Aplicando Pontos de Corte para Listas de Termos Extraídos

Lucelene Lopes, Renata Vieira

PPGCC – FACIN – PUCRS
Av. Ipiranga, 6681 – 90.619-900
Porto Alegre – RS – Brazil

{lucelene.lopes, renata.vieira}@pucrs.br

Abstract. *This paper presents a practical study to find out terms that are potentially relevant to a given domain. Considering the use of many term extraction process, it is assumed the availability of domain term list, duly ordered according to a relevance criteria. In such way, a term selection police is proposed through the definition of a cut-off point to the term list, i.e., a automatized choice of which terms to discard. The proposed police options were analyzed to terms extracted from a brazilian portuguese corpus and their results were quantitatively analyzed according to a previously defined reference list.*

Resumo. *Este artigo apresenta um estudo prático para escolher termos que são potencialmente relevantes para um domínio. Considerando diversos processos de extração de termos utilizados, assume-se a existência de uma lista de termos de um domínio, e que esses termos sejam ordenados de acordo com um critério de relevância. Dessa forma, propõem-se uma política de seleção de termos que consiste na definição de um ponto de corte, ou seja, uma escolha automática de quais termos descartar. As opções da política proposta foram analisadas sobre termos extraídos de um corpus em português brasileiro, e os resultados quantificados de acordo com uma lista de referência previamente definida.*

1. Introdução

A identificação automática de termos relevantes para um domínio é uma tarefa de extrema relevância em diversas aplicações de processamento de linguagem natural. Por exemplo, na construção automática de ontologias, determinar os conceitos de um domínio é uma etapa fundamental da qual depende toda a qualidade das etapas subsequentes [Maedche and Staab 2001, Cimiano 2006, Lopes 2012].

Uma parte importante da tarefa de identificação de termos relevantes é a extração de termos empregados em um domínio e o cálculo de um índice de relevância para cada termo extraído. Para a extração de termos, diversas ferramentas [Banerjee and Pedersen 2003, Lopes et al. 2009a] fornecem opções qualificadas para extrair termos de corpora de domínio. Igualmente, o cálculo de um índice de relevância para cada termo extraído também possui diversas opções [Manning and Schütze 1999, Chung 2003, Kit and Liu 2008, Kim et al. 2009] que fornecem a possibilidade de ordenações de termos de acordo com sua relevância para o domínio.

No entanto, mesmo assumindo uma lista de termos extraídos, devidamente ordenados segundo um índice de relevância, resta o problema de quantos termos considerar suficientemente relevantes para serem representativos do domínio. Todo processo de

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 79–87,
Fortaleza, CE, Brazil, October 21–23, 2013.

extração está sujeito a fornecer alguns termos que não são necessariamente relevantes para o domínio. A classificação segundo um índice de relevância tende a retirar das primeiras posições das listas esses termos.

Utilizando métricas de precisão e abrangência, usuais da área de recuperação de informação [van Rijsbergen 1975], fica claro que se quisermos maximizar a precisão devemos considerar apenas o mais relevante dos termos e descartar os demais. Se por outro lado, se quisermos maximizar a abrangência devemos considerar todos termos extraídos, ou seja, não descartar termo nenhum. Se quisermos um equilíbrio entre precisão e abrangência devemos ter uma política para escolha de ponto de corte que forneça um número adequado de termos classificados como relevantes.

Dessa forma, o objetivo desse artigo é propor uma política de escolha de ponto de corte que possa equilibrar a precisão e abrangência da lista de termos relevantes de um domínio. Para atingir esse objetivo, desenvolve-se um estudo prático onde aplica-se diversas políticas de ponto de corte a listas de termos extraídos de um corpus de Pediatria [Coulthard 2005] e compara-se a precisão e abrangência obtidos frente a uma lista de referência que assume-se como os termos relevantes desse domínio.

