Avaliando Diferentes Formas de Geração de Termos a partir de Coleções Textuais

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

[1]Instituto de Ciências Matemáticas e de Computação – 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.usp.br

Abstract. *With the fast expansion on the number of textual documents, the task of extracting representative terms of fixed domain becomes extremely important since they can be employed for various purposes. One purpose is the discrimination of texts in the processes of knowledge extraction, which allows the construction of topics taxonomies for specific areas. In this paper different methods of term generation from text collections are evaluated aiming at selecting the one(s) that best fits each domain. In order to reach this goal, subjective evaluations were carried out with some text collection domain specialists and the best results were obtained to stemming and substantivisation.*

Resumo. *Com a rápida expansão da quantidade de documentos textuais, a extração de termos representativos de domínio fixado torna-se extremamente importante, podendo ser empregados em diversos fins, como para discriminar textos nos processos de extração de conhecimento, permitindo a construção de taxonomias de tópicos de domínios específicos. Neste trabalho foram avaliadas diferentes formas de geração de termos a partir de coleções textuais com o objetivo de selecionar as mais adequadas para domínios específicos. Para isso, foram realizadas avaliações subjetivas junto a especialistas do domínio do conhecimento de cada coleção de texto analisada, tendo-se obtido melhores resultados para "radicalização" e substantivação.*

1. Introdução

A quantidade estimada de informação acrescida anualmente no universo digital no período de 2006 a 2010 corresponderá a 988 hexabytes [Gantz et al. 2007], sendo que aproximadamente 80% da informação das corporações no mundo é representada por documentos textuais[1]. A Mineração de Textos (MT) vem tornando possível transformar esse grande volume de documentos em conhecimento útil, muitas vezes inovador para as empresas. Para isso, é necessário que estes documentos sejam preparados e tratados de tal forma que se possa aplicar técnicas a fim de transformá-los em conhecimento útil. Isso é realizado em uma etapa extremamente relevante do processo de MT, denominada pré-processamento.

[1]Delphi Group - http://www.delphigroup.com/

A qualidade dos resultados, nesta etapa, depende diretamente do trabalho realizado na seleção e geração dos termos e medidas sobre os dados dos quais se pretende extrair conhecimento, uma vez que estes termos delimitam o domínio do problema trabalhado. Desta forma, o foco deste trabalho é a avaliação de técnicas de geração de termos, a fim de conseguir representações de domínios mais consistentes e, conseqüentemente, melhorar os resultados obtidos pelas demais etapas da MT.

Na próxima seção, discute-se o pré-processamento, com algumas das formas de geração de termos (atributos, palavras, características). Na Seção 3, é descrita a metodologia deste trabalho. Na Seção 4, são descritos os experimentos e os resultados obtidos, e as conclusões são apresentadas na Seção 5.

2. Pré-processamento de Textos

O processo de Mineração de Textos é tratado, neste artigo, como uma particularidade do processo de Mineração de Dados [Fayyad et al. 1996], sendo dividido em cinco grandes etapas: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Utilização do Conhecimento [Rezende et al. 2003]. O pré-processamento é crucial para o restante do processo, sendo o foco deste trabalho, especificamente na geração e avaliação de termos, que são os atributos utilizados pelas ferramentas de aprendizado no processo de MT. Nesta etapa, aplicam-se aos dados um processo de tratamento, limpeza e, em geral, redução de seu volume, sempre preservando as características necessárias para o cumprimento dos objetivos.

Um dos problemas importantes tratado nesta etapa é reduzir o número de termos, sendo este iniciado desde a padronização da coleção textual. Tal padronização é necessária, pois os documentos podem provir de bases distintas. Em seguida, deve-se analisar subjetivamente o número de documentos não danificados e, caso considerado insuficiente, deve-se buscar mais documentos. Todo o processo pode ser repetido até se obter uma coleção textual satisfatória para atingir os objetivos pré-estabelecidos para, então, fazer a limpeza dos textos (remoção de caracteres desnecessários e *stoplist*).

Em sistemas semi-automáticos ou automáticos, para a geração/identificação dos termos que descrevem o conteúdo do documento, busca-se por técnicas que simplifiquem as diversas formas de apresentação de termos mantendo o mesmo significado. **Termo**, neste trabalho, é usado no mesmo contexto que para recuperação de informação, pois se considera uma palavra ou composição de palavras, removidas as inflexões ou não. Tratam-se os termos com uma única palavra como *unigramas*, com duas palavras como *bigramas* e com três como *trigramas*; e as coleções sempre como *bag-of-words*. Para esta geração, busca-se por técnicas que simplifiquem as diversas formas de apresentação de termos mantendo o mesmo significado. Entre as técnicas mais utilizadas, encontram-se a radicalização, lematização e substantivação. Para exemplificar a aplicação de cada uma desta técnicas será utilizada a frase *"Técnicas relacionadas à Inteligência Artificial"* aplicando-na uma limpeza, ou seja, remoção de *stopwords*, que no caso corresponde ao artigo *"à"* e transformação de todos os caracteres para sua forma minúscula.

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 [Manning et al. 2008], sendo cada palavra analisada isoladamente (exemplo: *tecnic relacion inteligenc arti-*

fici). Segundo [Aranha 2007], 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". O processo de *stemming* pode depender da linguagem, pois existem algoritmos de "radicalização" que necessitam de conhecimento lingüístico [Silla Jr e Kaestner 2002]. Porém, os mais atuais não utilizam informações do contexto para determinar o sentido correto de cada palavra [Ebecken et al. 2003], devido à maioria das palavras poder ser considerada por um único significado, sendo que nestes casos, o contexto não ajudará no processo de *stemming*. No entanto, deve-se atentar aos possíveis erros resultantes de análise incorreta do sentido das palavras, já que tais algoritmos ignoram o significado dos termos levando, às vezes, a erros. Como algoritmos de radicalização para a língua inglesa, pode-se citar o de Lovins [Lovins 1968], Stemmer S [Harman 1991] e de Porter (*Porter Stemming Algorith*) [Porter 1980]; para o português e espanhol surgiram algumas adaptações a partir desses algoritmos, como a ferramenta PreTexT [Martins et al. 2003] e Stemmer [Orengo e Huyck 2001].

