Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006) - Poster Section, Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Expansão Automática de Consultas na RI com Análise Local de Sintagmas Nominais

João Marcelo Azevedo Arcoverde[1]
Maria das Graças Volpe Nunes[1]
Wendel Scardua[2]

[1] ICMC - Universidade de São Paulo - Campus de São Carlos
Caixa Postal 668, 13560-970 - São Carlos, SP - Brasil

[2] IME - Universidade de São Paulo - Campus de São Paulo
Caixa Postal 66.281, 13083-970 - São Paulo, SP - Brasil

Abstract. Realimentação cega de relevantes constitui uma técnica amplamente utilizada para melhorar a performance de sistemas de Recuperação de Informação. Este trabalho apresenta uma técnica de análise local para expansão automática de consultas através da utilização de sintagmas nominais extraídos do conjunto pseudo-relevante. Na condução de nossos experimentos, observou-se uma melhoria nos resultados sobre alguns tópicos, assim como uma depreciação sobre outros tópicos.

1 Introdução

Em um sistema de Recuperação de Informação (SRI), a consulta é definida como o processo de elaboração da necessidade de informação do usuário. Diante da dificuldade de elaboração de consultas, o processo de selecionar documentos relevantes normalmente envolve ciclos interativos entre o usuário e o sistema, compreendendo reformulações da consulta inicial. Uma estratégia para simplificar esse processo consiste em expandir a consulta inicial com termos relacionados, na tentativa de fornecer ao sistema um contexto mais elaborado como consulta, minimizando os problemas inerentes à linguagem humana.

Na Seção 2 é apresentada uma visão geral do processo de expansão de consultas; na Seção 3 é apresentado o método de extração dos sintagmans nominais; na Seção 4 descrevemos nosso experimento com pseudo-realimentação de relevantes; na Seção 5 avaliamos o método segundo medidas para sistemas de RI; na Seção 6 concluímos o artigo.

2 Expansão de Consultas

O processo de reformulação da consulta inicial, denominado expansão de consulta, deve contemplar um critério de seleção para os novos descritores que irão compor a consulta expandida, assim como uma estratégia para recalcular o peso desses descritores juntamente com os da consulta original. Quantos novos descritores selecionar é um problema que deve ser analisado experimentalmente. O importante é que os descritores selecionados reflitam uma carga semântica ou diferencial qualitativo para o contexto onde eles foram identificados, influenciando positivamente o processo.

A expansão da consulta pode ser feita utilizando-se a própria coleção de documentos ou uma base de conhecimento externa à coleção[1], sendo que o

custo de sua obtenção e manutenção geralmente os restringe para aplicações e domínios específicos. Nesse trabalho a própria coleção de documentos é a fonte de análise para expansão.

Existem duas abordagens para a expansão de consultas: a) interativa - quando o usuário interage com o sistema fornecendo informações sobre a relevância dos documentos retornados e; b) automática - quando não há interação do usuário no processo de expansão de consulta. Na primeira abordagem diz-se que há realimentação de relevantes (*relevance feedback*). Na segunda abordagem (automática), diz-se que há pseudo-realimentação de relevantes. O escopo de documentos analisados para expandir a consulta pode ser global, quando é contemplada toda a coleção de documentos, ou local, quando é analisado apenas um subconjunto da coleção, normalmente os "n" primeiros que já foram retornados em ordem de relevância pela consulta inicial.

3 Identificação de sintagmas nominais

Esse trabalho considera apenas os SNs lexicais - aqueles cujo núcleo é um substantivo. Para sua identificação, utilizou-se o sistema de Aprendizado de Máquina de Santos [2], baseado no algoritmo de aprendizado TBL (*Transformation Based Learning*) [3].

4 Descrição do experimento

O experimento utiliza a pseudo-realimentação de relevantes para expandir automaticamente a consulta inicial do usuário sem que haja interação desse com o processo de expansão. A análise local foi realizada apenas nos vinte primeiros documentos retornados em ordem de relevância[3] pela consulta inicial. Esse valor é empírico e varia em função da própria coleção, do tópico e sua relação com o número de relevantes existente.

