Avaliação Empírica da Expansão de Consultas Baseada em um Thesaurus: Aplicação em um Engenho de Busca da Web

Franklin Ramalho[1,2]
Jacques Robin [2]

Resumo: Neste artigo, são avaliados empiricamente os ganhos da precisão, cobertura e medida-F, obtidos a partir do uso de várias estratégias de expansão de consultas submetidas a um engenho de busca da Web. Estas expansões foram realizadas de forma automática e baseadas em um thesaurus: WordNet. Os resultados desta avaliação mostraram que a expansão com sinônimos e/ou hiperônimos de primeiro nível melhora a medida-F e a cobertura. Eles também mostraramm que a expansão com sinônimos associados ao significado mais comum das palavras-chave das consultas melhora a medida-F, a cobertura e a precisão para os primeiros 20, 30, 40 e 50 resultados ordenados da consulta.

Abstract: In this paper, we empirically measure the precision, recall and F-measure gains, captured from usage of several query expansion strategies applied to a web search engine. This expansion was done automatically based on the WordNet thesaurus. The evaluation results showed that: expansion with keyword synonyms and/or first level hypernyms improves f-measure and recall. It also showed that expansion with most common word sense synonyms improves F-measure, recall and precision, for the 20, 30, 40 and 50 top-ranked query results.

Palavras-chave: Recuperação da Informação, Expansão de consultas, Thesaurus, Engenhos de Busca, Web.

1 Introdução

Sistemas de Recuperação de Informação (SRIs) são sistemas cuja tarefa principal é a busca por documentos relevantes que atendam à necessidade de informação do usuário, expressa através de uma consulta composta por palavras-chave. Esta busca é realizada sobre uma base de dados proprietária conhecida como coleção de documentos. Para avaliar a eficiência (qualidade dos documentos retornados em resposta ao pedido de busca do usuário) dos SRIs, existem duas métricas principais: (1) *precisão*, a proporção realmente relevante

[1]Departamento de Sistemas e Computação, UFCG, Caixa Postal 10106
franklin@dsc.ufcg.edu.br
[2]Centro de Informática, UFPE, Caixa Postal 7851
[fsr,jr]@cin.ufpe.br

dentre os documentos recuperados; e (2) *cobertura*, proporção dos documentos relevantes recuperados dentre os documentos relevantes existentes na coleção do SRI.

Com a explosão do volume de informações disponibilizadas na Web, tornou-se cada vez mais difícil encontrar a informação desejada na Web. Nesta conjectura, surgiram os engenhos de busca, SRIs que indexam em sua coleção as páginas HTML presentes na WWW e oferecem um serviço de busca da informação aos navegadores Web. Ainda hoje, estas ferramentas possuem uma grande ineficiência no processo de recuperação de informação, apresentando baixos valores para precisão e cobertura.

Uma das causas desta ineficiência é decorrente do fato de que documentos e consultas são indexados apenas por palavras-chave, independentemente do seu significado. Desta forma, durante a comparação entre os termos existentes nos documentos e os termos da consulta, documentos relevantes podem não ser recuperados apenas por não possuírem o(s) termo(s) exato(s) da consulta. Suponha, por exemplo, que um usuário submeta a consulta '*carro*' ao engenho de busca. Documentos que não contenham este termo, mas que falam sobre carros, contendo o termo '*automóvel*' ou o termo '*veículo*', não seriam recuperados, degradando a eficiência do sistema. Para resolver este problema, tem-se utilizado uma técnica conhecida como expansão de consultas, que consiste em acrescentar termos, através da disjunção, à consulta original. Estes termos adicionados devem ser semanticamente relacionados com os termos originais da consulta. A consulta 'carro' poderia ser então expandida para '*carro OR automóvel OR veículo*', recuperando, assim, mais documentos relevantes.

