Avaliação da Influência da Remoção de *Stopwords* na Abordagem Estatística de Extração Automática de Termos

Ígor Assis Braga

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos, SP – Brasil

igorab@icmc.usp.br

Abstract. *The construction of terminological products is important to the organization and spreading of knowledge. This task can be leveraged by the automatic extraction of terms, which has been considered a Natural Language Processing problem. In this paper, the interaction between the statistical approach to term extraction and the process of stopword removal is investigated. Experiments conducted on two corpora show that stopword removal improves performance when extracting bigram terms, no matter if the removal is done before or after the application of a statistical metric. As a result of this investigation, it is possible to recommend more appropriate statistical metrics for the case where it is possible to remove stopwords and for the case that this removal cannot be done.*

Resumo. *A criação de produtos terminológicos é importante para a difusão e a organização de conhecimentos. Essa tarefa pode ser auxiliada pela extração automática de termos, que vem sendo tratada como um problema de Processamento de Língua Natural. Neste artigo, é investigada a interação entre a abordagem estatística de extração de termos e a remoção de stopwords. Experimentos sobre dois córpus mostram que é benéfico fazer a remoção de stopwords em bigramas, independente se antes ou depois da aplicação de uma medida estatística. Como resultado da investigação, pode-se recomendar as medidas estatísticas mais apropriadas para o caso em que é possível remover stopwords e para o caso em que essa remoção não pode ser feita.*

1. Introdução

O Dicionário Aurélio da Língua Portuguesa define terminologia como "um conjunto de termos próprios duma arte ou duma ciência". Em outras palavras, unidades terminológicas designam idéias, ações ou objetos de um campo especializado da atividade humana. Os pesquisadores que compartilham dessa visão concebem a terminologia como uma forma de saber, de conhecer ou de comunicar [Dias 2000]. Sendo assim, produtos terminológicos[1] constituem ferramentas muito úteis nestes tempos em que "estamos afogados em Informação e sedentos por Conhecimento".

A criação de produtos terminológicos tem sido auxiliada pela extração automática de candidatos a termos de um córpus [Almeida et al. 2006]. Entre as abordagens computacionais para extração automática de termos, encontram-se as estatísticas, as quais usam

[1]Exemplos de produtos terminológicos incluem ontologias, dicionários especializados e tesauros para indexação de documentos.

córpus para calcular medidas que reflitam características de unidades terminológicas. Além de serem utilizadas sozinhas para a extração de candidatos a termos, as medidas estatísticas também são utilizadas em conjunto com abordagens linguísticas para tentar melhorar a qualidade da extração [Teline 2004].

Apesar da abordagem estatística ter a vantagem de independência de língua, o seu uso sozinho acaba gerando muito ruído pela recuperação de expressões que não configuram termos (expressões idiomáticas, expressões verbais fixas etc). Com o intuito de aliviar esse problema, é comum a remoção de *stopwords*[2], que são palavras, geralmente funcionais, que não devem ser consideradas para a formação de um termo. A remoção de stopwords pode acontecer em três momentos distintos: antes ou depois da execução das abordagens estatísticas e no momento da *token*ização dos textos. Por alterar a contagem de frequências, a remoção de stopwords e o momento em que ela é feita podem alterar também o desempenho de uma medida estatística na caracterização de unidades terminológicas.

Neste artigo, será investigado o efeito da remoção de stopwords nos diferentes momentos em que ela pode ser feita. Para isso, na próxima seção, serão apresentadas as principais medidas estatísticas que já foram aplicadas à extração automática de termos. Em seguida, na Seção 3, são descritos os experimentos que foram realizados para analisar o efeito da remoção de stopwords nos diversos momentos em que ela pode ser aplicada. A apresentação e a análise dos resultados seguem na Seção 4. Por fim, são destacados os principais resultados obtidos e são feitas recomendações sobre qual tipo de medida utilizar no caso em que é possível remover stopwords e no caso em que isso não é possível.