A **lematização** reduz as palavras a seus lemas (exemplo: *tecnicas relacionar inteligencia artificial*). Para as línguas francesa e inglesa, tem-se o software proprietário Sphinx, versão 4 [Sphinx Brasil 2008]. Já para o português, tem-se o Lematizador de Nunes [Nunes 1996], além de outras ferramentas que etiquetam morfologicamente as palavras, como MXPOST (*Maximum entropy pos tagger*) [Ratnaparkhi 1996], TreeTagger [Schmid 1994] e o etiquetador de BRILL [Brill 1995], sendo necessário, em seguida, a utilização de ferramenta que lematize tais palavras.

Na **substantivação**, também conhecido por "Nominalização", as palavras passam a exibir um comportamento sintático/semântico semelhante àquele próprio de um nome[2] (exemplo: *tecnica relacionar inteligencia artificial*). Pode-se citar, para a língua portuguesa, a ferramenta FORMA [Gonzalez et al. 2006] que toqueniza e etiqueta morfologicamente as palavras dos textos, para que se possa aplicar ao resultado desta última a ferramenta CHAMA [Gonzalez et al. 2006], responsável pela nominalização de adjetivos, advérbios e verbos.

Ao optar por uma dessas técnicas, pode-se buscar combinações de termos (localização de colocações), 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-grams* (seqüência de 'n' *tokens* - palavras adjacentes) nos textos, selecionando os mais freqüentes. Para obter resultados mais satisfatórios, é necessário utilizar métodos estatísticos para verificar a importância das associações entre as palavras que compõem estas colocações. Para os métodos estatísticos, pode-se utilizar o pacote NSP (*Ngram Statistics Package*) [Banerjee e Pedersen 2003], e para as colocações, em português, tem-se tanto o NSP quanto a PreText.

Além disso, pode haver um passo *anterior* à aplicação destas técnicas, como o reconhecimento de sinônimos ou termos hierarquicamente superiores ou inferiores aos termos analisados, bem como a substituição dos termos analisados por esses, ou melhor,

[2]http://www.dacex.ct.utfpr.edu.br/paulo3.htm

o uso de *thesaurus* [Dias-da-Silva et al. 2000] ou taxonomias [Kilgarriff e Yallop 2000].

Após a geração dos termos, é necessário representá-los em um formato que obtenha uma descrição adequada da semântica do texto para possibilitar a execução correta da tarefa alvo, de acordo com as necessidades do usuário. Há duas principais abordagens para se trabalhar com termos: *análise semântica*, que avalia a importância dos termos na frase, utilizando técnicas baseadas no processamento de língua natural; e *análise estatística* que considera a importância dos termos correspondente ao número de vezes que os mesmos aparecem no texto, tendo esta última maior independência de idioma, e é a utilizada neste trabalho.

Na análise estatística, há alguns tipos de modelo de representação dos termos e geralmente são dependentes das ferramentas eleitas para a extração de padrões, como o **modelo matriz atributo-valor**. Ao usar algum dos métodos estatísticos, normalmente é suficiente representar os dados no modelo matriz atributo-valor com a representação *bag-of-words*, que considera os termos como um conjunto desordenado de palavras independentes, sendo este utilizado neste trabalho. No trabalho de Aranha [Aranha 2007], é apresentado um modelo de pré-processamento para MT utilizando tal representação juntamente com a *bag-of-lexems*, para avaliar o conteúdo lingüístico dos termos em sua geração. Já no presente trabalho, considera-se as diferentes formas de geração de termos, verificando a(s) mais adequada(s) para domínios específicos por meio do uso de testes estatísticos e avaliação subjetiva de especialistas da área, não tendo o objetivo de avaliar o conteúdo lingüístico, já que isto traria maior custo computacional.

3. Metodologia para Geração de Termos

Os termos gerados representativos de coleções textuais podem ser usados para diversos fins. Neste trabalho, os termos gerados têm como objetivo final apoiar à metodologia do Projeto TopTax [3], possibilitando auxiliar especialistas de um domínio específico a organizar e manter a informação do mesmo, o que é possível devido à criação de uma taxonomia de tópicos para domínios específicos. Tanto a citada metodologia quanto a abordagem deste trabalho são semi-automatizadas, permitindo a interação com o usuário.

Primeiramente, deve-se **recuperar** e **padronizar** os dados da coleção textual. Tal padronização consiste em adequar os formatos dos documentos, descartando os que são danificados; separá-los por idiomas; e avaliar subjetivamente o estado da coleção, ou seja, se deve ser analisada separadamente, por idiomas e/ou tamanhos, ou se é insuficiente, até que se obtenha uma coleção satisfatória para atingir os objetivos pré-estabelecidos.

Em seguida, efetua-se a **limpeza da coleção textual**, ou seja, a remoção de caracteres desnecessários, além de transformá-los para a sua forma minúscula; verificar a existência de metadados; e remover *stoplist* (lista de palavras que nada acrescentam à representatividade dos termos ou que sozinhas nada significam).

