
'convenção'	x	'nota'	x	'desemprego'	x
'partes'		'desenvolver'	x	'momento'	
'diferentes'		'grandes'		'investigadores'	x
'questões'	?	'antropóides'	x	'orientação'	x
'melhorar'	?	'mamíferos'	x	'produtividade'	x
'estado-membro'	x	'rendimento'	x	'orientação'	x
'pagamentos'	x	'be'		'possíveis'	
'anual'	?	'prossímios'	x	'9º'	
'nova'		'diálogo'	x	'interesses'	x
'sem'		'dimensão'	x	'pensões'	x
'frança'	x				

B.2 Exemplos de aplicação do critério de relevância — Caso Inglês

Tabela B.2: Lista aleatória de 200 palavras - Caso Inglês

Palavra	Class.	Palavra	Class.	Palavra	Class.
'2'		'does'		'prosimians'	x
'that'		'situation'	x	'finance'	x
'other'		'budget'	x	'furthermore'	
'their'		'particularly'		'brussels'	x
'services'	x	'potential'	x	'types'	?
'these'		'obligations'	x	'index'	x
'its'		'gdp'	x	'obligation'	x

'used'		'lisbon'	x	'dentistry'	x
'more'		'30'		'elements'	?
'council'	x	'now'		'agricultural'	x
'between'		'toxicity'	x	'regions'	x
'public'	x	'whether'		'rule'	x
'measures'	x	'various'		'u'	
'some'		'regards'	?	'coming'	
'data'	?	'form'		'statistics'	x
'law'	x	'n/a'		'involved'	
'were'		'procedures'	x	'decisions'	x
'-'		'considered'		'transfer'	x
'out'		'proposals'	x	'users'	x
'rules'	x	'financing'	x	'give'	
'specific'	x	'down'		'scientific'	x
'would'		'providing'	x	'facilitate'	x
'made'		'participation'	x	'learning'	x
'internal'	x	'average'	x	'achieve'	x
'year'	x	'external'	x	'infrastructure'	x
'one'		'fund'	x	'strong'	
'service'	x	'structural'	x	'27'	
'most'		'monitoring'	x	'standard'	x
'those'		'help'	x	'justice'	x
'taken'		'agriculture'	x	'once'	
'requirements'	x	'smes'	x	'organisation'	x
'different'	?	'taking'		'type'	?
'important'	?	'co-ordination'	x	'irl'	x
'both'		'implementing'	x	'agenda'	x
'un'	x	'oil'	x	'shows'	
'issues'	x	'analysis'	x	'cover'	x
'f'		'disposal'	x	'czech'	x
'procedure'	x	'review'	x	'generally'	
'candidate'	x	'prices'	x	'contracts'	x
'uk'	x	'recent'		'household'	x
'directives'	x	'regarding'		'revision'	x
'g'		'line'	?	'freedom'	x
'access'	x	'project'	x	'substantial'	x
'climate'	x	'devices'	x	'encourage'	x

'11'		'veterinary'	x	'cannot'	
'per'		'who'		'gas'	x
'actions'	x	'scope'	x	'transparency'	x
'instruments'	x	'rabbits'	x	'similar'	
'risk'	x	'conclusions'	x	'sample'	x
'make'		'1997'	x	'similar'	
'needs'	x	'fish'	x	'prevention'	x
'germany'	x	'versus'		'nl'	x
'belgium'	x	'rats'	x	'42'	
'20'		'open'		'staff'	x
'same'		'21'		'lead'	x
'sectors'	x	'cats'	x	'networks'	x
'france'	x	'guidelines'	x	'aims'	x
'ireland'	x	'launched'	x	'lack'	x
'society'	x	'statistical'	x	'processing'	x
'denmark'	x	'pigs'	x	'36'	
'transport'	x	'network'	x	'efficient'	x
'businesses'	x	'continued'		'norway'	x
'greece'	x	'changes'	x	'concern'	
'effective'	x	'accordance'		'communities'	x
'consumer'	x	'initiative'	x	'solution'	x
'view'	x	'cattle'	x		

Apêndice C

Listas de Resultados da Classificação pelo Método das Ilhas

Este apêndice mostra os resultados obtidos pela aplicação do Método das Ilhas às duas listas aleatórias de palavras apresentadas no apêndice B. Os resultados estão divididos de acordo com a língua; à esquerda encontram-se as palavras e à direita a sua classificação, onde "X" significa que a palavra em questão foi considerada relevante pelo critério de referência já referido e "?" que a palavra em questão pode ou não ser relevante, dependendo do contexto. Por fim, encontra-se a classificação dada pelo Método das Ilhas de acordo com cada uma das métricas utilizadas para gerar a lista de relevância utilizada pelo Método das Ilhas: o "1" marca que a palavra foi classificada como relevante tendo como base a lista de relevância dada pela métrica a que essa coluna pertence.