A próxima seção apresenta algum conceitos básicos sobre pontos de corte, além da definição formal das métricas usuais de precisão e abrangência. A terceira seção apresenta os resultados práticos obtidos com diversas opções de políticas de ponto de corte. A quarta seção propõem uma política de pontos de corte que julga-se a mais adequada frente aos estudos práticos desenvolvidos. Finalmente, as considerações finais sumarizam a contribuição desse artigo e sugere trabalhos futuros.

2. Conceitos Básicos

Nessa seção apresenta-se três tipos distintos de pontos de corte: Pontos de corte absolutos; Pontos de corte por limiar; e Pontos de corte relativos. Apresenta-se em seguida a definição formal das métricas de qualidade em recuperação de informação: Precisão, Abrangência e Medida F.

2.1. Pontos de Corte Absolutos

A maneira mais simples de se aplicar pontos de corte é escolher um número arbitrário de termos que serão considerados. Porém, é importante salientar que em muitos trabalhos da literatura [Yang and Callan 2008, Lopes et al. 2009b, Evert 2010, Ding et al. 2011], as listas geradas separam os termos segundo o número de palavras que os compõem, ou seja, trata-se separadamente listas de unigramas, bigramas, trigramas, *etc.*

Essa análise em separado faz sentido, uma vez que os termos tendem a apresentar variações distintas para os índices, segundo o número de palavras que os compõem. Dessa forma, o estudo de pontos de corte será feito escolhendo um ponto de corte para unigramas, outro para bigramas, e assim por diante.

2.2. Pontos de Corte por Limiar

Uma forma popular de descartar termos é o uso de pontos de corte através da determinação de limiares arbitrários de ocorrências de termos no *corpus*. Por exemplo, o trabalho de Bourigault e Lame [Bourigault and Lame 2002] sugere o uso de um número mínimo de

10 ocorrências para considerar um termo relevante. Essa forma de identificar termos relevantes, corresponde à escolha de um ponto de corte baseado em limiar, ou seja, organizar a lista de termos extraídos segundo um índice e considerar apenas os termos nos quais o seu índice possui um valor acima do limiar escolhido. No caso de Bourigault e Lame [Bourigault and Lame 2002], o índice escolhido foi a frequência absoluta de termos, porém qualquer índice poderia ser escolhido.

O uso de pontos de corte por limiar baseados na frequência absoluta é adotado com base em um raciocínio intuitivo, que sugere uma relação direta entre o tamanho do *corpus* e o ponto de corte a escolher [Wermter and Hahn 2005]. Esta intuição, ainda que verdadeira, não é uma relação linear, pois o número de ocorrências de termos em um *corpus* decresce exponencialmente [Spärck-Jones 1972].

O formato da curva de decréscimo exponencial pode variar bastante segundo o método de extração, por exemplo, para palavras extraídas segundo um processo puramente estatístico, o decréscimo segue a lei de Zipf[1] [Zipf 1935]. No entanto, para processos linguísticos de extração de termos, não se observa esta mesma lei, como poderá ser verificado pelo número de ocorrência dos 10 termos mais frequentes do *corpus* de Pediatria apresentado na próxima seção.

Por essa razão, é difícil propor uma fórmula que permita estimar automaticamente um limiar para ponto de corte a partir do tamanho do *corpus*. Logo, nesse artigo analisam-se diversos valores de limiar escolhidos de forma arbitrária.

2.3. Pontos de Corte Relativos

Uma alternativa de pontos de corte, encontrada na literatura [Maynard et al. 2008], é manter apenas um percentual da lista extraída.

Essa alternativa é denominada ponto de corte relativo, pois define-se o tamanho da lista a ser considerada proporcional ao total de termos extraídos. A vantagem mais clara dessa abordagem é o fato de que pontos de corte relativos podem ser facilmente aplicados a qualquer tipo de lista seja ela organizada por qualquer métrica.

No entanto, resta o problema de escolher o percentual adequado, razão pela qual nesse artigo analisa-se o comportamento de diversos valores de pontos de corte relativos escolhidos de forma arbitrária.

