O Efeito do uso de Diferentes Formas de Geração de Termos na Compreensibilidade e Representatividade dos Termos em Coleções Textuais na Língua Portuguesa

Merley da Silva Conrado[1], Ricardo Marcondes Marcacini[1], Maria Fernanda Moura[1,2], Solange Oliveira Rezende[1]

[1] Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP)
Caixa Postal 668 – 13560-970 – São Carlos – SP – Brasil
[2] Embrapa Informática Agropecuária
Caixa Postal 6041 – 13083-970 – Campinas – SP – Brasil
{merleyc,mnanda,solange}@icmc.usp.br, marcacini@grad.icmc.ups.br

Resumo A geração de termos em coleções textuais pode ser empregada para diversos fins nos processos de extração de conhecimento. Tendo em vista sua importância, neste trabalho, avaliou-se os efeitos produzidos na compreensibilidade e representatividade dos atributos quando utilizadas diferentes formas de geração de termos a partir de coleções textuais. Estas diferentes formas de geração de termos utilizam principalmente três técnicas de simplificação de termos: radicalização, lematização e substantivação. Para apoiar a avaliação da compreensibilidade dos termos foram utilizadas taxonomias de tópicos, possibilitando a avaliação subjetiva dos especialistas do domínio, tendo a técnica de substantivação, seguida da lematização, como a que obtém termos mais compreensíveis. Já para apoiar a avaliação objetiva da representatividade dos termos em relação as coleções textuais, utilizou-se a medida CTW, sendo que a técnica de radicalização mostrou-se mais eficaz na recuperação de termos em um vocabulário expandido, considerando para ambos o mesmo domínio.

Palavras-chave: Geração de termos, técnicas de simplificação de termos, compreensibilidade, representatividade

1 Introdução

Tarefas para geração de termos têm se destacado nas atividades de extração e organização do conhecimento em que a compreensibilidade, representatividade e o número de termos têm impacto direto na interpretação dos resultados. A geração de termos é uma das principais tarefas realizadas na etapa de Pré-Processamento de Mineração de Textos (MT); ela subsidia a interpretação dos termos e resultados obtidos.

Neste trabalho, considera-se como **termo**, **atributo** ou **característica** uma palavra (*unigrama* ou termo simples) ou composição de palavras (termo composto, como *bigrama* e *trigrama*), removidas as inflexões ou não. Para a tarefa de geração de termos destacam-se três passos principais: (i) a extração dos diferentes termos simples contidos nos documentos; (ii) a simplificação destes termos; e (iii) a combinação de termos simples consecutivos que juntos apresentem significado único, formando termos compostos. Nessa tarefa, três aspectos muito importantes devem ser observados: (i) o número de termos gerados, dado que um alto número de termos gerados pode ter impacto negativo na eficiência de processos de Mineração de Texto; (ii) a compreensibilidade dos termos, após sua simplificação; e (iii) a representatividade dos termos, devendo os termos gerados conseguir representar corretamente e de maneira mais completa possível o domínio trabalhado.

Existem diversas técnicas na literatura para efetuar a simplificação (remoção de inflexões e variações morfológicas) dos termos. Dentre as principais, destacam-se a radicalização, a lematização e a substantivação. Para a combinação de termos simples, diversas medidas estatísticas são utilizadas para determinar a significância dos termos compostos, como o teste da razão de máxima verossimilhança [7]. Devido a importância de se verificar o efeito de compreensibilidade e representatividade dos termos gerados e a ausência de trabalhos na literatura que efetuem uma avaliação abrangente das três técnicas de simplificação citadas juntamente com a posterior obtenção de termos compostos, no presente artigo é proposta uma avaliação dos termos gerados com estas técnicas para algumas coleções de textos na Língua Portuguesa, levando em consideração a quantidade de termos gerados por cada técnica. A avaliação da compreensibilidade dos conjuntos de termos gerados é feita por especialistas do domínio por meio da utilização de taxonomias de tópicos. Já a avaliação da representatividade destes termos em relação ao domínio trabalhado é feita de forma objetiva, utilizando, a medida CTW (*Context term weight*) [8], e subjetivamente pelos especialistas.

A seguir, discutem-se as técnicas para geração de termos. Na Seção 3, é descrita a metodologia de geração de termos. Na Seção 4, são descritos os experimentos e os resultados obtidos, e as conclusões são descritas na Seção 5.