C.1 Lista de resultados da classificação - Caso Português

Tabela C.1: Lista de resultados da classificação - Caso Português

Palavra	Classificação	Sílabas	*Sc*	*Sc* & Sil.	*SPQ*	*SPQ* & Sil.	*Tf-idf*	*Tf-idf* & Sil.	Zhou	Zhou & Sil.
'0'			1				1		1	
' '										
'das'										

'na'										
'2'									1	
'à'										
'4'									1	
'mais'									1	
'7'									1	
'países'	x	1							1	1
'aos'										
'the'					1		1		1	
'empresas'	x	1	1	1	1	1	1	1	1	1
'outros'										
'foi'										
'of'							1		1	
'nível'			1		1		1		1	
'sua'					1					
'ue'	x		1		1		1		1	
'nas'										
'também'									1	
'todos'			1						1	
'avaliação'	x	1	1	1	1	1	1	1	1	1
'relatório'	x	1	1	1	1	1		1	1	1
'regulamento'	x	1	1	1		1	1	1	1	1
'política'	x	1	1	1	1	1			1	1
'd'							1		1	
'está'										
'ainda'									1	
'resultados'	x	1	1	1	1	1	1	1	1	1
'parte'			1		1				1	
'legislação'	x	1	1	1	1	1	1	1	1	1
'apoio'	x	1	1	1	1	1	1	1	1	1
'relativamente'									1	
'novas'										
'maior'					1				1	
'investigação'	x	1	1	1	1	1	1	1	1	1
'segurança'	x	1	1	1	1	1	1	1	1	1
'12'										
'saúde'	x	1	1	1	1	1	1	1	1	1

'geral'										
'progressos'	x	1	1	1	1	1	1	1	1	1
'artigo'	x	1		1		1	1	1	1	1
'm'					1				1	
'11'										
'consumidores'	x	1	1	1	1	1	1	1	1	1
'conta'	x	1	1	1	1	1			1	1
'lei'	x		1		1		1		1	
'novos'										
'principais'					1	1				
'especial'		1	1	1	1	1				1
'será'										
'candidatos'	x	1	1	1	1	1	1	1	1	1
'três'										
'estes'										
'pequenas'		1	1	1			1	1	1	1
'convenção'	x	1	1	1	1	1	1	1	1	1
'partes'			1						1	
'diferentes'		1		1	1	1		1		1
'questões'	?				1				1	
'melhorar'	?	1	1	1		1	1	1	1	1
'estado-membro'	x	1	1	1	1	1	1	1	1	1
'pagamentos'	x	1	1	1	1	1		1	1	1
'anual'	?						1		1	
'nova'					1					
'sem'										
'frança'	x		1		1		1		1	
'outro'			1						1	
'qual'			1		1		1		1	
'relativas'		1	1	1						1
'durante'		1								
'orçamento'	x	1	1	1	1	1	1	1	1	1
'autoridades'	x	1	1	1	1	1			1	1
'competências'	x	1	1	1	1	1	1	1	1	1
'contra'									1	
'l'									1	
'requisito'	x	1	1	1	1	1	1	1	1	1