2. Medidas Estatísticas para Extração Automática de Termos

O primeiro passo para a extração automática de candidatos a termos é tentar caracterizá-los. Segundo Kageura e Umino [1996], unidades terminológicas possuem duas características importantes: unidade e *termhood*. Unidade se refere à indivisibilidade e à estabilidade de termos multipalavras, isto é, as palavras que compõem um termo estão fortemente ligadas. Termhood, por sua vez, se refere à capacidade de um termo em representar conceitos de um domínio específico.

Neste trabalho, será focado o uso de medidas estatísticas para tentar identificar termos em n-gramas, pois há um número maior de medidas desenvolvidas para esse caso. Para cada n-grama extraído de um determinado córpus, uma medida estatística dá um *score* que indica o quanto que o n-grama pode ser considerado um candidato a unidade terminológica. Geralmente, quanto maior o score dado, maior é a confiança da medida nesse fato. Para exemplificar com algumas das medidas que serão utilizadas neste trabalho, considere a matriz de contigência — Tabela 1 — para um bigrama extraído de um córpus. Nessa matriz, n_{11} indica o número de vezes em que o bigrama `São Paulo` aparece no córpus, n_{12} indica o número de vezes que a palavra "São" aparece como primeira palavra de um bigrama e a palavra "Paulo" não aparece como segunda palavra. O mesmo se aplica aos outros contadores, sendo que as margens são a soma dos contadores internos da matriz. As fórmulas para algumas das medidas apresentadas nesta seção podem ser encontradas na Tabela 2. Enquanto o coeficiente Dice tenta capturar a característica de

[2]O termo stopword será mantido pela falta de uma boa tradução para o português e por ser bastante difundido.

unidade, a frequência de ocorrência (TF) pode ser utilizada para capturar a característica de termhood.

	São	!Paulo	
São	n_{11}	n_{21}	n_{p1}
!Paulo	n_{12}	n_{22}	n_{p2}
	n_{1p}	n_{2p}	n_{pp}

Tabela 1. Exemplo da ocorrência de um bigrama em um córpus

$\frac{2n_{11}}{n_{1p}+n_{p1}}$	$\frac{n_{11}}{n_{pp}}$	$2 \cdot \frac{n_{11}}{n_{pp}} \cdot \frac{n_{11}}{n_{11}+n_{12}+n_{21}}$
Coeficiente Dice	Frequência	Mutual Expectation (bigramas)

Tabela 2. Algumas medidas estatísticas

A medida de *Mutual Information* (MI) foi proposta em [Pantel e Lin 2001] para a extração de termos levando em consideração a característica de unidade. Essa medida faz a suposição de que palavras que compõem n-gramas que não são termos ocorrem independentemente umas das outras. Já a medida de *Log-likelihood* (LL) aparece no mesmo trabalho, e é apontada como uma medida mais robusta que a MI por também levar em consideração a frequência do n-grama sendo analisado.

Em [Dias et al. 2000], é apresentada a medida *Mutual Expectation* (ME) para a extração automática de termos. Essa medida, ao contrário de Mutual Information e Log-likelihood, não faz pressuposições de independência. Um outro ponto favorável é que o score dado a um bigrama pode ser diretamente comparado àquele dado a um trigrama ou outro n-grama qualquer. Analisando a fórmula de ME, pode-se verificar que essa medida é o produto entre a frequência do n-grama e um termo que é diretamente proporcional ao coeficiente Dice para o n-grama.

3. Experimentos

Considerando uma lista de stopwords[3] contendo artigos, preposições, advérbios, pronomes e conjunções, foram extraídos todos os bigramas e trigramas de dois córpus: um córpus de artigos sobre revestimentos cerâmicos e um córpus de textos sobre Ecologia. Algumas medidas estatísticas foram, então, aplicadas separadamente nos bigramas e nos trigramas extraídos. Essas medidas ordenam os n-gramas para que, em seguida, essa ordenação seja comparada com uma lista de unidades terminológicas preparada por especialistas de cada área. Antes de detalhar os córpus e os critérios de avaliação, considere um córpus C e duas listas de referência L_2 e L_3 para esse córpus que sejam, respectivamente, compostas de duas e de três palavras. Esse exemplo será utilizado para descrever a metodologia seguida neste trabalho.