2.4. Métricas

A comparação de listas de conceitos nesse artigo é feita através de três índices oriúndos da área de teoria da informação e de uso frequente na área de recuperação de informação. Esses índices são as tradicionais medidas de precisão (em inglês: *precision* - P), abrangência (em inglês: *recall* - R) e medida F (em inglês: *f-measure* - F) [van Rijsbergen 1975].

Essas medidas são utilizadas para comparar dois conjuntos, por exemplo, duas listas de termos. Um desses conjuntos, denominado $\mathcal{LR}$ (lista de referência), contém os termos de referência considerados corretos para o propósito, ou seja, o alvo da identificação

[1]Segundo a lei de Zipf, a frequência de uma palavra em um *corpus* é inversamente proporcional a sua posição (*rank*). Dessa forma, a palavra mais frequente de um *corpus* possui: o dobro de ocorrências do que as ocorrências da segunda palavra mais frequente; o triplo de ocorrências do que as ocorrências da terceira palavra mais frequente; e assim por diante.

de conceitos. O outro conjunto, denominado $\mathcal{LE}$ (lista extraída), contém os termos a comparar com a referência, ou seja, os termos extraídos que por alguma política foram escolhidos como conceitos.

A precisão (P) é dada pela equação abaixo que expressa a razão entre o número de termos da lista de referência que foram extraídos e considerados (tamanho da intersecção entre os conjuntos $\mathcal{LR}$ e $\mathcal{LE}$) e o tamanho da lista de termos extraídos e considerados ($\mid \mathcal{LE} \mid$). Dessa forma, a precisão expressa o percentual de termos corretamente extraídos, ou seja, o percentual dos termos localizados como corretos, quantos são efetivamente corretos.

$$P = \frac{\mid \mathcal{LR} \cap \mathcal{LE} \mid}{\mid \mathcal{LE} \mid}$$

A abrangência (R) é semelhante à precisão, porém expressa a razão entre o número de termos da lista de extraídos e considerados ($\mathcal{LE}$) presentes na lista de referência ($\mathcal{LR}$) e o tamanho da lista de referência ($\mid \mathcal{LR} \mid$). Dessa forma, a abrangência expressa o percentual de termos da lista de referência coberta pela extração de termos feita.

$$R = \frac{\mid \mathcal{LR} \cap \mathcal{LE} \mid}{\mid \mathcal{LR} \mid}$$

A medida F (F) expressa o equilíbrio entre os valores de precisão e abrangência. A sua expressão numérica é a média harmônica entre os valores de P e R. Os valores da medida F são valores situados entre P e R, e quanto maior for a diferença entre esses valores, mais próxima a medida F será do menor valor entre eles.

$$F = \frac{2 \times P \times R}{P + R}$$

O uso desses índices de qualidade é bastante difundido em diversas áreas, *e.g.* [Boreczky and Rowe 1996, Thomas et al. 2000, Fernandes et al. 2010]. Na área de PLN, e em especial nas tarefas de extração de termos, diversos trabalhos justificam a sua validade baseados em seus resultados numéricos, *e.g.*, [Manning and Schütze 1999, Bell et al. 1999, Hulth 2004, Lopes et al. 2010b].

3. Resultados Práticos

Os resultados práticos desse artigo foram obtidos extraindo bigramas e trigramas do corpus de Pediatria [Coulthard 2005], um corpus formado por 281 textos do Jornal de Pediatria que possui 835.412 palavras, distribuídas em 27.724 frases. Para este corpus existe uma lista de referência com 1.534 bigramas e 2.660 trigramas gerada por uma equipe de linguístas e especialistas do domínio de forma semi-automática e cuidadosamente revisada [Finatto 2011].

O ponto de partida de todas as experiências com pontos de corte são listas de bigramas e trigramas extraídos utilizando a ferramenta EχATO$_{lp}$ [Lopes et al. 2009a], uma ferramenta de extração baseada em abordagem linguística (extração de sintagmas nominais) organizada segundo a relevância estimada através do índice *tf-dcf* [Lopes et al. 2012], um índice que se baseia no uso de corpora contrastantes.

