Manuel Soares Lourenço

Extracção de palavras compostas por bootstrapping

Universidade da Beira Interior

Departamento de Informática

Julho 2009

Manuel Soares Lourenço

Extracção de palavras compostas por bootstrapping

Tese submetida à Universidade da Beira Interior para o preenchimento dos requisitos para a concessão do grau de Mestre em Engenharia Informática efectuada sob a supervisão do Doutor Gaël Harry Dias, Professor Auxiliar no Departamento de Informática da Universidade da Beira Interior, Covilhã, Portugal

Universidade da Beira Interior
Departamento de Informática
Julho 2009

Resumo

Nesta dissertação foi proposto um novo método, que altera o funcionamento de um sistema existente para extracção de palavras compostas. Este sistema, o SENTA apresenta limitações para as quais apresentamos uma possível solução, extraíndo assim palavras compostas que não seriam extraídas pelo SENTA original. Propomos assim usar um algoritmo de *bootstrapping* para fazer o sistema SENTA trabalhar de forma recursiva, alterando o corpus a cada iteração.

Abstract

In this master thesis we propose a new method, to change the way, an existing system to extract multiword units works. This system, named SENTA shows some limitations. For which we propose a possible solution plus extracting new multiword units that the original SENTA would not extract. For that propose, we propose to use a bootstrapping algorithm to make the system SENTA work recursively changing the corpus at each iteration.

Agradecimentos

Os primeiros agradecimentos vão para os meus pais por serem eles os principais financiadores e por me permitirem chegar até esta etapa, contando sempre com o apoio de ambos para as minhas decisões pessoais.

Ao Doutor Gaël Harry Dias por me ter apresentado esta proposta de tese e pela orientação prestada, os meus agradecimentos.

Finalmente a todos os meus amigos e colegas de curso, que também contribuíram sempre quando necessitei deles, quer em momentos de trabalho, quer em momentos de lazer, o meu obrigado a todos.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A informação digital tem sofrido um enorme crescimento nas últimas décadas, o que tem promovido que o desenvolvimento de uma área de pesquisa de informação de documentos, a IR[1] seja extremamente necessária e que novas abordagens e novos algoritmos tenham ajudado a melhorar a forma como se classificam, filtram, traduzem e se resumem documentos, tornando esta área cada vez mais complexa, mas também mais eficiente e mais automatizada. Ferramentas de extracção automática de informação de documentos, como é exemplo a extracção automática de palavras compostas surgem para responder a estas necessidades.

Nesta dissertação abordar-se-á um novo método de extracção de palavras compostas que combina um algoritmo já existente para este fim, actualmente conhecido como SENTA[2] [4] e um método de programação, o *bootstrapping*, que consiste essencialmente num método recursivo que vai alterando o *corpus*[3] a cada iteração. Um dos principais problemas do SENTA é uma restrição imposta pelo algoritmo *GenLocalMaxs* que impossibilita a extracção de uma palavra composta de tamanho *n* se extrair uma correspondente (que contenha as mesmas palavras) de tamanho *n-1* ou *n+1*. A introdução do *bootstrapping* procura resolver este problema do SENTA para obtenção de palavras compostas que não seriam extraídas utilizando o SENTA normalmente.

O capítulo 2, Extracção de palavras compostas, explica em que contexto o nosso problema se enquadra na área da IR, analisando a evolução da extracção de palavras compostas

[1] do inglês *Information Retrieval*.

[2] Software for the Extraction of N-ary Textual Associations.

[3] Conjunto de textos utilizados para análise, retirados de um conjunto disponibilizado pela Reuters em 2000.

e enunciando trabalhos relacionados.

O capítulo 3, SENTA, explica o funcionamento do software SENTA desenvolvido pelo Alexandre Gil como complemento da sua dissertação: "Extracção eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas" [7], que foi baseada no trabalho teórico de Gaël Dias, Sylvie Guilloré e José Gabriel Lopes em "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora" [4].

O capítulo 4, SENTA por *bootstrapping*, explica o funcionamento do método proposto por esta dissertação para combinar o SENTA com o *bootstrapping*.