[3] `http://www.icmc.usp.br/~igorab/stopList.txt`

3.1. Metodologia

A primeira etapa da execução de um experimento consiste da *token*ização dos textos no córpus C, que é o processo de dividir o texto em palavras que formarão os bigramas e os trigramas. A tokenização é um dos possíveis pontos de remoção de stopwords. Quando uma stopword é removida nessa etapa, ela não é considerada na formação de um n-grama. Desse modo, se "de" é uma stopword, então a sequência "`casa de tábua`" formará somente o bigrama "`casa tábua`".

Após a formação dos n-gramas, é gerada uma lista com frequências absolutas relacionadas aos n-gramas extraídos do córpus C. Considere, por exemplo, "`São<>Paulo<>50 100 75`". Nesse exemplo, está indicado que, no córpus C, o bigrama "São Paulo" apareceu 50 vezes em todo o córpus, a palavra "São" apareceu 100 vezes na primeira posição de um bigrama e a palavra "Paulo" apareceu 75 vezes como a segunda palavra de um bigrama. No caso de um trigrama, haverá também contadores indicando a frequência de todas as combinações possíveis de duas palavras que o compõem.

O segundo ponto possível de remoção de stopwords é exatamente antes da geração da lista de frequências. Os n-gramas que porventura contiverem stopwords como uma de suas palavras são eliminados. Por exemplo, se "de" é uma stopword, e existem bigramas como "`casa de`" e "`de tábua`", então esses bigramas não aparecem na lista de frequência. Mais ainda, os valores das frequências dos outros n-gramas são atualizados para não considerar as ocorrências dos bigramas removidos.

Eliminando stopwords ou não, a lista de frequência é então processada por cada uma das medidas estatísticas analisadas. A saída são listas em que cada n-grama recebe um score de acordo com a medida utilizada. Os n-gramas nessas listas são, então, ordenados decrescentemente em relação ao score obtido. Logo após a formação das listas ordenadas por score, vem o terceiro e último momento em que é possível remover stopwords. O processo é o mesmo daquele feito para o segundo momento, com exceção de que nenhuma frequência é atualizada. Isso significa que a remoção no terceiro momento preserva a ordenação feita pelas medidas estatísticas, enquanto que fazer a remoção no segundo momento pode influenciar na ordem final.

A partir da lista ordenada gerada pelas medidas estatísticas, é gerada uma última lista para a avaliação das medidas. Cada uma dessas listas contém, em cada linha, o score obtido por um n-grama e um rótulo "0" ou "1". O rótulo "0" indica que o n-grama em questão não está na lista de referência L_n, e o rótulo "1" indica o inverso. A partir dessas listas, é feita, como será visto a seguir, a avaliação das medidas estatísticas.

3.2. Critérios de Avaliação

Vários trabalhos anteriores utilizaram as medidas de precisão (P) e de revocação (R) — Tabela 3 — ou uma combinação das duas utilizando F-*measure* como critérios de avaliação das abordagens estatísticas [Zavaglia et al. 2007, Alegría et al. 2004, Teline 2004, Pantel e Lin 2001]. No entanto, esse tipo de avaliação acaba sendo "pontual", pois envolve definir um score de corte (ou limiar) para classificar n-gramas como termos ou não termos. Assim, somente os n-gramas que tivessem score maior ou igual ao limiar seriam considerados termos. A princípio, não existe um limiar específico com o qual se queira trabalhar, por isso, é interessante fazer uma avaliação independente de limiar.