Segue-se, então, para a aplicação de diferentes formas de geração dos termos, conforme mostradas na Figura 1 e descritas conceitualmente na Seção 2, visando escolher os que melhor representem a coleção de textos. Inicialmente, obtém-se os termos simples (unigramas), aplicando-se as técnicas de "radicalização", por meio da ferramenta PreText;

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

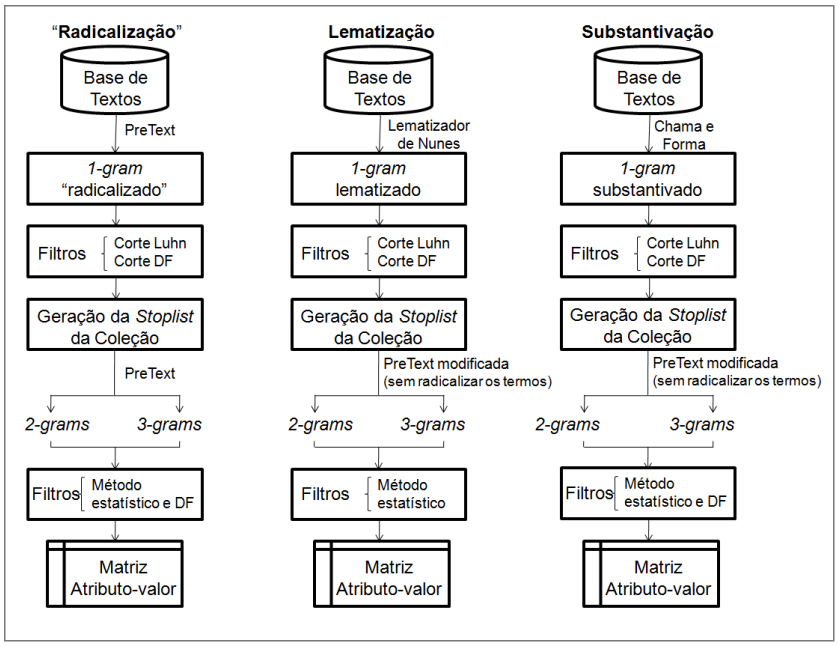

Figura 1. Metodologia para Geração de Termos.

Lematização, usando o Lematizador de Nunes; Substantivação, utilizando as ferramentas FORMA e CHAMA.

Considerando o elevado número de termos obtidos e a necessidade de trabalhar com os que mais representam os documentos, para unigramas pode-se utilizar os cortes de Luhn [Luhn 1958]. Consideram-se, também, apenas os termos que aparecerem, no mínimo, em dois documentos na coleção textual, ou seja, apenas os termos que contêm $DF >= 2$, sendo DF a *Document Frequency*, pois, potencialmente, esses termos podem ser responsáveis por discriminar um agrupamento de documentos. Com o resultado destes filtros juntamente com a remoção da *stoplist* da coleção - lista de palavras que pouco ou nada acrescentam para a representatividade da respectiva coleção - são geradas **combinações de termos** (bigramas e trigramas). Como o número dessas combinações é ainda alto e grande parte delas não tem significado semântico, aplicam-se métodos estatísticos, como o teste da razão de máxima verossimilhança, disponível no pacote NSP, para eliminar as combinações com pouco significado semântico na coleção textual; além de considerar também $DF >= 2$. Na geração de *n-grams*, pode-se citar o trabalho de Silic et al. [Silic et al. 2007], no qual o uso de *n-grams* é comparado com normalização morfológica[4], mas com foco diferente ao deste e não aplicável à língua portuguesa.

Os termos gerados são, então, **representados** na matriz atributo-valor, na qual o documento a que o termo pertence é colocado como linha da matriz, os termos gerados como colunas, e a matriz é preenchida, neste caso, com a freqüência absoluta deste termo no respectivo documento. Às técnicas de Lematização e Substantivação, no decorrer do processo, são aplicadas adaptações e cálculos para possibilitar a geração da matriz atributo-valor no formato da ferramenta PreText. O objetivo é efetuar a clusterização usando similaridade de cosseno (algoritmo average) e rotulação pelo método desenvolvido

[4]Técnicas de lematização, "radicalização" e normalização derivacional

no Projeto TopTax, para, então, construir a taxonomia para avaliação dos resultados finais. Na **avaliação dos termos gerados** é verificado se os resultados obtidos estão em conformidade com os resultados esperados, por meio da (i) geração de diferentes taxonomias de tópicos, com as diferentes formas de geração dos termos; (ii) submissão dessas taxonomias à análise subjetiva dos especialistas do domínio, com uma forma clara de apresentação dos termos gerados, permitindo avaliar qual forma é mais adequada para apoiar os especialistas do domínio em seus objetivos e possibilitar a implementação de melhorias na abordagem como um todo. Além disso, efetua-se o (iii) cálculo de precisão da representatividade dos termos gerados em relação ao domínio em questão. Por fim, o conhecimento gerado (taxonomia gerada) pode ser finalmente utilizado para o objetivo pré-estabelecido.

4. Experimentos e Análise de Resultados

Com o objetivo de avaliar as formas de geração de termos em domínios específicos, foram feitos experimentos com duas coleções de textos: base do ENIA (Encontro Nacional de Inteligência Artificial) e base do TIL (Workshop de Tecnologia da Informação e da Linguagem Humana). Ambas pertencentes à base textual composta por 743 artigos completos não-rotulados em português da área de *Inteligência Artificial*, provenientes de congressos, *workshops* e conferências.