O capítulo 5, Resultados, mostra os resultados obtidos e as comparações entre os dois métodos testados e nas várias configurações que permitem.

Finalmente, o último capítulo, mostra as conclusões às quais chegámos com esta nova abordagem para a extracção de palavras compostas.

Capítulo 2

Extracção de palavras compostas

A expansão da *world wide web* e um uso cada vez mais comum de documentos em formato digital em detrimento de documentos físicos, como livros, revistas ou jornais tem provocado um crescimento exponencial de informação em formato digital, Actualmente existe uma quantidade imensurável de documentos que necessitam de ser analisados, classificados ou filtrados de uma forma automática devido ao facto de se ter tornado impossível tratá-los manualmente. Esta realidade levou ao crescimento de métodos automáticos de selecção, tratamento e classificação de documentos, que cada vez mais necessitam de algoritmos rápidos e eficientes, e ao desenvolvimento de uma área de pesquisa de informação em Documentos, a *IR*, uma área que no passado estava mais centrada na indexação e na procura de documentos úteis numa colecção e que hoje inclui uma pesquisa em modelação, classificação e filtragem de documentos, interfaces com o utilizador, visualização de dados [1]. Os motores de busca, o *Google*, o *Yahoo*, e muito recentemente o *Bing* da *Microsoft*, são os principais responsáveis pela investigação acelerada nesta área, em grande parte devido à concorrência que existe entre eles.

Uma das áreas ligadas a esta evolução do tratamento automático de documentos é a extracção automática de palavras compostas, palavras compostas são grupos de palavras que ocorrem frequentemente juntas e que têm um significado diferente do que se ocorressem separadas. Bons exemplos deste tipo de grupos de são, nomes compostos (*Presidente da República*), verbos compostos (*correr riscos*), locuções adverbiais (*de modo algum*), locuções preposicionais (*a respeito de*) ou locuções conjunctivas (*a fim de que*). Esta área não está muito desenvolvida e são ainda poucos os métodos com resultados que possam ser considerados muito bons. Segundo a comunidade cientifica [5] existem três abordagens para extrair palavras compostas, uma primeira usando técnicas baseadas em métodos

linguísticos, por exemplo etiquetagem morfossintática [9] [5] [1] ou utilização de padrões ou modelos linguísticos[8], uma segunda usando métodos puramente estatísticos onde a extracção das palavras compostas é um processo totalmente independente da língua dos documentos [2] [12], e uma terceira abordagem usando um misto das duas abordagens anteriores, onde se procura encontrar certos padrões textuais [13] [11]. A primeira e a terceira abordagens são dependentes da língua e obrigam à existência de bases de dados, actualizadas, de padrões linguísticos [6] [10]. A segunda abordagem obriga igualmente ao desenvolvimento de medidas e dos respectivos limiares de aceitação para palavras compostas de duas palavras não generalizando para N palavras.

Um bom exemplo para um método com bons resultados e o qual analisaremos nesta dissertação, é o SENTA [4], um método baseado puramente em estatísticas para uma extracção em massa de palavras compostas, que usa uma medida de associação que mede a força entre um conjunto de palavras, a *Expectativa Mútua*, e um algoritmo de selecção, o *GenLocalMaxs*. Este método tem uma boa taxa de eficiência para a extracção de palavras compostas, compostas por duas palavras. Um dos pontos chaves para o estudo realizado nesta dissertação reside nesta eficiência e o objectivo deste trabalho pretende utilizar o SENTA por *bootstrapping* ou por outras palavras, utilizá-lo duma forma recursiva, a partir de um *corpus* que será em cada iteração dessa recursividade, modificado integrando as palavras compostas identificadas na etapa anterior.Exemplificando: utilizando o SENTA para retirar palavras de tamanho 2 retiramos a seguinte palavra composta [A B] do seguinte corpus [A B C A D A B C A B] numa primeira iteração. Depois modificando o corpus para uma segunda iteração teremos o seguinte corpus [A_B C A D A_B C A_B] que será utilizado novamente para o SENTA que detectará a palavra composta [A_B C] como sendo de tamanho 2, que na realidade é de tamanho absoluto 3. O ponto chave para este estudo é a limitação do algoritmo de selecção, o *GenLocalMaxs*, relativa ao facto de este algoritmo detectar máximos locais para uma palavra composta de tamanho *n* descartando a possibilidade de extrair as suas palavras compostas vizinhas, ou seja, as palavras compostas com tamanhos *n-1* e *n+1* que contenham as mesmas palavras do que aquela de tamanho *n*.