2 Técnicas para Geração de Termos

As técnicas para geração de termos incluem desde as técnicas de simplificação de termos, como a radicalização, lematização e substantivação; até a combinação de termos simples e o uso de *thesaurus* e taxonomias. Para exemplificar a aplicação de cada uma das técnicas de simplificação de termos utilizada, neste trabalho, será considerada a frase *"Técnicas relacionadas à Inteligência Artificial"* após sua limpeza, ou seja, remoção de *stopwords*, que no caso corresponde à contração do artigo "a" e da preposição "a" ("*à*"), e , e transformação de todos os caracteres para sua forma minúscula. *Stopwords* são palavras que nada acrescentam à representatividade dos termos ou que sozinhas nada significam.

A **radicalização** ("stemmização" ou *stemming*) reduz as palavras às suas formas inflexionáveis e, às vezes, às suas derivações, ou seja, eliminação de prefixos e sufixos das palavras ou à colocação de um verbo em sua forma infinitiva [7], sendo cada palavra analisada isoladamente (exemplo: *tecnic relacion inteligenc artifici*). Segundo Aranha [2], a radicalização pode ser vista como *radicalização inflexional*, em que se considera apenas as remoções de flexões verbais, ou *radicalização para a raiz* que se realiza a remoção de todas as formas de prefixos e sufixos dos termos, sendo esta última a forma mais agressiva de radicalização. Como algoritmos de radicalização para o português, cita-se o Porter - Português[3], PortugueseStemmer [10], Pegastemming[4] e STEMBR [1].

A **lematização** visa agrupar as variantes de um termo em um único lema (exemplo: *tecnicas relacionar inteligencia artificial*). Considera-se como lema o conjunto de palavras com a mesma raiz e a mesma classe léxico-gramatical. Já para o português, tem-se o Lematizador de Nunes [9], além de outras ferramentas que etiquetam morfologicamente as palavras, como MXPOST [11], TreeTagger [12] e o etiquetador de BRILL [4], sendo necessário, em seguida, a utilização de ferramenta que lematize tais palavras.

Na **substantivação** ou "Nominalização", as palavras passam a exibir um comportamento sintático/semântico semelhante àquele próprio de um nome[5] (exemplo: *tecnica relacionar inteligencia artificial*). Pode-se citar, para o português, a ferramenta FORMA [6] que reduz as palavras dos textos aos seus *tokens* e os etiqueta morfologicamente, para, então, aplicar ao resultado desta última a ferramenta CHAMA [6], responsável pela nominalização das palavras.

Ao optar por uma dessas técnicas, pode-se buscar combinações de termos, ou seja, seqüências de duas ou mais palavras que possuem características sintática e semântica de uma unidade, como *inteligência* e *inteligência artificial*. Um método simples para encontrá-las é a contagem de ocorrências de *n-gramas* (seqüência de 'n' *tokens* - palavras adjacentes) nos textos, selecionando os mais freqüentes. Entretanto, nem sempre os termos compostos mais freqüentes são semanticamente significativos. Para obter maior significância, é necessário utilizar métodos estatísticos para verificar a importância das associações entre as palavras que compõem estas combinações, utilizando medidas como os testes do χ^2 ou da razão de máxima verossimilhança. Para aplicá-los, pode-se utilizar o pacote NSP (*Ngram Statistics Package*) [3] e, para encontrar os *n-gramas*, em português, tem-se tanto o NSP quanto a ferramenta PreText [13].

A geração de termos compostos com maior significância visa adicionar maior compreensibilidade aos resultados da extração de conhecimento, uma vez que estes termos compostos trazem consigo uma carga semântica maior que apenas termos simples, tornando a interpretação do conhecimento mais intuitiva. O processo de geração de termos torna-se mais importante quando a compreensibilidade dos termos obtidos afeta os resultados, como em uma análise de uma taxonomia de tópicos obtida automaticamente. Um eficiente processo de geração

[3] *Snowball* - http://snowball.tartarus.org/index.php
[4] *Pegastemming* - http://www.inf.pucrs.br/ gonzalez/ri/pesqdiss/analise.htm
[5] Gramática Tradicional - http://www.dacex.ct.utfpr.edu.br/paulo3.htm

de termos, por gerar somente termos representativos ao domínio, acarreta na diminuição do número de atributos utilizados pelos algoritmos de aprendizado.