Conforme descrito na Seção 3, os documentos foram padronizados e limpos, reduzindo a base do ENIA de 220 para 216 artigos, e a do TIL de 99 para 94 artigos. Em seguida, foi aplicada a abordagem proposta neste trabalho para obtenção de unigramas, bigramas e trigramas juntos, utilizando as três diferentes formas de geração de termos (Figura 1). Na Tabela 1, é mostrada a redução do número de termos com a aplicação de tal abordagem, utilizando em todos os casos a *stoplist* padrão para português junto à ferramenta PreText e $DF >= 2$; sendo que: a) corresponde ao número de termos gerados sem corte; b) número de termos retirando a *stoplist* da coleção c) número de termos após o corte de Luhn para unigramas e $p_value = 0.01$ para o teste da razão de máxima verossimilhança para bigramas e trigramas.

	Bases de Textos					
	ENIA			**TIL**		
Técnicas	***a***	b	***c***	***a***	***b***	***c***
"Radicalização"	558930	405271	22135	190397	171688	7984
Lematização	417080	342336	16210	176458	139395	5809
Substantivação	406033	336551	16049	232613	206344	8006

Tabela 1. Número de termos para as diferentes formas de geração

Com os termos do item *c*, foram geradas seis taxonomias, uma para cada forma de geração de termos; como o conjunto de termos utilizado é diferente, para as mesmas coleções de documentos, também as hierarquias obtidas ficam diferentes. Devido a essas diferenças, os autores escolheram ramos que julgaram subjetivamente mais semelhantes, em relação aos documentos sob eles e semântica dos rótulos, para serem avaliados por seis especialistas do domínio. Dessa forma, para cada hierarquia gerada foram escolhidos: 11 ramos para a "radicalização" do ENIA (Figura 2), 12 ramos para "radicalização" do TIL

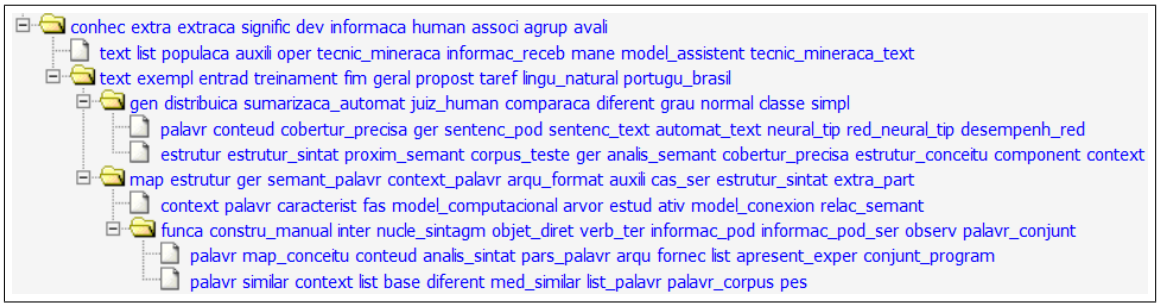

Figura 2. Ramos para a "radicalização" utilizando a base do ENIA.

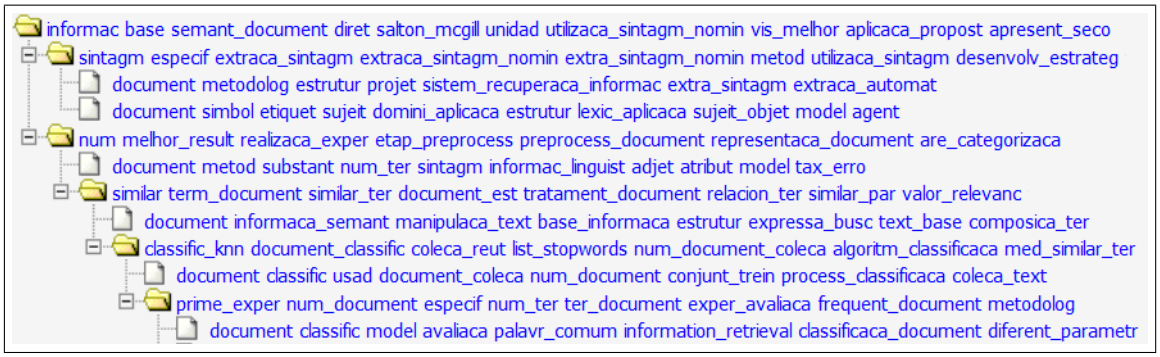

Figura 3. Ramos para a "radicalização" utilizando a base do TIL.

(Figura 3), 16 ramos para a lematização do ENIA (Figura 4), 12 ramos para lematização do TIL (Figura 5), 13 ramos para a substantivação do ENIA (Figura 6) e 12 ramos para a substantivação do TIL (Figura 7).

Os especialistas observavam as hierarquias obtidas (Figuras 2 a 7), e para cada nó das hierarquias foram atribuídas duas notas que abordaram a (i) avaliação do ramo selecionado, ou seja, foi observado a representatividade dos termos para os documentos referente ao ramo selecionado, considerando seus ascendentes também; e a (ii) avaliação geral, isto é, avaliou-se com uma "visão geral do todo" o ramo selecionado em relação a toda hierarquia, conforme ilustrado na Figura 8.

Para cada hierarquia foi dada uma nota geral, entre um (1) e quatro (4). E, ainda, para cada ramo, os especialistas examinaram os documentos e descritores obtidos, pois foi-lhes pedido que a partir de cada observação dessem uma nota ao ramo, correspondente a:

1. o conjunto de palavras indicado atrapalha a identificação do tópico;
2. o conjunto de palavras ajuda pouco na identificação do tópico;
3. o conjunto de palavras realmente ajuda a identificação do tópico;
4. o conjunto de palavras aproxima-se do ideal na identificação do tópico.

Além disso, foi avaliada a qualidade da hierarquia gerada, ou seja, avaliaram-se "como um todo" todos os ramos da hierarquia, atribuindo uma única nota para toda a hierarquia, conforme mostrado na Figura 9.