[1] Classificação das palavras segundo a sua morfologia sintáctica.

Capítulo 3

SENTA

O SENTA é um software criado para a extracção automática de palavras compostas que usa um sistema baseado em estatísticas e que é independente da língua do *corpus*. A versão do SENTA utilizada nessa dissertação é a versão desenvolvida por Alexandre Gil para a sua dissertação de Mestrado, "Extracção Eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas" [7]. O SENTA consiste num sistema que constrói listas com n-gramas e a frequência em que cada um ocorre no corpus, atribuindo um valor baseado nessa frequência definindo uma medida de associação entre as palavras do n-grama, a Expectativa Mútua, e selecciona desse grupo as palavras compostas utilizando um processo de aquisição, o *GenLocalMaxs* que utiliza os máximos locais.

3.1 N-gramas

Os n-gramas são grupos de palavras que respeitam a ordem e a posição nas quais estas ocorrem no corpus. Estes podem ser contíguos ou não-contíguos. São gerados fazendo uma leitura ao corpus por grupos de N palavras, e gerados de cada um desses grupos, os sub-ngramas possíveis. Por exemplo, (ver a figura 3.1). Os n-gramas serão representados nos próximos capítulos com a forma [$p_{11}w_1$,$p_{12}w_2$,$p_{13}w_3$, ... , $p_{1i}w_i$] onde p_{1i} representa a distância entre a palavra w_i e a palavra w_1 ou simplificado $\vec{w}$. Nesta representação foi escolhido o primeiro elemento como o elemento pivot, mas qualquer elemento poderia ser escolhido para tal.

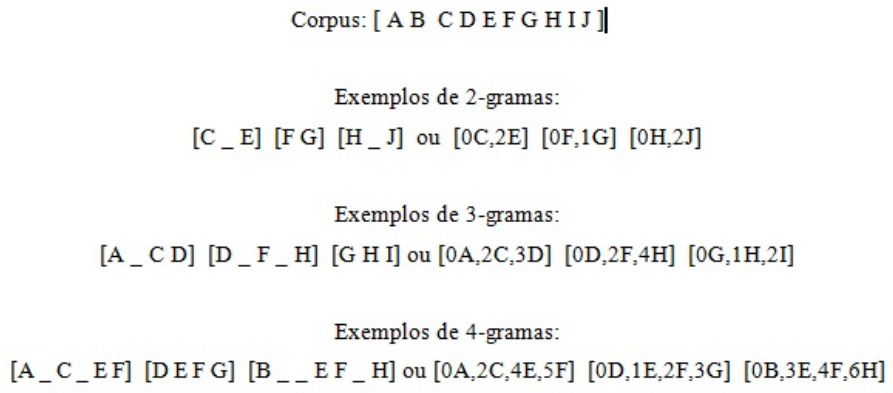

Figura 3.1: Exemplos de construção de n-gramas

3.2 Expectativa Mútua

A Expectativa Mútua é uma medida de associação para n-gramas, que permite identificar a força de ligação entre as palavras de um n-grama, avaliando o impacto da perda de cada uma das palavras no valor do conjunto. Outras medidas de associação utilizadas até então eram insatisfactórias, pois elas só avaliavam o grau de associação entre duas palavras enquanto que a Expectativa Mútua avalia o grau de coesão para N palavras.