Foi utilizada a base de coleções disponibilizada pelo Clef 2006[4] para a atividade de RI ad-hoc monolíngue para o português do Brasil e de Portugual. A base é composta pelas edições completas dos anos de 1994 e 1995 dos jornais PÚBLICO (www.publico.pt) e Folha de São Paulo (www.folha.com.br)[5]. Foram disponibilizados 50 tópicos de pesquisa com temas variados, com suas respectivas descrições, as quais foram submetidas a um tratamento manual para formulação do conjunto de consultas iniciais, uma para cada tópico.

A coleção necessitou de três etapas distintas de pré-processamento antes de iniciar a indexação. Primeiramente o texto de cada documento foi segmentado em sentenças, estruturando-as uma por linha. Em seguida os *tokens* de cada sentença foram analisados morfologicamente. Nessa fase foram feitas as disjunções morfológicas necessárias para o pré-procesamento dos textos. Na segunda fase foi feita uma etiquetação morfosintática (*POS-tagger*) do texto utilizando-se o programa *MXPOST*[6]. Na terceira fase, utilizou-se o sistema de identificação de SNs desenvolvido por Santos [2], que tem *F-measure* de aproximadamente 87%.

De posse dos textos pré-processados inicia-se a fase de indexação que, por sua vez, também requer operações de pré-processamento comumente utilizadas, a exemplo de *case-folding*, remoção dos acentos e de *stopwords*. Uma vez que os

[3] foi utilizada a equação Okapi BM25

[4] www.clef-campaign.org

[5] compiladas pela Linguateca (www.linguateca.pt)

[6] Maximum Entropy, de Adwait Ratnaparkhi

termos estão sintaticamente etiquetados, é possível tomar algumas decisões para a redução da dimensionalidade do espaço de descritores através de indexação com controle de vocabulário.

A fim de identificar quais SNs fornecem os melhores contextos para contribuir eficientemente na reformulação da consulta, optamos por selecionar apenas aqueles candidatos que estejam próximos da ocorrência sinalizada pela consulta inicial. O objetivo da escolha é diminuir o ruído causado por descritores que estejam distantes do contexto e, portanto, provavelmente referenciam-se a um tópico diferente.

Para calcular o peso dos SNs apenas os seus núcleos foram considerados, desprezando-se seus determinantes e modificadores. O peso de um SN s em um documento d segue a Equação (1), inspirada em Gonzalez [4].

$$w_{s,d} = f_{s,d} \times \sum_{i=1}^{n} w_{t_i,d} \tag{1}$$

- $f_{s,d}$ é a frequência de ocorrência de s em d e;
- $w_{t_i,d}$ é o peso do i-ésimo termo t_i do núcleo de s em d;

Cada SN das sentenças escolhidas do documento tem o seu núcleo (normalmente substantivos) segmentado por termos unigrama. Esses termos sofrem um processo de lematização, a fim de proporcionar uma conflação natural entre eles. Os pesos dos lemas são calculados em função da sua frequência nos SNs do documento. A frequência de ocorrência do SN s em d é a soma de quantas vezes essa estrutura multi-termos ocorre no documento.

Tendo calculado o peso de cada nucleotídeo e a frequência de cada SN no documento d, o peso desse sintagma nesse documento é o produto de sua frequência pelo somatório do peso de seus nucleotídeos. Ordenam-se os SNs do pseudo-conjunto de documentos em ordem decrescente desses pesos e capturam-se os primeiros colocados para compor a consulta expandida.

5 Avaliação

Dois lotes de processamento foram gerados para sua análise comparativa: *i*) NILC01 - sem o uso de expansão de consultas e; *ii*) NILC02 - com uso de expansão de consultas. Os lotes são avaliados pelo programa *trec_eval*[7], que os processa individualmente contra uma base de julgamentos relevantes elaborada por especialistas.