'1998'	x		1		1				1	
'valor'	x								1	
'grande'					1					
'5%'	?				1				1	
'orçamental'	x	1			1	1	1	1	1	1
'promover'	x	1	1	1		1	1	1	1	1
'fed'	x				1		1		1	
'métodos'	x	1	1	1		1	1	1	1	1
'coordenação'	x	1	1	1	1	1	1	1	1	1
'anexo'	x	1		1				1	1	1
'específicos'	x	1		1		1		1		1
'24'										
'conformidade'	x	1	1	1	1	1	1	1	1	1
'embora'		1								1
'curso'	x		1		1	1	1		1	
'19'										
'22'										
'*'							1		1	
'estatísticas'	x	1	1	1			1	1	1	1
'concorrência'	x	1	1	1	1	1	1	1	1	1
'análise'	x	1	1	1	1	1	1	1	1	1
'taxas'	x		1	1			1	1	1	1
'assistência'	x	1	1	1		1	1	1	1	1
'julho'	?	1	1	1	1	1	1	1	1	1
'25'										
'avaliações'	x	1	1	1		1		1	1	1
'livro'	x	1	1	1			1	1	1	1
'humanos'	x	1	1	1	1	1	1	1	1	1
'peixes'	x	1	1	1	1	1		1	1	1
'funcionamento'	x	1	1	1	1	1	1	1	1	1
'elaboração'	x	1	1	1	1	1	1	1	1	1
'eurostat'	x	1	1	1	1	1	1	1	1	1
'21'										
'resposta'	x	1	1	1	1	1	1		1	1
'privado'	x	1	1	1	1	1	1	1	1	1
'riscos'	x		1		1		1		1	
'cães'	x		1		1				1	

'luxemburgo'	x	1	1	1	1	1	1	1	1	1
'origem'	x	1	1	1	1	1	1	1	1	1
'espacial'	x	1					1	1	1	1
'coluna'	x	1	1	1		1	1	1	1	1
'há'										
'decisões'	x	1	1	1	1	1	1	1	1	1
'planos'	x		1	1			1		1	1
'comparação'	x	1	1	1	1	1	1	1	1	1
'nota'	x								1	
'desenvolver'	x	1	1	1	1	1	1	1	1	1
'grandes'									1	
'antropóides'	x	1	1	1	1	1		1	1	1
'mamíferos'	x	1	1	1	1	1		1	1	1
'rendimento'	x	1	1	1	1	1	1	1	1	1
'be'					1				1	
'prossímios'	x	1	1	1	1	1		1	1	1
'diálogo'	x	1	1	1	1	1	1	1	1	1
'dimensão'	x	1	1	1	1	1			1	1
'z'			1				1		1	
'rural'	x		1		1		1		1	
'utilizadores'	x		1		1	1	1	1	1	1
'respeita'	x	1	1	1	1	1	1	1	1	1
'23'										
'meses'	x		1		1		1		1	
'função'	x	1	1	1	1	1	1		1	1
'últimos'		1	1	1	1	1		1	1	1
'operações'	x	1	1	1	1	1		1	1	1
'criar'	x		1				1		1	
'certos'									1	
'existentes'	x	1		1				1		1
'máximo'	x	1	1	1	1	1		1	1	1
'alimentares'	x	1	1		1	1	1	1	1	
'seguimento'	x	1	1	1	1	1	1	1	1	1
'comerciais'	x	1		1	1	1		1		1
'previsto'	x	1		1						1
'construção'	x	1	1	1	1	1	1	1	1	1
'civil'	x								1	

'inocuidade'	x	1	1	1	1	1		1	1	1
'era'										
'regulamentação'	x		1		1	1	1		1	
'cidadãos'	x	1	1	1	1	1	1	1	1	1
'documento'	x	1	1	1	1	1	1	1	1	1
'novembro'	x	1	1	1	1	1	1	1	1	1
'primavera'	x	1	1	1	1	1	1	1	1	1
'profissional'	x	1	1	1		1			1	1
'with'									1	
'terceiros'	x	1	1	1	1	1	1	1	1	1
'n°'					1		1		1	
'500'										
'2000/01'	x		1		1		1		1	
'carácter'	x		1		1	1			1	
'pouco'									1	
'27'										
'phare'	x		1		1		1	1	1	
'económicos'	x	1		1		1				1
'irl'	x						1		1	
'despesa'	x	1	1	1			1	1	1	1
'metade'		1	1	1			1	1	1	1
'comunidades'	x		1	1						
'desafios'	x	1	1	1	1	1	1	1	1	1
'utilizar'	x	1	1	1	1	1		1	1	1
'fonte'	x		1						1	
'concessão'	x	1	1	1	1	1	1	1	1	1
'região'	x	1	1	1	1	1	1	1	1	1
'resolução'	x	1	1	1		1	1	1	1	1
'experiência'	x	1	1	1	1	1	1	1	1	1
'nl'	x		1		1				1	
'decision'	x	1	1	1	1	1		1	1	1
'televisão'	x	1	1	1		1	1	1	1	1
'intercâmbio'	x	1	1	1	1	1	1	1	1	1
'information'	x	1		1		1		1	1	1
'discriminação'	x	1	1	1	1	1	1	1	1	1
'desemprego'	x	1	1	1	1	1	1	1	1	1
'momento'		1	1	1	1	1	1	1	1	1