Segundo [3], a expectativa mútua é calculada da seguinte forma:

$$ME(\vec{w}) = p(\vec{w}) \times NE(\vec{w})$$

onde $\vec{w}$ é um *n-grama*, p($\vec{w}$) é a probabilidade de ocorrência do *n-grama* $\vec{w}$ no *corpus* em análise e NE($\vec{w}$) é a medida normalizada da expectativa associada a $\vec{w}$. [1]

A NE($\vec{w}$) associada a um *n-grama* é definida como sendo a expectativa média da ocorrência duma palavra numa determinada posição, conhecendo-se a ocorrência das outras palavras desse *n-grama*, igualmente restringidos às suas posições [3], e é calculada da seguinte forma:

$$NE(\vec{w}) = \frac{p(\vec{w})}{FPE(\vec{w})}$$

onde FPE($\vec{w}$) do inglês *Fair Point of Expectation* é a média das probabilidades de ocorrência de cada *sub-ngrama* de $\vec{w}$ no *corpus* e é calculado da seguinte forma:

[1] ME vem do inglês *Mutual Expectation* e NE *Normalized Expectation*.

$$FPE(\vec{w}) = \frac{p(p_{22}[w_2 p_{23} w_3 p_{24} w_4 ... p_{2t} w_t]) + \sum_{i=2}^{t} p([p_{11} w_1 p_{12} w_2 ... \hat{p_{1i}} \hat{w_i} ... p_{1t} w_t])}{t}$$

onde o símbolo " ˆ " representa a palavra a eliminar em cada passo, para a formação de cada *sub-ngrama*. O valor de *t* corresponde ao número de *sub-ngramas* válidos gerados a partir do *n-grama* w.

3.3 Algoritmo GenLocalMaxs

O *GenLocalMaxs* é um algoritmo de selecção independente da linguagem do *corpus*. O *GenLocalMaxs* necessita de uma medida de associação crescente, ou seja, onde os valores mais altos correspondem a casos mais relevantes. Neste caso a medida de associação que é a Expectativa Mútua.

Para que um *n-grama* seja considerado uma palavra composta, este terá de ser um máximo local, ou seja sendo $\vec{w}$ um *n-grama* terá de verificar as seguintes condições [2]:

$$\forall \vec{x} \in \Omega_{n-1} \; \forall \vec{y} \in \Omega_{n+1}$$

$$tamanho(\vec{w}) = 2 \; e \; g(\vec{w}) > g(\vec{y})$$

$$ou$$

$$tamanho(\vec{w}) > 2 \; e \; g(\vec{x}) \leq g(\vec{w}) \; e \; g(\vec{w}) > g(\vec{y})$$

onde *tamanho(.)* representa o número de palavras de cada *n-grama*, *g(.)* a função da medida de associação, $\vec{w}$ o *n-grama*, $\vec{x}$ um *(n-1)-grama*, $\vec{y}$ um *(n+1)-grama*, $\Omega_{\text{n-1}}$ é o conjunto dos *sub-ngramas* de tamanho (*n-1*) e $\Omega_{\text{n+1}}$ o conjunto de *super-ngramas* de tamanho (*n+1*), calculados a partir do ngrama $\vec{w}$.

Capítulo 4

SENTA por Bootstrapping

Explicado o funcionamento do SENTA no capítulo anterior, pode-se passar a este novo capítulo que demonstra o trabalho realizado, abordando um novo sistema que combina o SENTA com um método recursivo, o *bootstrapping*, um método que vai alterando o *corpus* e que vai executando várias vezes o SENTA com esse novo *corpus*, ou seja, o *corpus* a ser utilizado vai sendo refinado e reutilizado em cada nova iteração. Este processo é repetido até o SENTA não extrair novas palavras compostas do *corpus* em análise.

O SENTA utilizado está modificado para assimilar os *underscores* "_" como sendo letras normais, assim considera por exemplo a seguinte palavra composta: "Presidente_da_república" como sendo uma única palavra, e é esta a chave para a construção do novo *corpus*.

O *corpus* utilizado para os testes é um *corpus* público na língua inglesa lançado pela *Reuters* em 2000, que é um conjunto de ficheiros de texto de um apanhado de inúmeras notícias. Nesta dissertação é utilizada uma pequena parcela desse *corpus* contendo aproximadamente 1.2 milhões de palavras.