3.1. Pontos de Corte Absolutos

A aplicação de pontos de corte absolutos foi feita para valores de 100 a 3.500 termos conforme descrito na Figura 1. Nessa figura mostra-se os valores de precisão, abrangência e medida F para cada valor de ponto de corte. Observa-se nesses gráficos que para bigramas o ponto de corte absoluto de 2.000 termos tem o melhor equilíbrio, enquanto que para trigramas 3.300 termos foi o melhor ponto de corte absoluto.

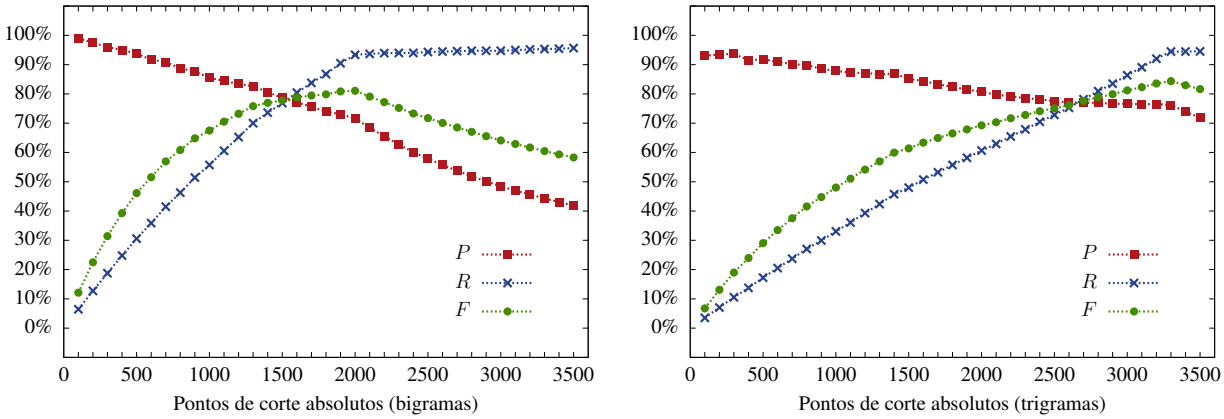

Figure 1. Precisão (P), abrangência (R), medida F (F) obtidos com pontos de corte absolutos.

3.2. Pontos de Corte por Limiar

A aplicação de pontos de corte por limiar foi feita para valores de 0 a 15 ocorrências conforme descrito na Figura 2. Cabe salientar que o limiar 0 não corresponde a nenhum descarte de termos. Observa-se nesses gráficos que para bigramas o ponto de corte por limiar com o valor 3 tem o melhor equilíbrio, enquanto que para trigramas 2 é o limiar que representa o melhor ponto de corte.

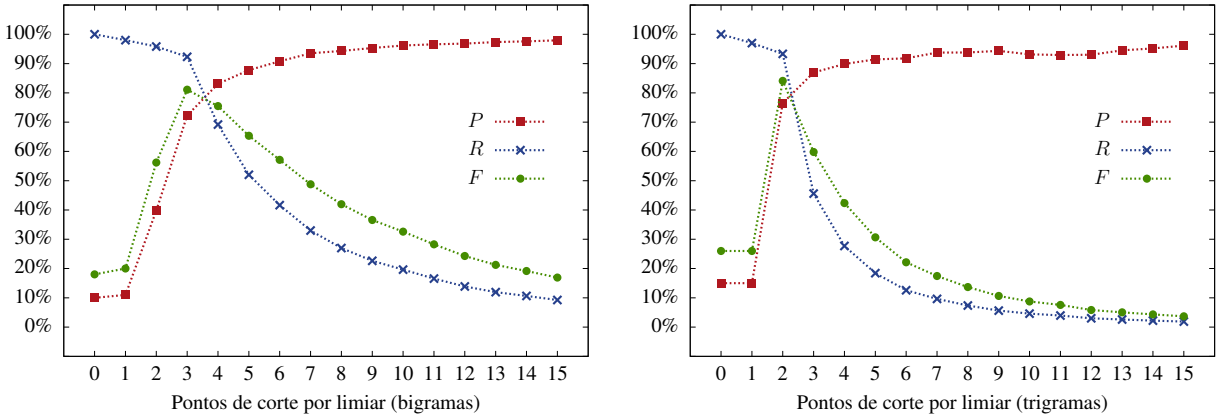

Figure 2. Precisão (P), abrangência (R), medida F (F) obtidos com pontos de corte por limiar.