3 Metodologia para Geração de Termos

A tarefa de geração de termos deve ser efetuada cuidadosamente, de tal forma que um conjunto de etapas seja executado convenientemente para alcançar um objetivo pré-estabelecido. Assim, neste trabalho, adotou-se a metodologia proposta em [5], na qual é descrita uma seqüência de etapas para obter termos representativos de um determinado domínio. Como exemplo de aplicação desta metodologia, pode-se citar o Projeto TopTax[6], que é um ambiente que viabiliza a construção de taxonomias de tópicos, sendo que os termos gerados aqui com o uso da metodologia são utilizados para melhorar a compreensão de taxonomias para domínios específicos.

Tal metodologia, ilustrada na Figura 1, inicia-se com a delimitação da **base de textos** sendo que sua qualidade deve ser obtida por meio da padronização dos formatos dos dados e limpeza dos textos. O segundo passo desta metodologia é a **aplicação das técnicas** descritas na Seção 2 que, neste trabalho, são: radicalização por meio da ferramenta PreText; lematização, usando o Lematizador de Nunes; substantivação, utilizando as ferramentas FORMA e CHAMA.

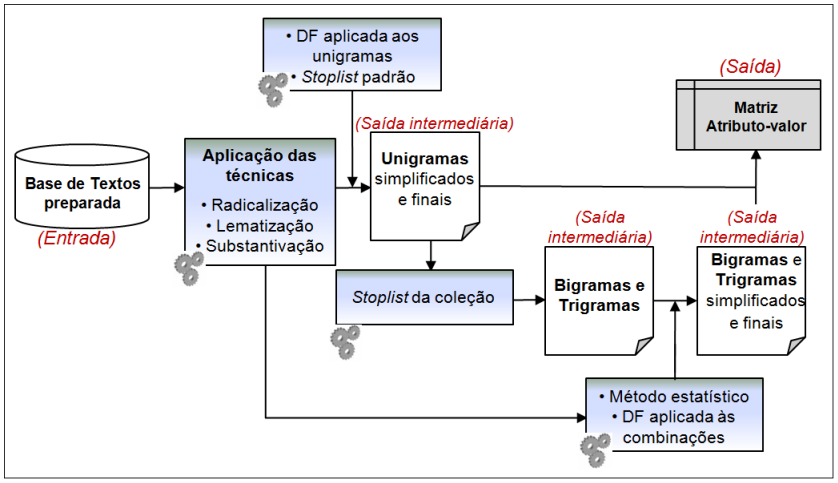

Figura 1. Metodologia para Geração de Termos

Mesmo com as bases de textos contendo somente termos simplificados de acordo com cada técnica, tem-se um elevado número de unigramas, porém nem todos representam os documentos de uma forma adequada. Para aplicar um

[6] TopTax - http://labic.icmc.usp.br/projects/researchproject.2008-06-04.9415524093

filtro simples a esses unigramas e os correspondentes *n-gramas* que deles são derivados, neste trabalho foram considerados apenas os termos que aparecem, no mínimo, em dois documentos na coleção textual (*Document Frequency* - **DF** >= 2). Além disso, remove-se uma **lista de *stopwords* padrão** para português (como artigos, interjeições, etc), obtendo após estes dois filtros os **unigramas simplificados e finais**. A partir dos unigramas originais e os finais obtidos, cria-se a ***stoplist* da coleção** - lista de *stopwords* da coleção.

Neste trabalho optou-se por utilizar unigramas, bigramas e trigramas, obtidos com a aplicação de cada uma das técnicas de simplificação de termos, eliminando a *stoplist* da coleção. Como o número dessas combinações é ainda alto e grande parte delas não tem significado semântico, **métodos estatísticos** são aplicados, como o teste da razão de máxima verossimilhança para eliminar as combinações com pouco significado semântico na coleção textual e o filtro $DF >= 2$.

A base de textos é, então, representada por uma **matriz atributo-valor**, sendo que o sucesso de tarefas de extração de padrões é diretamente afetado pela qualidade dos termos que compõem esta matriz. Ressalta-se, portanto, a importância de se obter termos representativos do domínio, sendo necessário escolher uma técnica adequada para geração de tais termos.

4 Experimentos e Análise de Resultados

Os experimentos têm como objetivo avaliar as formas de geração de termos em domínios específicos, apontando, por meio de avaliações subjetivas e objetivas, algumas das características positivas e negativas da utilização das técnicas de simplificação de termos (radicalização, lematização e substantivação), quando a compreensibilidade dos termos é importante para a análise dos resultados.