4.1 Arquitectura do sistema

O sistema como se pode ver na figura é constituído por 4 partes distintas, o SENTA que é a parte mais importante do sistema onde é feita a selecção das palavras compostas, o *MwuExtract* que faz uma selecção das palavras compostas retiradas do resultados do SENTA, o *CorpusBuilder* que cria um novo *corpus*, e uma parte de verificação da existência de palavras compostas, que fará o programa parar quando não encontrar mais expressões ou palavras compostas (ver figura 4.1).

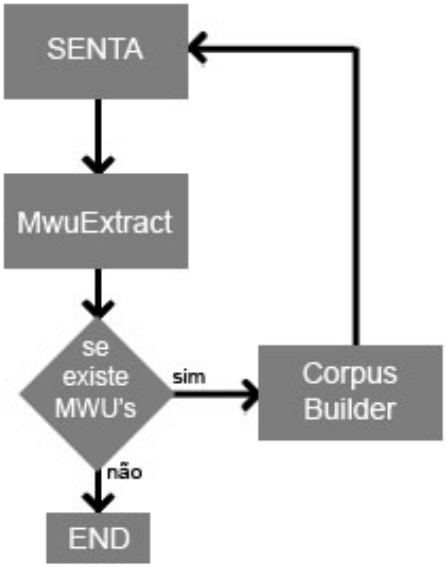

Figura 4.1: Funcionamento do SENTA por *bootstrapping*

4.2 MwuExtractor

O *MwuExtractor* selecciona as palavras compostas que tenham um tamanho igual ou inferior ao pretendido, e ignora todas as palavras compostas não contíguas, pois só se testará este sistema para palavras compostas contíguas.

4.3 CorpusBuilder

O *CorpusBuilder* utiliza a lista criada pelo *MwuExtractor* e analisa a existência de cada uma das palavras compostas seleccionadas em todos os ficheiros do *corpus* substituindo cada palavra composta encontrada por uma equivalente, mas separada por um *underscore* na vez de um espaço, tendo em conta que o SENTA está preparado para ignorar os *underscores*, que os considera como se fossem uma letra, fazendo com que as palavras compostas que foram modificadas no *corpus* sejam interpretadas na próxima iteração do sistema como se fossem uma única palavra.

Capítulo 5

Resultados

Neste capítulo apresentam-se e analisam-se os resultados dos testes realizados com o SENTA por *bootstrapping* e com o SENTA Original. O objectivo é comparar os resultados obtidos em ambos os casos, para saber até que ponto a recursividade do SENTA por *bootstrapping* consegue resolver a limitação do SENTA na sua versão original, uma limitação que advém do algoritmo *GenLocalMaxs* que não permite a extracção de uma palavra composta com *n* palavras se já tiver sido extraída uma palavra compostas com *n-1* ou *n+1* dessas palavras.

Como foi explicado no capítulo anterior o SENTA por *bootstrapping* utiliza duas variáveis que podem ser ajustadas para obter resultados diferentes. Estas duas variáveis que serão enunciadas por *N* e *M* são respectivamente, o tamanho dos n-gramas utilizados pelo SENTA e o tamanho máximo das palavras compostas que no novo corpus construído serão consideradas como uma única palavra. Na tabela 5.1 é demonstrado como são formados os n-gramas, seleccionando uma palavra e as suas vizinhas. Por essa razão é que são utilizados os valores ímpares 3, 5 e 7 para o *N*, porque as palavras vizinhas seleccionadas são sempre na mesma quantidade para a esquerda e direita.