3.3. Pontos de Corte Relativos

A aplicação de ponto de corte relativo foi feita levando em consideração que a ferramenta EχATO$_{lp}$ fornece 15.487 bigramas e 18.174 trigramas, e analisou-se pontos de corte relativos de 1% a 30% conforme descrito na Figura 3. Observa-se nesses gráficos que para bigramas o ponto de corte relativo de 13% tem o melhor equilíbrio, enquanto que para trigramas 18% é o melhor ponto de corte relativo.

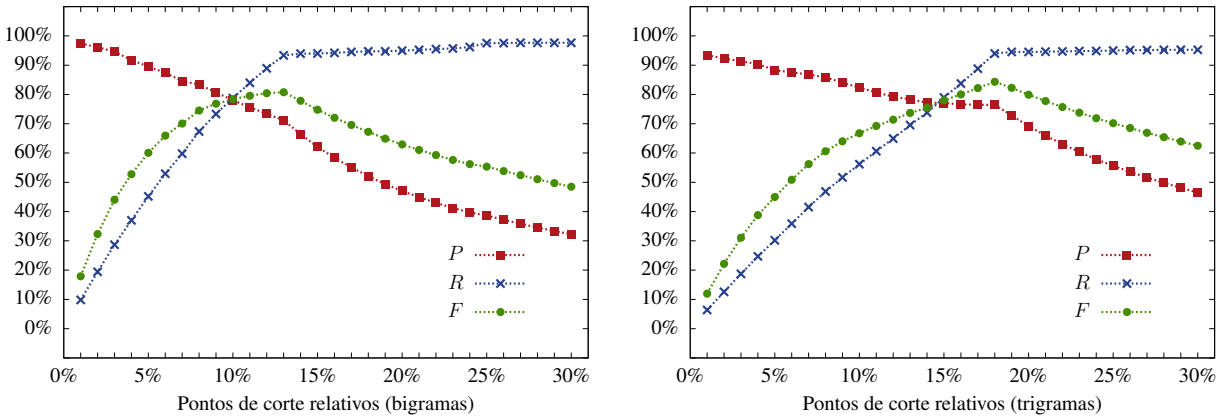

Figure 3. Precisão (P), abrangência (R), medida F (F) obtidos com pontos de corte relativos.

4. Política Proposta

Na seção anterior foram feitos experimentos sobre bigramas e trigramas extraídos do corpus de Pediatria. No entanto, o que se busca não é descobrir uma forma adequada de aplicar pontos de corte exclusivamente a essas listas de termos, mas sim, a todas as listas extraídas de todos os possíveis corpora, ou seja, uma política de ponto de corte generalizável.

A análise de um ponto de corte único, seja absoluto, por limiar, ou relativo, não parece ser possível, pois mesmo para os bigramas e trigramas do corpus de Pediatria não foi possível estabelecer um ponto de corte único que permitisse valores mais equilibrados de precisão e abrangência.

Dessa forma, propõe-se um método híbrido para estimar um ponto de corte que seja próximo dessas situações ótimas. Esse método é dito híbrido, pois segue alternativamente várias formas de ponto de corte. Especificamente, a proposta é aplicar em conjunto: um ponto de corte por limiar e um ponto de corte relativo. Cabe salientar que, o uso de ponto de corte absoluto não faz sentido nesse contexto de uma generalização, pois um ponto de corte absoluto é a definição de um número fixo e arbitrário de termos, enquanto que os demais (relativo e por limiar) são naturalmente dependentes do tamanho das listas de termos extraídos.