Para isso, foram utilizadas quatro bases de textos reais em português do domínio de agronegócio, sendo que seus documentos estão disponíveis no site da Embrapa - Empresa Brasileira de Pesquisa Agropecuária, especificamente sob a Agência de Informação EMBRAPA[7]. Estas bases são referentes a quatro diferentes produtos agrícolas: *milho*, *cana*, *feijão* e *caju* e a quantidade de documentos pertencente a cada base é 510, 391, 348 e 40, respectivamente.

Primeiramente, para cada base de textos e técnica de geração de termos, foram gerados (a) os termos iniciais da metodologia (unigramas, bigramas e trigramas), removendo-se somente a lista de *stopwords* padrão para português disponível na PreText, as conjugações do verbo SER e as palavras compostas por apenas um caracter. Em seguida, geraram-se (b) os termos finais, que são obtidos seguindo a metodologia descrita na Seção 3, sendo que para o teste da razão de máxima verossimilhança adotou-se $p\_value = 0.05$, visando manter somente os bigramas e trigramas com algum significado semântico. Com esta etapa, conforme mostrado na Figura 2, pode-se verificar a redução da quantidade de termos obtidos quando comparados os termos iniciais e finais, e que o uso da radicalização (R) obtém, geralmente, menos termos do que a lematização (L) e substantivação (S), já que esta é mais agressiva ao simplificar os termos.

[7] EMBRAPA - Veja http://www.agencia.cnptia.embrapa.br/.

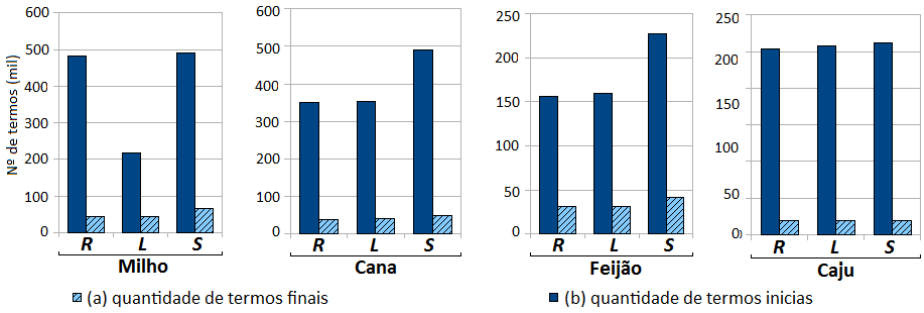

Figura 2. Redução do número de termos nas diferentes técnicas aplicadas

Com os termos finais gerados, avaliou-se objetivamente a representatividade dos mesmos em relação ao domínio em questão. Para isso, utilizou-se como suporte a medida CTW (*Context term weight*) [8], que avalia a quantidade de vezes em que um termo gerado (freqüência do termo) aparece em um determinado contexto. Neste trabalho, o contexto é representado por um vocabulário expandido, ou seja, os termos representativos e consagrados do domínio juntamente com seus sinônimos, que são obtidos em um *thesaurus* do mesmo domínio, sendo que nesta avaliação os termos foram recuperados neste vocabulário. Isso é possível, pois as bases utilizadas aqui são do domínio de agronegócio, o que permite o uso de um *thesaurus* consagrado do mesmo domínio, o Thesagro[8]. Já a freqüência de cada termo na coleção de textos utilizada no cálculo da medida CTW é obtida pela metodologia de geração de termos proposta neste artigo. A descrição formal da medida CTW adaptada para este artigo é:

$$CT(a) = \sum_{d \epsilon T_a} f_a(d) \tag{1}$$

na qual a é o termo obtido pela metodologia de geração de termos,
T_a é o conjunto de termos do vocabulário expandido,
d é a palavra do vocabulário expandido e
$f_a(d)$ é a freqüência de d na coleção de textos utilizada.

Na Tabela 1, é apresentada a pontuação CTW para cada técnica aplicada em cada base de textos e seus respectivos vocabulários expandidos, na qual pode-se observar que a técnica de radicalização geralmente é mais eficaz na recuperação de termos do vocabulário do domínio. Isto pode ser explicado pelo fato que a mesma é mais agressiva na simplificação dos termos em relação às outras duas técnicas. Ressalta-se que a escala das pontuações é proporcional ao tamanho do vocabulário expandido.