As notas foram tabuladas de acordo com a nota geral da hierarquia, as notas dadas a cada ramo em cada hierarquia e cada coleção de textos. Dessa forma, foi possível realizar testes comparativos, com base em análise de variância das notas em função de

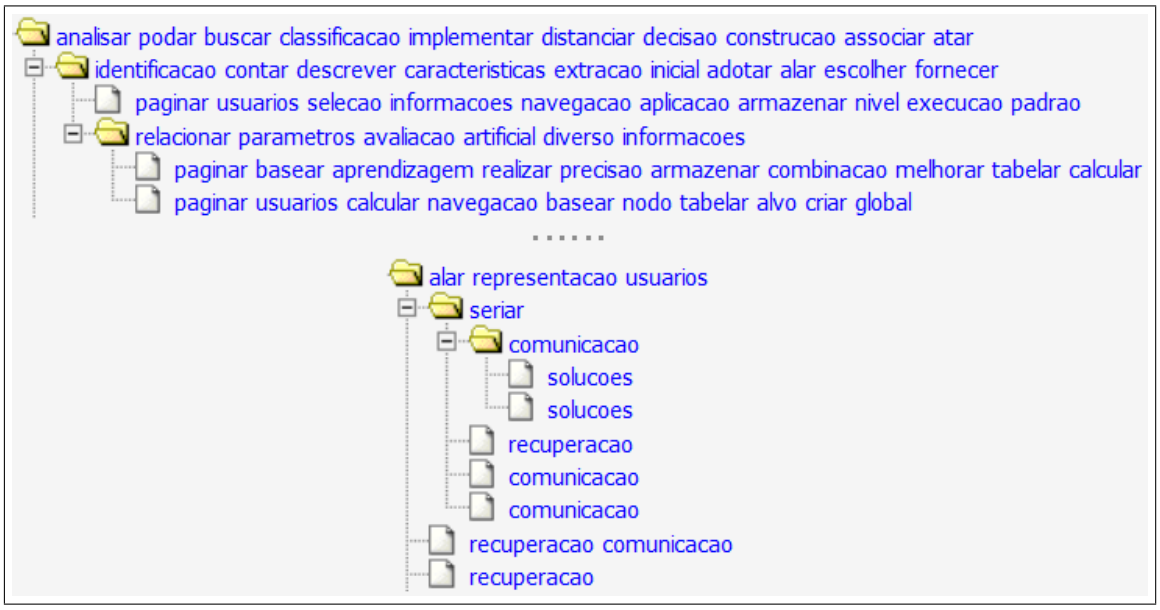

Figura 4. Ramos para a lematização utilizando a base do ENIA.

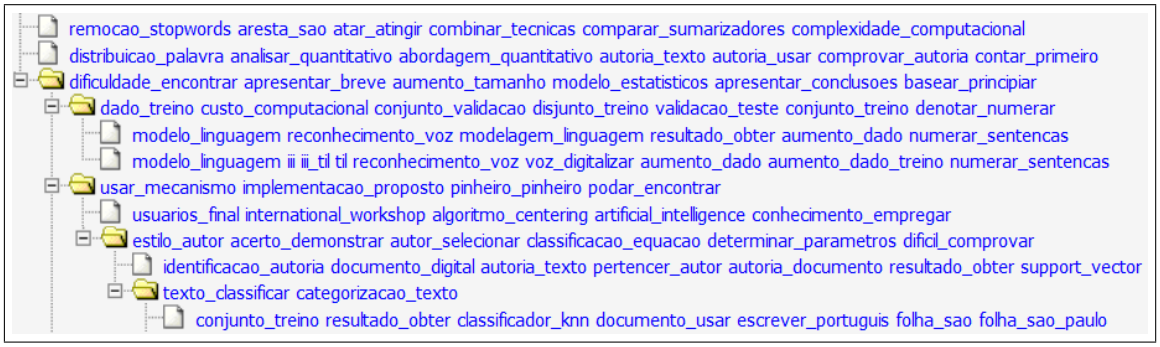

Figura 5. Ramos para a lematização utilizando a base do TIL.

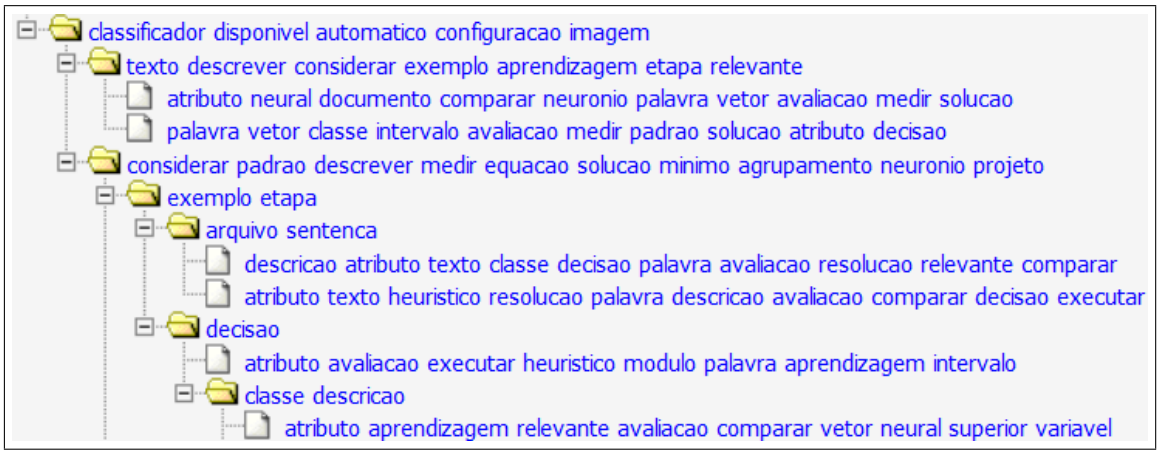

Figura 6. Ramos para a substantivação utilizando a base do ENIA.