Aplicando um ponto de corte por limiar a todas as listas de termos extraídos, parece ser razoável descartar termos que não atingem um valor mínimo segundo o índice proposto. Dessa forma, baseado nas análises feitas, sugere-se, como primeiro passo do método híbrido proposto, descartar termos que tenham um índice *tf-dcf* inferior a 2.

Essa escolha de um limiar 2 é conservadora, pois para bigramas do corpus de Pediatria um limiar 3 foi mais adequado (melhor medida F). Cabe lembrar que, para trigramas desse mesmo corpus um limiar 2 foi mais adequado, logo a escolha de um limiar 3 iria descartar trigramas relevantes para o corpus de Pediatria.

Analogamente ao ponto de corte com limiar 2 para o índice *tf-dcf*, a segunda etapa do método híbrido proposto é o descarte de termos por um ponto de corte relativo. Com base nos resultados apresentados, onde um ponto de corte de 13% para bigramas e de 18% para trigramas, mostrou os melhores valores da medida F, escolheu-se utilizar o ponto de corte intermediário (15%) para ser aplicado a todas as listas de termos extraídos.

5. Considerações Finais

Nesse artigo foram feitas diversos experimentos com pontos de corte sobre um corpus de Pediatria para o qual possui-se uma extração de termos bastante qualificada e uma lista de referência confiável. Nesse contexto, os experimentos práticos indicaram que é possível aproximar do ponto ótimo de equilíbrio utilizando uma política de ponto de corte híbrido, onde considera-se 15% da lista extraída e exclui-se os termos onde o índice *tf-dcf* é inferior ao limiar 2.

Uma possibilidade de extensão do trabalho realizado seria considerar formas mais sofisticadas de análise baseada em um conjunto maior de dados, além do índice de relevância único, que no caso desse artigo foi a frequência *tf-dcf*. Este tipo de iniciativa equivaleria a uma análise multivalorada que é comum a métodos de aprendizagem de máquina [Mitchell 1997]. Dentre essas técnicas pode-se imaginar diversos métodos de classificação [Bauer and Kohavi 1999, Lopes et al. 2008, Witten et al. 2011] que possam trazer uma forma distinta de escolher os termos relevantes.

Apesar desses interessantes trabalhos futuros, a política proposta nesse artigo mostra-se adequada para os experimentos feitos, logo outro trabalho futuro natural é realizar um número maior de experimentos. Esses novos experimentos podem incluir o uso de outras ferramentas de extração, *e.g.*, [Bourigault and Lame 2002, Banerjee and Pedersen 2003, Lopes et al. 2010a], ou ainda de outros índices de relevância, *e.g.*, [Wu et al. 2008, Park et al. 2008, Kim et al. 2009]. Eventualmente, também pode ser interessante, caso exista disponibilidade de listas de referência, experimentos com outros corpora de domínio.

A contribuição desse artigo já pode ser percebida por fornecer algumas conclusões baseadas em um estudo prático, pois os autores desconhecem trabalhos que forneçam estimativas semelhantes de políticas de pontos de corte para escolher termos relevantes de um domínio. Nesse sentido, o presente trabalho oferece algumas estimativas iniciais que podem auxiliar pesquisadores e profissionais na definição de pontos de corte para seus próprios experimentos da área de identificação de termos relevantes.

References

Banerjee, S. and Pedersen, T. (2003). The design, implementation and use of the ngram statistics package. In *4th ITPCL*, pages 370–381.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 36(1-2):105–139.

Bell, T., Witten, I., and Moffat, A. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco.

Boreczky, J. S. and Rowe, L. A. (1996). Comparison of video shot boundary detection techniques. *Journal of Electronic Imaging*, 5(2):122–128.

Bourigault, D. and Lame, G. (2002). Analyse distributionnelle et structuration de terminologie. application a la construction d'une ontologie documentaire du droit. *Traitement automatique des langues*, 43(1).

Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology*, 9(2):221–246.