$P = \frac{UR}{TR}$	$R = \frac{UR}{UT}$
Unidades terminológicas recuperadas (UR)	n-gramas recuperados (TR)
Total de unidades terminológicas na lista de referência (UT)	

Tabela 3. Cálculo de precisão (P) e de revocação (R)

Uma medida bastante apropriada para a avaliação de "ordenadores", e que é independente de limiar, é a área sobre a curva ROC (AUC). Os detalhes de como é feita a construção da curva ROC podem ser encontrados em [Fawcett 2004]. O valor de AUC aproxima-se de 1 quanto melhor for a ordenação, e de 0,5 quanto pior for a ordenação (Tabela 4). No contexto deste trabalho, isso quer dizer que o valor de AUC vai ser maior quanto mais unidades terminológicas tiverem um score maior que os outros n-gramas.

1	25.6288
1	23.1143
1	22.5635
1	22.1866
1	22.0209
0	22.0209
⋮	
0	12.0209
0	11.6117
AUC = 1	

1	25.6288
0	23.1143
1	22.5635
0	22.1866
1	22.0209
0	22.0209
⋮	
1	12.0209
0	11.6117
AUC = 0,5	

Tabela 4. Valores de AUC para a melhor ordenação possível (esq.) e para a pior ordenação possível (dir.)

Para ainda se ter uma ideia de precisão e revocação, será utilizada a medida *Precision-Recall Breakeven Point* (BP). Para calculá-la, deve-se obter pares do tipo (p, r) para todos os limiares possíveis. Esses pontos geram a curva "precisão $\times$ revocação". A medida BP é então definida como o valor em que a precisão se iguala à revocação (p = r). Além da medida BP, será considerado o número de termos que se encontram entre os 100 primeiros n-gramas com maior score (T100). Essa última medida pode ser interessante no caso em que um humano esteja avaliando os n-gramas extraídos.

3.3. Recursos e Ferramentas

Os experimentos foram conduzidos em dois córpus preparados no Núcleo Interinstitucional de Linguística Computacional (NILC). Um deles é o *CórpusEco* [4], que tem 260.921 palavras e é composto de textos em português extraídos de livros que versam sobre Ecologia. A lista de referência para esse córpus conta com 136 unidades terminológicas bigramas e 62 unidades terminológicas trigramas, todas elas ocorrendo nos textos.

O segundo córpus utilizado foi criado a partir de artigos sobre revestimentos cerâmicos[5] da Revista Cerâmica Industrial [Teline 2004]. Após uma etapa de limpeza

[4] `http://www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm`.

[5] `http://www.icmc.usp.br/~igorab/revCer.zip`

nos artigos, o tamanho do córpus ficou em 448.352 palavras. A lista de referência para esse córpus contém termos que já foram avaliados por especialistas da área, e conta com 74 unidades terminológicas bigramas e 43 unidades terminológicas trigramas.

Para a contagem de n-gramas e a aplicação das medidas estatísticas, foi utilizado o *N-gram Statistics Package* (NSP) [Banerjee e Pedersen 2003]. Todas as medidas, com exceção de Mutual Expectation, já estavam disponíveis no programa. A medida de Mutual Expectation para bigramas e para trigramas foi implementada pelo autor deste trabalho dentro do *framework* do NSP. Para obter o valor de BP, de T100 e de AUC, foi utilizado a ferramenta *Perf*[6].

4. Resultados e Análise

As Tabelas 5 e 6 mostram os resultados da extração de unidades terminológicas bigramas nos dois córpus considerados neste estudo. As medidas aplicadas foram Log-likelihood (LL), Mutual Expectation (ME), Mutual Information (MI), frequência de n-gramas (TF) e coeficiente Dice. Os experimentos foram conduzidos sem a remoção de stopwords, com a remoção durante a tokenização (stop-token), com a remoção antes da aplicação da medida estatística (stop-antes) e com a remoção depois da aplicação da medida estatística (stop-depois).