Neste artigo, é apresentado um conjunto de experimentos em Recuperação de Informação (RI) que avaliam quantitativamente várias estratégias de expansão automática de consultas submetidas aos engenhos de busca na Web. Estas estratégias são avaliadas através das métricas: precisão, cobertura e medida-F. Foram utilizados nos experimentos:

- A coleção de documentos do TIPSTER[3] [5]
- O thesaurus WordNet[4]
- Um clone de um engenho de busca comercial

Os resultados retornados pelo engenho foram comparados sobre a coleção TIPSTER. As consultas submetidas ao engenho são baseadas somente no uso de palavras-chave (termos) puras, com vários modos de expansão semântica destas palavras-chave. Neste processo de expansão, cada termo da consulta foi substituído por uma disjunção constituída dele mesmo mais um conjunto de palavras semanticamente relacionadas, obtidas a partir do WordNet.

Estes experimentos resultaram em quatro resultados principais:

[3] http://www.nist.gov
[4] http://www.cogsci.princeton.edu/∼wn

1. A medida-F e a cobertura foram melhoradas para todas as estratégias de expansão baseadas em sinônimos e hiperônimos[5] consideradas;

2. Limitando o conjunto de documentos recuperados àqueles pertencentes às primeiras páginas Web retornadas pelo engenho, a precisão também foi melhorada para todas as estratégias de expansão baseadas em sinônimos e hiperônimos;

3. A estratégia que alcançou melhores resultados foi a que expandiu cada palavra-chave da consulta apenas pelos sinônimos de seu significado mais comum[6];

4. Muito embora os resultados das consultas expandidas automaticamente tenham sido melhores que os obtidos das consultas padrões, sem expansão, a medida-F continua ainda assim, em termos de resultados absolutos, sendo muito baixa.

O complemento deste artigo está organizado como segue. Primeiro, é apresentada uma motivação para o trabalho, na seção 2. A seguir, a seção 3 detalha e discute os experimentos realizados, destacando os recursos e métricas utilizadas. Na seção 4 são apresentados os trabalhos correlatos e, finalmente, na seção 5, são feitas conclusões e menções sobre trabalhos futuros.

2 Motivação: Medindo o Impacto de Conhecimento Lingüístico em Engenhos de Busca da Web

Engenhos de busca têm se transformado numa ferramenta bastante difundida na Web, e assim, em nossas vidas. Usuários fazem uso deles em suas casas e em seus escritórios, com o objetivo de acessarem informações contidas na rede internet global, bem como nas redes internas de pequenas ou grandes organizações. Estas ferramentas são úteis ainda em um amplo conjunto de tarefas como *e-business*, *e-shopping*, *e-entertainment* e *e-education*. Eles são responsáveis pelo movimento de mais de um bilhão de dólares e possuem uma das maiores bases de usuários da indústria de software.

Todavia, a eficiência destes engenhos ainda é precária. Este é um problema que atinge a maioria dos SRIs que trabalham com volumosas bases de dados e que possuem recuperação baseada tão somente na comparação de palavras-chave contidas nesta base com àquelas incorporadas na consulta submetida pelo usuário. Assim, eles ainda retornam um expressivo

[5]Palavras relacionadas semanticamente através da relação *é-um-tipo-de*. Por exemplo, *bus* é um tipo de *car*, logo, *car* é um hiperônimo de *bus*.

[6]Uma palavra pode possuir mais que um significado (sentido), constituindo-se assim em uma palavra ambígua. O significado mais freqüente de uma palavra é aquele mais comumente utilizado por falantes da língua a qual ela pertence. É importante destacar que, para cada sentido da palavra, existe um diferente conjunto de sinônimos associado.

número de documentos irrelevantes (baixa precisão) e pecam também por deixarem de recuperar muitos dos documentos relevantes que estão contidos em sua base de dados (baixa cobertura). Isto se deve, principalmente, ao fato deles se limitarem a realizar apenas comparações "superficiais" de unidades lingüísticas, tratando apenas variações morfológicas e sintáticas das palavras contidas nos documentos e nas consultas do usuário, deixando de lado comparações "profundas" como o tratamento cognitivo e pragmático-semântico, que identificam a intenção comunicativa do autor do documento e da consulta.