Cimiano, P. (2006). *Ontology learning and population from text: algorithms, evaluation and applications*. Springer.

Coulthard, R. J. (2005). *The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus*. PhD thesis, UFSC.

Ding, J., Zhou, S., and Guan, J. (2011). mirfam: an effective automatic mirna classification method based on n-grams and a multiclass svm. *BMC Bioinformatics*, 12(1):216.

Evert, S. (2010). Google web 1t 5-grams made easy (but not for the computer). In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, WAC-6 '10, pages 32–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fernandes, P., Lopes, L., and Ruiz, D. D. A. (2010). The impact of random samples in ensemble classifiers. In *SAC'10: Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1002–1009, New York, USA. ACM.

Finatto, M. J. (2011). Textecc – textos técnicos e científicos. `http://www.ufrgs.br/textecc`. (último acesso em 13 dezembro 2011).

Hulth, A. (2004). Enhancing linguistically oriented automatic keyword extraction. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT/NAACL, pages 17–20, New York, USA. ACM.

Kim, S. N., Baldwin, T., and Kan, M.-Y. (2009). Extracting domain-specific words - a statistical approach. In Pizzato, L. and Schwitter, R., editors, *Proceedings of the 2009 Australasian Language Technology Association Workshop*, pages 94–98, Sydney, Australia. Australasian Language Technology Association.

Kit, C. and Liu, X. (2008). Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2):204–229.

Lopes, L. (2012). *Extração automática de conceitos a partir de textos em língua portuguesa*. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Lopes, L., Fernandes, P., Vieira, R., and Fedrizzi, G. (2009a). ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora. In *Proceedings*

of the 4th Language & Technology Conference (LTC '09), pages 427–431. Faculty of Mathematics and Computer Science of Adam Mickiewicz University.

Lopes, L., Oliveira, L. H., and Vieira, R. (2010a). Portuguese term extraction methods: Comparing linguistic and statistical approaches. In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*.

Lopes, L., Scalabrin, E. E., and Fernandes, P. (2008). An empirical study of combined classifiers for knowledge discovery on medical data bases. In *APweb 2008 Workshops (LNCS 4977)*, pages 110–121.

Lopes, L., Vieira, R., Finatto, M., and Martins, D. (2010b). Extracting compound terms from domain corpora. *Journal of the Brazilian Computer Society*, 16:247–259. 10.1007/s13173-010-0020-4.

Lopes, L., Vieira, R., Finatto, M. J., Zanette, A., Martins, D., and Ribeiro Jr., L. C. (2009b). Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area. *RECIIS*, 3(1):72–84.

Maedche, A. and Staab, S. (2001). Learning ontologies for the semantic web. In *SemWeb*.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population. In *Proceedings of the 2008 conference on Ontology Learning and Population*, pages 107–127, Amsterdam, The Netherlands. IOS Press.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Park, Y., Patwardhan, S., Visweswariah, K., and Gates, S. C. (2008). An empirical analysis of word error rate and keyword error rate. In *INTERSPEECH*, pages 2070–2073.

Spärck-Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. (2000). Automatic extraction of protein interactions from scientific abstracts. In *Pacific Symposium on Biocomputing*, volume 5, pages 538–549.

van Rijsbergen, C. J. (1975). *Information Retrieval*. Butterworths, London.

Wermter, J. and Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proc. of the Conf. on Human Language Technology*, HLT '05, pages 843–850, Stroudsburg, PA, USA. Assoc. for Comput. Ling.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 3 edition.

Wu, H. C., Luk, R. W. P., Wong, K. F., and Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Trans. Inf. Syst.*, 26:13:1–13:37.

Yang, H. and Callan, J. (2008). Ontology generation for large email collections. In *Proceedings of the 2008 international conference on Digital government research*, dg.o '08, pages 254–261. Digital Government Society of North America.

Zipf, G. K. (1935). *The Psycho-Biology of Language - An Introduction to Dynamic Philology*. Houghton-Mifflin Company, Boston, USA.