Para a avaliação subjetiva, a partir dos termos obtidos, foram geradas doze taxonomias, uma para cada forma de geração de termos e para cada base. Devido

[8] *Thesaurus* Nacional Agrícola - http://www.agricultura.gov.br/portal/page?pageid=-33,959135&dad=portal&schema=PORTAL

Bases	*# Docs*	*# Termos do Vocabulário Expandido*	*Radicalização*	*Lematização*	*Substantivação*
Milho	510	1.028	358	306	349
Cana	391	11.161	1.561	1.139	1.526
Feijão	348	8.128	809	499	588
Caju	40	65	11	3	7

Tabela 1. Pontuação CTW para cada técnica

ao tamanho das mesmas, foram escolhidos, para cada taxonomia, 10 ramos julgados subjetivamente mais semelhantes aos documentos. Estes 120 ramos foram avaliados por cinco especialistas do domínio que julgaram a representatividade dos termos em relação aos documentos destes ramos, atribuindo-lhes notas de 1 a 4, na qual 1 indica nada representativo e 4 totalmente representativo. Devido ao objetivo desta avaliação, as taxonomias serviram apenas como auxílio à avaliação, já que estas não foram avaliadas e sim os termos contidos nelas. Na Figura 3, são mostrados alguns ramos da base de *milho* avaliados pelos especialistas utilizando as técnicas de (a) radicalização, (b) lematização e (c) substantivação. Com esta figura, pode-se observar a diferente forma de simplificação de termos utilizada por cada técnica.

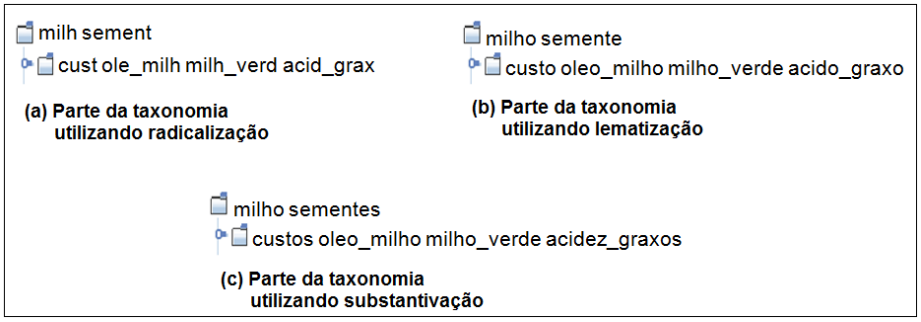

Figura 3. Alguns ramos da base de *milho* avaliados pelos especialistas

Considerando as coleções textuais como únicas, obteve-se como médias das avaliações das técnicas de radicalização, lematização e substantivação, os valores 2,65; 2,97; e 3,13, respectivamente, ou seja, a técnica de substantivação obteve as maiores notas dadas pelos especialistas. No entanto, foi realizada uma análise estatística para verificar se há diferença significativa entre os resultados das técnicas nesta avaliação. Na Tabela 2, encontram-se os resultados das comparações das notas das três técnicas utilizadas, por meio do teste estatístico não-paramétrico de *Kruskal-Wallis* para amostras não pareadas, aplicando-se o pós-teste de múltiplas comparações de *Dunn*. Observa-se que não houve evidências que indiquem diferenças significativas entre as técnicas de lematização e subs-

tantivação. No entanto, ambas técnicas diferenciaram em relação a radicalização, sendo que os ramos desta obtiveram pior avaliação entre os especialistas.

Comparação das técnicas	*Diferença estatística significativa*	*p_value*
Radicalização *vs.* Lematização	sim	$p < 0,001$
Radicalização *vs.* Substantivação	sim	$p < 0,001$
Lematização *vs.* Substantivação	não	$p > 0,05$

Tabela 2. Diferenças estatística entre os resultados das técnicas

Além disso, os especialistas indicaram, por meio de um formulário, qual das técnicas usadas, para cada base, geraram termos mais compreensíveis e qual a técnica de preferência para ser utilizada neste domínio. Para ambas perguntas, a técnica de substantivação foi a mais indicada, seguido da lematização.