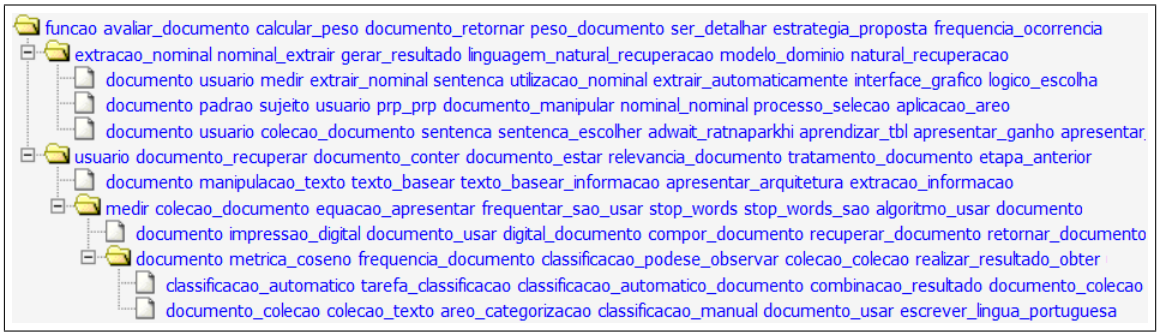

Figura 7. Ramos para a substantivação utilizando a base do TIL.

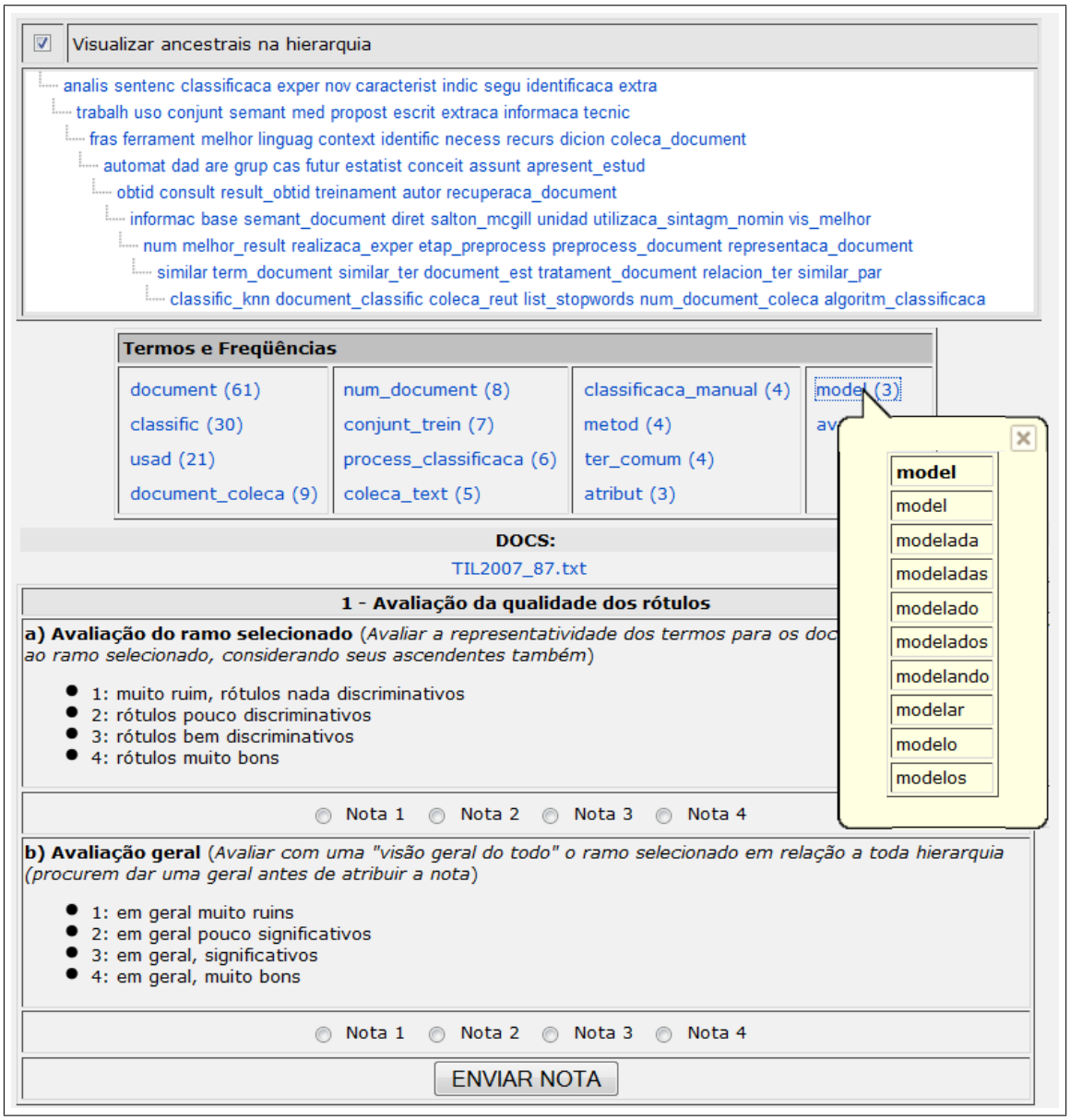

Figura 8. Avaliação do ramo selecionado.