	sem stopwords			stop-token			stop-antes			stop-depois		
	BP	T100	AUC	BP	T100	AUC	BP	T100	AUC	BP	T100	AUC
LL	9,09	9	**89,2**	12,8	**12**	**87,9**	**15,1**	**14**	83,6	**16,3**	**14**	**85,5**
ME	**12,5**	**11**	89,1	10,5	11	84,0	11,6	12	78,7	12,8	12	81,2
MI	9,09	9	77,6	**13,5**	**12**	83,0	**15,1**	**14**	81,7	15,4	**14**	77,2
TF	1,14	1	80,2	10,5	10	86,1	**15,1**	13	**84,3**	15,1	13	84,3
Dice	0	0	87,5	0	0	77,3	0	0	70,5	0	0	74,2

Tabela 5. Extração de unidades terminológicas bigramas no córpus de revestimentos cerâmicos

	sem stopwords			stop-token			stop-antes			stop-depois		
	BP	T100	AUC	BP	T100	AUC	BP	T100	AUC	BP	T100	AUC
LL	3,80	3	83,4	**10,8**	**12**	**75,6**	**12,0**	**16**	70,8	12,0	**16**	72,8
ME	**6,96**	**8**	**83,5**	6,96	10	69,9	10,1	12	65,9	7,60	11	66,7
MI	3,53	3	76,3	10,7	11	72,1	11,9	**16**	69,5	**12,1**	15	70,6
TF	0	0	68,4	8,86	11	74,0	10,4	13	**73,1**	10,4	13	**73,1**
Dice	0,26	0	81,7	0,18	0	62,4	0,18	0	57,4	0,26	0	58,9

Tabela 6. Extração de unidades terminológicas bigramas no *CórpusEco*

Com relação a se fazer ou não a remoção de stopwords, os resultados das Tabelas 5 e 6 mostram que a remoção de stopwords foi muito benéfica para a recuperação de bigramas que sejam terminologia. Em todos os casos de remoção, os valores de BP e de T100 são maiores. Já em relação aos três momentos de remoção, os resultados não indicam grande diferença entre eles. Ainda assim, os melhores resultados foram para a remoção depois da aplicação das medidas.

Em relação aos valores de AUC[7], o coeficiente Dice apresenta desempenho próximo ao melhor desempenho para o caso em que não é feita a remoção de stopwords.

[6] `http://kodiak.cs.cornell.edu/kddcup/software.html`

[7] É necessário ressaltar que AUC é comparável somente dentro de uma mesma coluna das Tabelas 5 e 6.

No entanto, nos casos de remoção, o valor de AUC para Dice se distancia bastante do melhor desempenho. Isso indica que o coeficiente Dice dá os menores scores para boa parte dos bigramas que são retirados nos casos de remoção. Fato inverso ocorre com a frequência de n-gramas: o valor de AUC é bastante ruim sem a remoção de stopwords e comparável ao melhor desempenho no caso de remoção. Esses resultados para Dice e para TF explicam os resultados mais estáveis de Mutual Expectation, pois esta combina explicitamente o resultado daquelas.

A Tabela 7 mostra os resultados da extração de unidades terminológicas trigramas. No contexto da língua portuguesa, remover um trigrama que contenha uma stopword pode não ser apropriado. Isso acontece porque várias unidades terminológicas trigramas em português tem uma preposição, e preposições geralmente são consideradas stopwords (o que é o caso neste trabalho). Desse modo, inicialmente sem realizar a remoção de stopwords, foi utilizada a medida Mutual Expectation, que obteve um desempenho mais consistente para bigramas, e a medida Log-likelihood e frequência de n-gramas, que obtiveram os melhores resultados no geral. De acordo com os resultados, a medida ME e a TF se mostram as mais apropriadas no caso da avaliação por BP e T100, sendo que a TF é melhor em AUC.

	BP	T100	AUC
LL	0	0	**92,0**
ME	**11,8**	**7**	77,7
TF	**11,8**	**7**	86,0

	BP	T100	AUC
LL	0	0	**76,1**
ME	**4,20**	**4**	47,3
TF	**4,20**	3	73,8

Tabela 7. Extração de unidades terminológicas trigramas sem remoção de stopwords no córpus de revestimentos cerâmicos (esq.) e no *CórpusEco* (dir.)