5 Conclusões

A geração de termos não é um trabalho trivial e afeta diretamente a qualidade dos resultados no processo de MT. Neste trabalho, avaliou-se o efeito das diferentes formas de geração de termos em domínios específicos na língua portuguesa, apontando algumas das características positivas e negativas do uso das técnicas de simplificação de termos utilizadas para auxiliar a geração de atributos no processo de MT. As técnicas avaliadas foram a radicalização, a lematização e a substantivação.

Nos experimentos realizados, percebe-se que a aplicação de toda a metodologia de geração de termos foi extremamente relevante para a redução da quantidade de atributos, mantendo os termos mais representativos da coleção textual. Tal característica é muito importante quando se trabalha com bases de textos pequenas, sendo que a importância de tal característica aumenta proporcionalmente com o tamanho da base.

A avaliação das diferentes formas de geração de termos mostrou que, para estas bases textuais, quando analisada objetivamente, o uso da técnica de radicalização obtém melhores resultados em relação à recuperação dos termos no vocabulário expandido. Por outro lado, a compreensão dos termos gerados pela mesma foi pior avaliada subjetivamente, devido a sua maior agressividade na simplificação dos termos. Com estas avaliações, conclui-se que o uso da técnica de radicalização é mais indicado para gerar de termos do domínio do agronegócio segundo uma medida objetiva (a CTW), porém não auxilia a compreensibilidade dos termos gerados, ou seja, subjetivamente é pior avaliada. Além disso, o uso da radicalização contribui, na maioria dos casos, para a geração de uma menor quantidade de termos se comparada às técnicas de lematização e substantivação. A técnica eleita como a preferida para ser utilizada neste domínio foi a substantivação, uma vez que os termos aqui obtidos são mais compreensíveis.

Para trabalhos futuros, pretende-se adicionar ao processo de geração de termos uma análise semântica dos atributos de forma a eliminar os unigramas com menor contribuição para a representação do domínio a ser trabalhado, como por exemplo, os verbos.

Agradecimentos: os autores agradecem ao CNPq pelo apoio financeiro.

Referências

1. R. V. Alvares, A. C. Garcia, and I. Ferraz. Stembr: A stemming algorithm for the brazilian portuguese language. In *Progress in Artificial Intelligence - 12th Portuguese Conference on Artificial Intelligence (EPIA) 2005, Covilha, Portugal*, volume 3808/2005, pages 693–701. Springer Berlin / Heidelberg, 2005.
2. C. N. Aranha. *Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: sob o Enfoque da Inteligência Computacional*. PhD thesis, Departamento de Engenharia Elétrica - PUC - Rio de Janeiro, 2007.
3. S. Banerjee and T. Pedersen. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 370–381, Mexico, 2003.
4. E. Brill. Transformation-based error-driven learning of natural language: A case study in part of speech tagging. *Computational Linguistics 21*, pages 543–565, 1995.
5. M. S. Conrado, M. F. Moura, R. M. Marcacini, and S. O. Rezende. Avaliando diferentes formas de geração de termos a partir de coleções textuais. Technical Report 334, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP - São Carlos, Janeiro 2009.
6. M. A. I. Gonzalez, V. L. S. de Lima, and J. V. de Lima. Tools for nominalization: An alternative for lexical normalization. *Proceedings of the Seventh Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR) - Springer Berlin / Heidelberg*, 3960:100–109, 2006.
7. C. D. Manning, P. Raghavan, and H. Schütze. Language models for information retrieval. In *An Introduction to Information Retrieval*, chapter 12. Cambridge University Press, 2008.
8. D. Maynard and S. Ananiadou. Term extraction using a similarity-based approach. In *In Recent Advances in Computational Terminology. John Benjamins*, pages 261–278, 1999.
9. M. e. a. Nunes. The design of a lexicon for brazilian portuguese: Lessons learned and perspectives. *Proceedings of the II Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR)*, pages 61–70, 1996.
10. V. M. Orengo and C. Huyck. A stemming algorithm for portuguese language. In *Proceedings of Eigth Symposium on String Processing and Information Retrieval (SPIRE) - Chile*, pages 186–193, November 2001.
11. A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP), University of Pennsylvania*, pages 491–497, 1996.
12. H. Schmid. Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing (NeMLaP)*, pages 44–49, 1994.

13. M. V. Soares, R. C. Prati, and M. C. Monard. Pretext II: Descrição da reestruturacão da ferramenta de pré-processamento de textos. Technical Report 333, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP - São Carlos, São Carlos - SP, 2008.