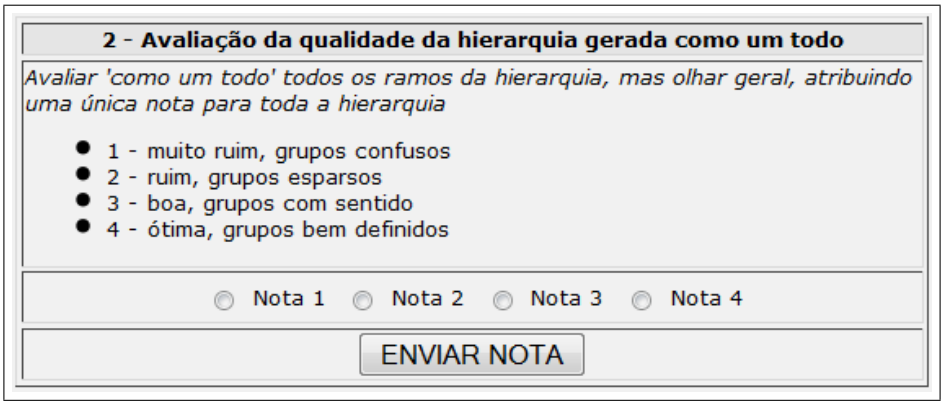

Figura 9. Avaliação geral da hierarquia em relação ao ramo selecionado.

cada um desses efeitos; em relação às diferentes formas de obtenção das hierarquias, ou seja, as técnicas utilizadas, e as técnicas em cada domínio, isto é, nas diferentes coleções de textos. As comparações múltiplas de médias foram realizadas com o uso do teste SNK [Snedecor e Cochran 1967], com 0.05 de erro; no qual os grupos de notas foram representados por letras, ou seja, nas tabelas que se seguem, letras iguais correspondem a grupos estatisticamente iguais.

Na Tabela 2 encontram-se os resultados das comparações das notas das três técnicas utilizadas considerando-se as diferentes coleções textuais como únicas. Consideraram-se as notas dadas na parte "*A*" da Figura 8, ainda como se todas as coleções correspondessem a uma única coleção e cada coleção fosse um "efeito" do modelo de análise de variância, nota-se que apesar da expectativa inicial de que a substantivação deveria ser melhor classificada, isso não ocorreu. A "radicalização" foi ligeiramente melhor classificada que as demais, sendo que as duas outras técnicas foram consideradas estatisticamente iguais.

Técnicas	Quantidade de notas dadas	Ramos	Grupos
"Radicalização"	137	2.941606	a
Substantivação	145	2.634483	b
Lematização	164	2.469512	b

Tabela 2. Agrupamento das notas nos ramos

Na Tabela 3 encontram-se os resultados das comparações das notas das técnicas para cada coleção. As notas consideradas correspondem àquelas dadas na parte *B* da Figura 8, sendo cada coleção considerada como um "efeito" do modelo de análise de variância. Este resultado permitiu avaliar (i) o quão os termos gerados representam seus respectivos textos, bem como seus ancestrais, mostrando que a técnica de "radicalização" gerou um melhor resultado geral. Já quanto a (ii) representatividade do ramo avaliado em relação a toda a hierarquia correspondente, notas dadas na Figura 9, os agrupamentos mostram que a taxonomia gerada com a técnica de "radicalização" para a base do ENIA foi considerada tão boa quanto as demais taxonomias para a base do TIL.

Para avaliar a (iii) compreensibilidade dos termos gerados ao utilizar cada técnica e em relação a (iv) preferência geral subjetiva comparando todas as taxonomias, de acordo

Técnicas	Quantidade de nós avaliados	Ramos	Grupos
"Radicalização" ENIA	66	2.954545	a
"Radicalização" TIL	71	2.929577	a
Lematização TIL	69	2.826087	a
Substantivação TIL	69	2.811594	a
Substantivação ENIA	76	2.473684	b
Lematização ENIA	95	2.210526	b

Tabela 3. Agrupamento das notas nos ramos para cada taxonomia

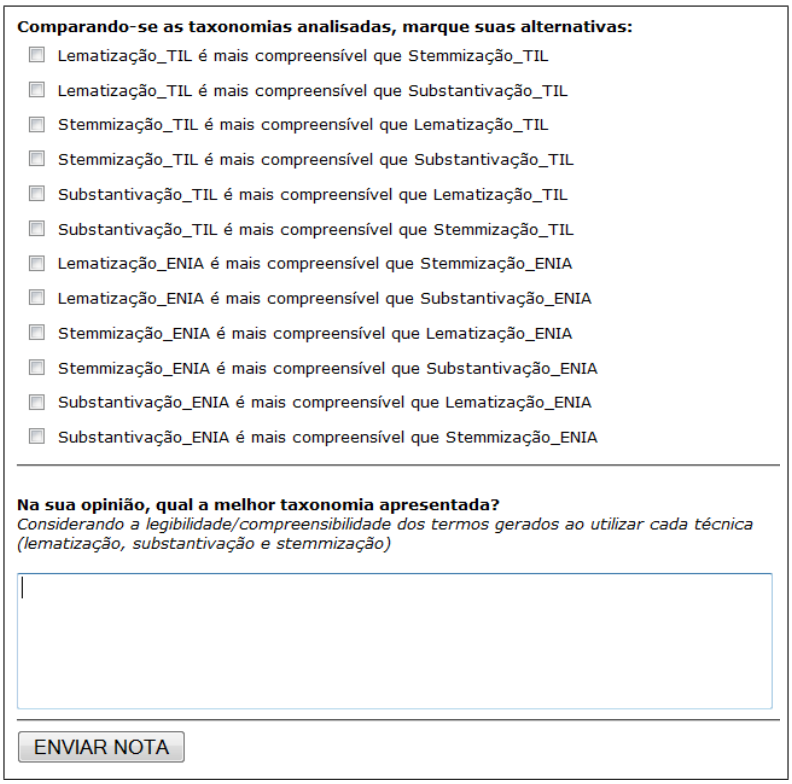

Figura 10. Avaliação geral das taxonomias.

com os especialistas, o questionário de avaliação possuía uma questão aberta, conforme mostrado na Figura 10. Apesar desta questão não ser obrigatória, quase todos os especialistas a responderam. Na avaliação como um todo, a substantivação ficou em primeiro lugar com a base do TIL e, em seguida, com a base do ENIA.