'investigadores'	x	1	1		1	1	1	1	1	1
'orientação'	x	1	1	1	1	1	1	1	1	1
'produtividade'	x	1	1	1	1	1	1	1	1	1
'possíveis'		1	1	1	1	1		1	1	1
'9º'			1						1	
'interesses'	x	1	1	1	1	1	1	1	1	1
'pensões'	x		1		1	1	1		1	

C.2 Lista de resultados da classificação - Caso Inglês

Tabela C.2: Lista de resultados da classificação - Caso Inglês

Palavra	Classificação	Sílabas	Sc	Sc & Sil.	SPQ	SPQ & Sil.	*Tf-idf*	*Tf-idf* & Sil.	Zhou	Zhou & Sil.
'2'									1	
'that'										
'other'		1								
'their'					1					
'services'	x	1		1	1	1	1	1	1	1
'these'										
'its'					1					
'used'		1	1			1			1	1
'more'		1							1	
'council'	x	1	1	1	1	1			1	1
'between'		1								
'public'	x								1	
'measures'	x	1		1		1	1	1	1	1
'some'		1							1	1
'data'	?	1					1	1	1	1
'law'	x		1		1		1		1	
'were'		1							1	1
'-'									1	

'out'					1					
'rules'	x	1					1	1	1	1
'specific'	x	1		1		1				1
'would'									1	
'made'		1	1	1		1			1	1
'internal'	x	1	1	1			1	1	1	1
'year'	x		1		1		1		1	
'one'		1		1		1			1	1
'service'	x	1		1	1	1	1	1	1	1
'most'										
'those'		1			1	1				
'taken'		1	1	1		1	1		1	1
'requirements'	x	1					1	1	1	1
'different'	?	1		1	1	1				1
'important'	?	1	1	1	1	1				1
'both'									1	
'un'	x		1		1		1	1	1	
'issues'	x	1							1	1
'f'							1		1	
'procedure'	x	1		1	1	1	1	1	1	1
'candidate'	x	1	1	1			1	1	1	1
'uk'	x		1		1		1		1	
'directives'	x	1	1	1	1	1	1	1	1	1
'g'			1		1					
'access'	x	1	1	1			1		1	1
'climate'	x	1	1	1			1	1	1	1
'11'					1					
'per'									1	
'actions'	x	1					1	1	1	1
'instruments'	x	1		1	1	1		1	1	1
'risk'	x		1				1		1	
'make'		1	1	1	1	1			1	1
'needs'	x		1							
'germany'	x	1	1	1		1	1	1	1	1
'belgium'	x	1					1	1	1	1
'20'									1	
'same'		1	1	1	1	1				1

'sectors'	x	1			1	1			1	1
'france'	x	1	1	1		1	1	1	1	1
'ireland'	x	1	1	1	1	1	1	1	1	1
'society'	x	1	1	1	1	1	1		1	1
'denmark'	x	1	1	1			1	1	1	1
'transport'	x	1		1	1	1	1	1	1	1
'businesses'	x	1	1	1	1	1	1	1	1	1
'greece'	x	1	1	1					1	1
'effective'	x	1			1	1			1	1
'consumer'	x	1	1	1		1			1	1
'view'	x		1		1		1		1	
'does'			1							
'situation'	x	1	1	1	1	1			1	1
'budget'	x	1	1	1	1	1	1	1	1	1
'particularly'		1		1						1
'potential'	x	1	1	1		1				1
'obligations'	x	1		1	1	1			1	1
'gdp'	x		1		1		1		1	
'lisbon'	x	1	1	1	1	1	1	1	1	1
'30'									1	
'now'					1				1	
'toxicity'	x	1		1		1	1	1	1	1
'whether'		1							1	1
'various'		1		1	1	1				1
'regards'	?	1	1	1	1	1			1	1
'form'			1						1	
'n/a'			1		1	1	1	1	1	
'procedures'	x	1			1	1	1	1	1	1
'considered'		1	1	1		1	1	1	1	1
'proposals'	x	1	1	1			1	1	1	1
'financing'	x	1	1	1	1	1	1	1	1	1
'down'					1				1	
'providing'	x	1		1		1			1	1
'participation'	x		1	1			1		1	
'average'	x	1	1	1	1	1		1	1	1
'external'	x	1		1		1	1	1	1	1
'fund'	x		1		1		1		1	