Tabela 5.1: N-gramas utilizados pelo SENTA

	contexto esq.	palavra pivot	contexto dir.
N = 3	- 1	palavra	+ 1
N = 5	- 2	palavra	+ 2
N = 7	- 3	palavra	+ 3

5.1 SENTA por *bootstrapping* Vs SENTA Original

Nesta primeira secção, iremos analizar e comparar os valores de obtidos em quantidade de palavras extraídas, no SENTA Original e no SENTA por emphbootstrapping. No SENTA Original e iremos fazer 3 testes para valores de N, de 3, 5 e 7. No SENTA por emphbootstrapping. iremos fazer mais testes para valores de M/N, de 2/3, 2/5, 3/5, 4/5, 2/7, 3/7, 4/7, 5/7 e 6/7. Verifique-se que o valor de M é o tamanho máximo para as palavras compostas extraídas que serão reutilizadas na construção do novo *corpus*, logo, por exemplo, quando M = 5, reutilizamos as palavras compostas de tamanhos 2, 3, 4 e 5. Os valores de M podem ser no máximo N-1, porque o SENTA Original só extrai palavras compostas com um máximo de N-1.

Nas seguintes tabelas 5.2 e 5.3 estão as quantidades de palavras compostas extraídas. Utilizamos para estes testes o *corpus Reuters* de 1.2 milhões de palavras e os valores entre os parêntises são as quantidades de palavras compostas extraídas a cada iteração. Para valores de M = N-1, as palavras compostas extraídas na primeira iteração pelo SENTA por emphbootstrapping são as mesmas que são extraídas no SENTA Original usando o N.

Tabela 5.2: Resultados do SENTA Normal

N	Quantidade de Palavras compostas extraídas
3	139
5	319
7	327

Na tabela 5.4 temos as palavras que foram encontradas pelo SENTA por *bootstrapping* depois da primeira iteração concluída, palavras que não usaram as palavras encontradas nas iterações anteriores, mas que devido à modificação do corpus passaram a ter medidas/valores suficientes para serem seleccionadas pelo *GenLocalMaxs*.

5.2 Precisão da Extracção

Uma palavra composta extraída só é considerada uma boa palavra composta se as palavras que a compõem tiverem um significado em conjunto diferente do que o que teriam separadas. Nas próximas tabelas 5.5 e 5.6 mostramos as frequências e percentagens de acertos

Tabela 5.3: Quantidades de palvras compostas extraídas pelo SENTA por *bootstrapping* para vários valores de *M*

M	N	Quantidade de Palavras compostas extraídas
2	3	169 (139 + 26 + 4)
2	5	129 (116 + 13)
3	5	365 (258 + 86 + 18 + 3)
4	5	439 (319 + 103 + 16 + 1)
2	7	122 (107 + 15)
3	7	349 (245 + 86 + 15 + 3)
4	7	420 (299 + 108 + 13)
5	7	447 (317 + 119 + 11)
6	7	466 (327 + 126 + 11 + 2)

Tabela 5.4: Quantidade de palavras compostas extraídas pelo SENTA por *bootstrapping* que foram extraídas sem usar as extraídas nas iterações anteriores.

M	N	Quantidade de Palavras compostas extraídas
2	3	24
2	5	8
3	5	29
4	5	27
2	7	9
3	7	31
4	7	29
5	7	27
6	7	28

do SENTA original e do SENTA por *bootstrapping*.

Analizando estes resultados podemos verificamos que para valores em que o M está mais distante do N, a precisão aumenta, mas a quantidade de resultados obtidos é menor. Sendo o teste para M/N igual a 3/5 provavelmente o resultado melhor, pois tem mais uma frequência de precisão superior ao dobro dos resultados dos teste que tenham maior percentagem de precisão.

Tabela 5.5: Precisão de extracção do SENTA Original

N	Frequencia	Percentagem
3	111	79,9 %
5	209	65,5 %
7	211	64,5 %

Tabela 5.6: Precisão de extracção do SENTA por *bootstrapping*

M	N	Frequencia	Percentagem
2	3	132 (111 + 19 + 2)	78,1 %
2	5	108 (97 + 11)	83,7 %
3	5	259 (174 + 69 + 14 + 2)	71,0 %
4	5	287 (209 + 67 + 10 + 1)	65,3 %
2	7	102 (90 + 12)	83,6 %
3	7	215 (158 + 57 + 10 + 2)	61,6 %
4	7	265 (192 + 73 + 8)	63,1 %
5	7	287 (207 + 80 + 6)	64,2 %
6	7	290 (211 + 79 + 6 + 2)	62,2 %

Para as palavras compostas extraídas a partir da segunda iteração que não foram construídas a partir das extraídas nas iterações anteriores, mostramos na tabela 5.7 a frequência e a percentagem de precisão. Para estes casos podemos verificar que os resultados não são assim tão bons a nível de precisão, mas sempre retiramos palavras compostas que o SENTA Original não detecta, o que por si só já é algo significativo.