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 as diferentes técnicas de geração de termos a fim de selecionar a(s) mais adequada(s) a domínios específicos. Nos experimentos realizados, percebe-se que o uso de DF e *stoplist* da coleção foi extremamente relevante para a redução do número de termos, sendo que, com esta última, os termos mais representativos da coleção de textos foram mantidos. Percebe-se a viabilidade e necessidade do uso do teste da razão de máxima verossimilhança para remover as colocações menos discriminantes da coleção, já que utilizar apenas cortes de Luhn pode-se remover

termos representativos da coleção, por não considerar o significado semântico dos mesmos. Observa-se também que o uso de colocações melhora a compreensibilidade final dos resultados, além de manter termos compostos importantes para o domínio, auxiliando na avaliação subjetiva feita pelos especialistas do domínio.

As diferentes formas de geração de termos mostraram que, para estas bases textuais, o uso da "radicalização" e da substantivação obtiveram melhores resultados em relação à lematização. Para a substantivação, este fato foi devido a ocorrência de maior quantidade de colocações e sua menor agressividade aos termos em relação as outras técnicas, sendo eleita pelos especialistas como a de maior compreensibilidade e preferência para formação das taxonomias. A "radicalização" gerou termos mais representativos, tendo resultados melhores classificados quando avaliados os ramos em relação a toda a hierarquia.

Referências

Aranha, C. N. (2007). *Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Portuguuês: sob o Enfoque da Inteligência Computacional*. Tese de Doutorado, Departamento de Engenharia Elétrica - PUC - Rio de Janeiro.

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*, p. 370–381, Mexico.

Brill, E. (1995). Transformation-based error-driven learning of natural language: A case study in part of speech tagging. *Computational Linguistics 21*, p. 543–565.

Dias-da-Silva, B. C., Oliveira, M. F., Moraes, H. R., Hasegawa, R., Amorim, D., Paschoalino, C., e Nascimento, A. C. (2000). A construção de um thesaurus eletrônico para o português do brasil. *V Encontro para o processamento computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000)*.

Ebecken, N. F. F., Lopes, M. C. S., e Aragão, M. C. d. (2003). Mineração de textos. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 13, p. 337–364. Manole, 1 ed.

Fayyad, U. M., Piatetsky-Shapiro, G., e Smyth, P. (1996). From data mining to knowledge discovery: An overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., e Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining - Artificial Intelligence Magazine*, capítulo 1, p. 1–34. The American Association for Artificial Intelligence - AAAI Press, Menlo Park, CA, USA.

Gantz, J. F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., e Manfrediz, A. (2007). The expanding digital universe: A forecast of worldwide information growth through 2010. *External Publication of IDC (Analyse the Future) Information and Data*, p. 1–21.

Gonzalez, M. A. I., de Lima, V. L. S., e de Lima, J. V. (2006). 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.

Harman, D. (1991). How effective is suffixing? *Journal of of the American Society for Information*, 42:7–15.

Kilgarriff, A. e Yallop, C. (2000). What's in a thesaurus? *Second Conference on Language Resources and Evaluation - LREC*, p. 100–108.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Manning, C. D., Raghavan, P., e Schutze, H. (2008). Language models for information retrieval. In *An Introduction to Information Retrieval*, capítulo 12. Cambridge University Press.

Martins, C. A., Matsubara, E. T., e Monard, M. C. (2003). Um estudo de caso utilizando uma ferramenta computacional que auxilia na redução da dimensão da representação de documentos em tarefas preditivas de mineração de textos. *IV Workshop on Advances and Trends in AI, Chillan - In Proceedings of Fourth Workshop on Advances & Trends in AI for Problem Solving*, p. 21–27.

Nunes, M. (1996). The design of a lexicon for brazilian portuguese: Lessons learned and perspectives. *Proceedings of the Second Workshop on Computational Processing of Written and Spoken Portuguese*, p. 61–70.

Orengo, V. M. e Huyck, C. (2001). A stemming algorithm for the portuguese language. *Eighth International Symposium on String Processing and Information Retrieval (SPIRE)*, p. 183–193.

Porter, M. (1980). An algorithm for suffixing stripping. *Program*, 14(3):130–137.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania*, p. 491–497.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., e Paula, M. F. (2003). Mineração de dados. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 12, p. 307–335. Manole, 1 ed.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, p. 44–49.

Silic, A., Chauchat, J.-H., Basic, B. D., e Morin, A. (2007). N-grams and morphological normalization in text classification: A comparison on a croatian - english parallel corpus. *Springer Berlin / Heidelberg*, 4874:671–682. ISBN 978-3-540-77000-8.

Silla Jr, C. N. e Kaestner, C. A. A. (2002). Estudo de métodos automáticos para sumarização de textos. *Anais do Simpósio de Tecnologias de Documentos - STD, São Paulo-SP.*, 1:45–48.

Snedecor, G. e Cochran, W. (1967). *Statistical methods*. Arnes: lowa State University Press, 7 ed.

Sphinx Brasil (2008). Sphinx. http://www.sphinxbrasil.com.br/po/.