'structural'	x	1	1	1		1		1	1	1
'monitoring'	x	1		1		1		1	1	1
'help'	x		1		1				1	
'agriculture'	x	1	1	1	1	1		1	1	1
'smes'	x		1		1		1		1	
'taking'		1	1	1						1
'co-ordination'	x	1	1				1	1	1	1
'implementing'	x	1	1	1	1	1			1	1
'oil'	x		1		1				1	
'analysis'	x	1	1	1		1			1	1
'disposal'	x	1	1	1	1	1		1	1	1
'review'	x	1	1	1	1	1	1	1	1	1
'prices'	x	1			1	1	1	1	1	1
'recent'		1		1	1	1				1
'regarding'		1	1	1					1	1
'line'	?	1	1	1	1	1	1	1	1	1
'project'	x	1		1		1		1	1	1
'devices'	x	1	1	1	1	1	1	1	1	1
'veterinary'	x	1	1	1					1	1
'who'										
'scope'	x	1	1	1			1	1	1	1
'rabbits'	x	1	1	1	1	1		1	1	1
'conclusions'	x	1	1	1		1		1	1	1
'1997'	x		1		1				1	
'fish'	x		1						1	
'versus'			1	1					1	1
'rats'	x				1	1	1		1	
'open'		1	1	1		1	1	1	1	1
'21'										
'cats'	x		1						1	
'guidelines'	x	1	1	1		1			1	1
'launched'	x	1	1	1			1	1	1	1
'statistical'	x	1	1	1				1	1	1
'pigs'	x		1		1				1	
'network'	x	1	1	1			1		1	1
'continued'		1	1	1		1	1	1	1	1
'changes'	x	1	1						1	1

'accordance'		1	1	1	1	1	1	1	1	1
'initiative'	x	1					1	1	1	1
'cattle'	x	1	1	1		1		1	1	1
'prosimians'	x	1	1	1	1	1		1	1	1
'finance'	x	1	1	1	1	1		1	1	1
'furthermore'		1		1						1
'brussels'	x	1	1	1	1	1				1
'types'	?		1							
'index'	x	1			1	1	1	1	1	1
'obligation'	x	1	1	1		1			1	1
'dentistry'	x	1	1	1	1	1	1	1	1	1
'elements'	?	1	1	1		1	1	1	1	1
'agricultural'	x						1		1	
'regions'	x	1	1	1	1	1			1	1
'rule'	x	1			1	1	1	1	1	1
'u'							1		1	
'coming'			1	1						
'statistics'	x	1		1	1	1		1	1	1
'involved'		1	1	1		1		1		1
'decisions'	x	1		1		1			1	1
'transfer'	x	1	1	1	1	1	1	1	1	1
'users'	x	1	1	1	1	1	1	1	1	1
'give'		1	1	1	1	1			1	1
'scientific'	x	1		1		1	1	1	1	1
'facilitate'	x	1	1	1	1	1			1	1
'learning'	x				1	1	1	1	1	
'achieve'	x	1	1	1	1	1		1	1	1
'infrastructure'	x						1		1	
'strong'					1					
'27'										
'standard'	x	1		1	1	1			1	1
'justice'	x	1	1	1	1	1		1	1	1
'once'		1		1					1	1
'organisation'	x	1	1						1	1
'type'	?		1						1	
'irl'	x						1		1	
'agenda'	x	1		1		1			1	1