A Tabela 8 mostra os resultados da extração de unidades terminológicas quando é feita a eliminação de trigramas que contenham duas ou três stopwords. De acordo com os resultados, a medida ME e a TF continuam a ser as mais apropriadas no caso da avaliação por BP e T100, sendo que a TF ainda é melhor em AUC. No entanto, os ganhos em se fazer a remoção não são tão expressivos quanto o caso de bigramas. Isso pode ter sido causado pelo fato de ainda haver trigramas com uma única stopword.

	BP	T100	AUC
LL	0	0	**93,2**
ME	**11,8**	**9**	75,6
TF	**23,5**	**9**	86,5

	BP	T100	AUC
LL	0	1	**79,5**
ME	**5,63**	**4**	37,1
TF	4,23	**4**	74,7

Tabela 8. Extração de unidades terminológicas removendo trigramas com duas ou três stopwords no córpus de revestimentos cerâmicos (esq.) e no *CórpusEco* (dir.)

5. Conclusão

Este trabalho analisou os efeitos de se fazer a remoção de stopwords na abordagem estatística de extração automática de termos, uma das etapas consideradas na criação de produtos terminológicos assistida por computador. Através de experimentos no córpus de revestimentos cerâmicos e no *CórpusEco*, foi possível analisar os efeitos de considerar ou não considerar a remoção de stopwords, inclusive verificando os três momentos possíveis de se fazer tal remoção.

Os experimentos mostraram que, em geral, as medidas estatísticas para bigramas se beneficiaram da remoção de stopwords. O momento em que ela é feita, no entanto, não teve efeito no resultado final. Os resultados da remoção depois da aplicação das medidas estatísticas não foram muito melhores do que a remoção antes da aplicação das medidas ou no momento da tokenização. Em relação às medidas para trigramas, no entanto, não há ganho de desempenho fazendo a remoção de stopwords.

Outro resultado interessante é que a remoção de stopwords faz a frequência dos n-gramas ser a medida mais efetiva no caso geral. No caso em que não exista uma lista de stopwords disponível ou não seja possível aplicar a remoção de stopwords, verifica-se que a medida *Mutual Expectation* é a mais apropriada para bigramas, principalmente quando há interesse nos bigramas com maior score (por exemplo, quando um humano precisa inspecionar os termos extraídos).

Agradecimentos

O autor deseja agradecer à FAPESP e ao CNPq pelo apoio recebido durante a realização deste trabalho.

Referências

Alegría, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S., e Urizar, R. (2004). "Linguistic and stadistical approaches to basque term extraction". *Actes de GLAT-Barcelona 2004*, páginas 235–246.

Almeida, G., Oliveira, L., e Aluísio, S. (2006). "A terminologia na era da informática". *Ciência e Cultura*, 58(2).

Banerjee, S. e Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, páginas 370–381, Mexico City.

Dias, C. A. (2000). "Terminologia: conceitos e aplicações". *Ci. Inf*, 29(1).

Dias, G., Guillore, S., Bassano, J. C., e Lopes, J. G. P. (2000). Combining linguistics with statistics for multiword term extraction: A fruitful association". In *Proceedings of Recherche d'Informations Assistee par Ordinateur*.

Fawcett, T. (2004). "ROC graphs: Notes and practical considerations for researchers". *Machine Learning*, 31.

Kageura, K. e Umino, B. (1996). "Methods of automatic term recognition: A review". *TERMINOLOGY AMSTERDAM*, 3:259–290.

Pantel, P. e Lin, D. (2001). A statistical corpus-based term extractor. In *Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, páginas 36–46. Springer-Verlag.

Teline, M. F. (2004). Avaliação de métodos para a extração automática de terminologia de textos em português.

Zavaglia, C., Oliveira, L., Nunes, G. V., e Aluísio, S. M. (2007). Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio da ecologia. In *Anais do XXVII Congresso da Sociedade Brasileira de Computação e do V Workshop em Tecnologia da Informação e da Linguagem Humana*, páginas 1575–1584.