Tabela 5.7: Precisão de extracção das palavras compostas sem usar as extraídas nas iterações anteriores pelo SENTA por *bootstrapping*

M	N	Frequencia	Percentagem
2	3	15	62,5 %
2	5	6	75,0 %
3	5	22	75,9%
4	5	15	55,6 %
2	7	6	66,7 %
3	7	22	71,0 %
4	7	16	55,2 %
5	7	15	55,6 %
6	7	14	50,0 %

Capítulo 6

Conclusão

Neste trabalho procuramos corrigir uma limitação de um sistema de detecção de palavras compostas, o SENTA, usando um algoritmo de *bootstrapping* para implementar uma forma recursiva de refinar os resultados a cada iteração. Este método para extrair palavras compostas apresentado é um método pesado e que necessita de bastante tempo de processamento comparado com o SENTA original, este poderá ser optimizado visto que não era objectivo desta dissertação implementar um software optimizado e sim funcional que esclarecesse se realmente conseguiria responder aos objectivos de corrigir a falha, que foi explorada, do SENTA original. Depois de analisados os resultados obtidos por este método, podemos concluir que o SENTA por *bootstrapping* consegue resolver a falha detectada no SENTA Original e extrair palavras compostas que não seriam extraídas, embora sejam em pequenas quantidades poderão ser suficientes para fazer do SENTA por *bootstrapping* um sistema mais completo. Mesmo assim, a eficiência do sistema não apresenta resultados satisfatórios, pois também extraímos bastantes palavras compostas que não o podem ser considerado. Os melhores resultados a nível de precisão/quantidade são obtidos na configuração M/N = 3/5. Para trabalho futuro poderemos considerar para além de uma melhoria de optimização do sistema, também a sua implementação para o HELAS, um sistema similar ao SENTA mas que usa etiquetagem morfossintática acopulada ao sistema do SENTA, uso da Expectativa Mútua e do *GenLocalMaxs*, um sistema já dependente da língua do *corpus*.

Bibliografia

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999.

[2] Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of 9th Portuguese Conference in Artificial Intelligence (EPIA 1999)*, pages 113–132, 1999.

[3] Gaël Dias, Sylvie Guilloré, Jean claude Bassano, José Gabriel, and José Gabriel Pereira Lopes. Combining linguistics with statistics for multiword term extraction: A fruitful association? In *Proceedings of 6ème Conférence sur la Recherche d'Informations Assistée par Ordinateur (RIAO 2000)*, 2000.

[4] Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In *Proceedings of 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, 1999.

[5] Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. Multilingual aspects of multiword lexical units. In *Proceedings of Workshop on Language Technologies of the 32th annual meeting fo the Societas Linguistica Europea*. S. Vintar (eds), 1999.

[6] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured texts, 2000.

[7] Alexandre Nuno Capinha Gil. Extracção eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas. Master's thesis, Universidade Nova de Lisboa, 2002.

[8] Blank I. Computer-aided analysis of multilingual patent documentation. In *Proceedings of the First LREC*, pages 765–771, 1998.

[9] Dagan I. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 23–29, 1994.

[10] Satoru Ikehara, Satoshi Shirai, and Hajime Uchino. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *Proceedings of the 16th conference on Computational linguistics*, pages 574–579, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[11] Church K. Word association norms mutual information and lexicography. In *Computational Linguistics*, pages 23–29, 1990.

[12] Feldman R. Text mining at the term level. In *Proceedings of PKDD98*, 1998.

[13] Frank Smadja. Retrieving collocations from text: Xtract. In *Computational Linguistics*, volume 19, pages 143–177, 1993.