'shows'			1							
'cover'	x	1	1	1	1	1				1
'czech'	x		1				1		1	
'generally'		1		1	1	1				1
'contracts'	x	1					1	1	1	1
'household'	x	1	1	1		1			1	1
'revision'	x	1	1	1		1	1	1	1	1
'freedom'	x	1	1	1			1	1	1	1
'substantial'	x	1		1	1	1				1
'encourage'	x	1	1	1	1	1				1
'cannot'		1	1	1						
'gas'	x		1		1				1	
'transparency'	x	1	1	1		1		1	1	1
'similar'		1		1		1				1
'sample'	x	1	1				1	1	1	1
'similar'		1		1		1				1
'prevention'	x	1	1	1	1	1			1	1
'nl'	x		1		1		1		1	
'42'									1	
'staff'	x						1		1	
'lead'	x		1						1	
'networks'	x	1					1	1	1	1
'aims'	x		1							
'lack'	x		1						1	
'processing'	x	1	1	1	1	1	1	1	1	1
'36'									1	
'efficient'	x	1	1	1	1	1		1	1	1
'norway'	x	1	1		1		1	1	1	1
'concern'		1	1	1					1	1
'communities'	x		1	1	1	1				
'solution'	x	1	1	1	1	1	1	1	1	1

Bibliografia

[Benoit 04] Benoit Mathieu, Romanic Besancon and Christian Fluhr. *Multilingual document clusters discovery.* RIAO'2004, p. 1–10.

[Dagan & Church 94] I. Dagan and K. Church. Termight: Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, 1994.

[Daille 96] B. Daille. *Study and Implementation of Combined Techniques from Automatic Extraction of Terminology. Cap. 3 of "The Balancing Act": Combining symbolic and statistical approaches to language*, pages 49–66. MIT Press, 1996.

[Das et al. 02] A. Das, M. Marko, A. Probst, M. A. Porter and C. Gershenson. *Neural Net Model for featured word extraction.* CoRR, cs. NE/0206001, 2002.

[Feldman et al. 06] R. Feldman and B. Rosenfeld and M. Fresko. *TEG - A hybrid approach to information extraction. Springer-Verlag, 2006.*

[Gael & Spela 05] Gael Dias and Spela Vintar. Unsupervised Learning of Multiword Units from Part-of-Speech Tagged Corpora: Does quantity mean quality? *Lecture notes in computer science.* Vol. 3808, pp. 669-679. Portuguese Conference on Artificial Intelligence, 2005. Covilhã, PORTUGAL December 2005.

[Gao & Zhao 03] Y. Gao and G. Zhao. *Knowledge-based Information Extraction: A case study of recognizing emails of Nigerian frauds*, 2003.

[Heid 99] Ulrich Heid. *A linguistic bootstrapping approach to the extraction of term candidates from German text.* Terminology 5(2).

[Jone & Paynter 02] S. Jones and G. W. Paynter. Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. *Journal of the American Society for Information Science and Technology*, (8):653–677, 2002.

[Luhn 58] H. P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2: 159-165, 1958

[Ortuno et al. 02] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz and A.M. Somoza. *Europhys. Lett. 57 5 (2002), pp. 759-764.*

[Salton & Buckley 88] G. Salton and C. Buckley. *Term-weighting approaches in automatic text retrieval. In Information Processing & Management, 24(5): 513-523.*

[Silva et al. 99a] J. F. Silva and G. Dias and S. Guilloré and G. P. Lopes. Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Progress in Artificial Intelligence*, volume 1695 of *Lecture Notes in Artificial Intelligence*, pages 113–132. Springer-Verlag, 1999.

[Silva et al. 01] J. F. Silva, J. T. Mexia, C. A. Coelho, and G. P. Lopes. Multilingual document clustering, topic extraction and data transformation. In *Progress in Artificial Intelligence*, volume 2258 of *Lecture Notes in Artificial Intelligence*, pages 74–87. Springer-Verlag, 2001.

[Silva et al. 01a] J. F. Silva, J. T. Mexia, C. A. Coelho, and G. P. Lopes. Document clustering and cluster topic extraction in multilingual corpora. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 513–520, San Jose, California, November 2001.

[Smadja 93] F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, (19):143–177, 1993.

[Taniza 01] Taniza Afrin. *Extraction of Basic Noun Phrases from Natural Language Using Statistical Context-Free Grammar*. Master of Science Thesis In Electrical Engineering, submitted to the faculty of Virginia Polytechnic Institute and State University. 2001.

[Ventura & Silva 07] João Ventura and Joaquim Silva. New Techniques for Relevant Word Ranking and Extraction. In *Progress in Artificial Intelligence*, volume 4874 of *Lecture Notes in Artificial Intelligence*, pages 691–702. Springer-Verlag, 2007.

[Zhou 03] H. Zhou and G. W. Slater. A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, Vol 329, issues 1-2, pages 309-327. Elsevier.