Apenas 19 dos 50 tópicos (38%) apresentaram ganho de MAP[8] em relação à consulta inicial. Houve empate em apenas 1 tópico, que não retornou resultado sem expansão de consulta. No total, verificou-se que 30 tópicos apresentaram uma perda de MAP em relação à consulta inicial. Isso significa que, apesar da expansão ter retornado mais documentos relevantes na grande maioria dos tópicos, ela também retornou um número muito maior de documentos irrelevantes, pulverizando os relevantes entre eles, prejudicando o *ranking* do conjunto retornado.

O MAP do lote NILC01 é de $35,20\%$, enquanto que para o lote NILC02 é de $29,01\%$. A *precisão* e *revocação* são mapeadas no gráfico de área da Figura (1), que analisa o *trade-off* entre a precisão para cada nível de revocação consumido, em uma escala percentual, para todos os tópicos.

[7] Chris Buckley - http://trec.nist.gov/

[8] *Mean Average Precision* - expressa a média da precisão após cada documento relevante ter sido recuperado.

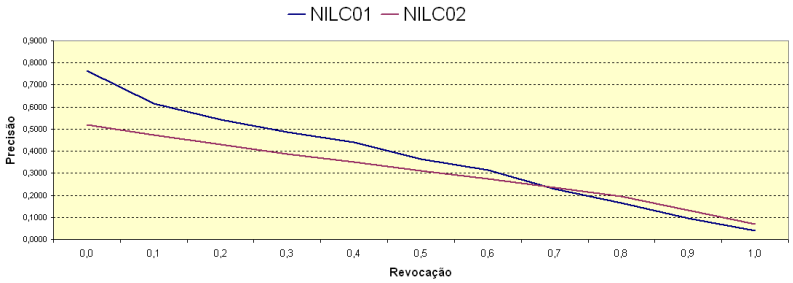

Fig. 1. Precisão a cada 10% de revocação obtida

Observou-se que a qualidade da consulta inicial é o fator que mais influencia a expansão de consutlas. Outros fatores também são responsáveis por essa influência sobre cada tópico, individualmente: *i*) a quantidade de SNs escolhidos; *ii*) a quantidade de sentenças escolhidas para extração dos SNs e; *iii*) a quantidade de documentos do conjunto pseudo-relevante.

6 Conclusão

Este trabalho investigou evidências que fundamentam a hipótese de que a aplicação de métodos que utilizam conhecimento lingüístico é viável, contribuindo com os métodos estatísticos tradicionais. Foi apresentada uma técnica de expansão de consultas com análise local sem a intervenção do usuário, sobre um modelo linguisticamente motivado por sintagmas nominais. Faz-se necessário experimentar a manipulação individual da consulta expandida para cada tópico, antes de submetê-la ao SRI, a fim de que se possa formular a melhor combinação dos parâmetros do sistema. A observação desse comportamento certamente revelará resultados mais conclusivos a respeito do experiemento.

References

1. Pizzato, L.A.S., Strube de Lima, V.L.: Evaluation of a thesaurus-based query expansion technique. In Mamede, N.J., Baptista, J., Trancoso, I., Nunes, M.V., eds.: Proceedings of the 6th Workshop on Computacional Processing of the Portuguese Language - Written and Spoken. Lecture Notes in Computer Science 2721, Universidade do Algarve-FCHS, Faro, Portugal., Springer-Verlag (2003) 251–258
2. Santos, C.N.: Aprendizado de Máquina na identificação de sintagmas nominais: o caso do português brasileiro. PhD thesis, Instituto Militar de Engenharia (IME), Rio de Janeiro (2005) Dissertação de Mestrado.
3. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics **21**(4) (1995) 543–565
4. Gonzalez, M.: Termos e Relacionamentos em Evidência na Recuperação de Informação. PhD thesis, Universidade Federal do Rio Grande do Sul (UFRGS) (2005) Tese de Doutorado.