O processamento de cadeias de palavras para obter a intenção comunicativa é uma tarefa central de Processamento de Linguagem Natural (PLN) e embutir PLN dentro de engenhos de busca da Web parece ser a chave para que se alcance uma melhor eficiência nestas ferramentas. Enquanto na década de 80 o uso de PLN era ainda essencialmente confinado a um nicho de aplicações com domínios de discurso bem restritos, o recente desenvolvimento, em larga escala, de recursos de conhecimento lingüístico (simbólicos ou probabilísticos) tem tornado o PLN mais robusto, de forma que seu uso passou a ser uma realidade para aplicações independentes de domínio, como engenhos de busca[7].

Muito embora existam inúmeros recursos lingüísticos e técnicas de PLN, eles podem ser combinados de diferentes maneiras no intuito de melhorar a eficiência de SRIs. Isto, porém, eleva o custo computacional de tais aplicações. Surge assim, um problema novo que é determinar quais são exatamente as técnicas de PLN que, na prática, realmente ajudam no processo de recuperação de informação no mundo da Web. Assim, antes de decidir investir no desenvolvimento e incorporação de PLN em engenhos de busca, seus desenvolvedores precisam ter uma precisa idéia de quais ganhos em eficiência eles podem esperar como retorno de cada aplicação de PLN incorporada ao engenho.

Enquanto existe um amplo conjunto de trabalhos prévios sobre a avaliação de técnicas de PLN aplicadas à RI, são poucas as publicações que relatam experimentos cujo objetivo seja avaliar *individualmente* estas técnicas, e este conjunto fica ainda mais reduzido quando se trata de avaliá-las em um cenário de busca da informação na Web. Muitos experimentos anteriores ao nosso, medem o impacto *global* de *combinações* de técnicas de PLN (p.e., [6]), deixando ao leitor, a tarefa de adivinhar qual a *contribuição* individual de cada técnica sobre os resultados relatados. Os poucos experimentos que medem tais contribuições individuais são realizados com consultas que são mais sofisticadas e longas que aquelas que são usualmente submetidas aos engenhos de busca da Web. Por exemplo, [8] mede o impacto do processamento morfológico em consultas, cujo tamanho médio é de 9.65 palavras, enquanto [15] mede o impacto da expansão de consultas através de sinônimos, em três experimentos, para consultas com tamanho médio de 52.5, 29.2 e 11.0 palavras, respectivamente. Em contraste, nenhuma das 1000 consultas mais freqüentemente submetidas ao engenho de busca

[7] Desde que engenhos de busca da Web mantêm de forma confidencial quais técnicas de RI eles usam, torna-se impossível também, saber, precisamente, quais técnicas de PLN eles utilizam. Todavia, suas baixas precisão e cobertura nos sugerem que investigações e melhoramentos ainda precisam ser realizados nesta direção.

comercial, que utilizamos em nossos experimentos, possuía mais que cinco palavras: 60.7% possuíam apenas 1 palavra, 33.1% eram formadas por apenas duas palavras, 5% continham três palavras e 0.2% das consultas eram compostas por cinco palavras.

Neste artigo, é dada continuação ao trabalho de Voorhees [15], que concluiu que a expansão de consultas com sinônimos era ineficiente quando aplicada às consultas longas. Nós concluimos que, para consultas compostas por uma pequena quantidade de palavras (formadas por até quatro palavras), como as consultas tipicamente submetidas aos engenhos de busca da Web, tal expansão melhora significativamente não apenas a cobertura, mas também a precisão para os documentos recuperados e apresentados nas primeiras páginas de resultado da busca.

3 Os Experimentos

3.1 Coleção de Testes

Para os experimentos, foi utilizado o disco 2 da coleção de testes TIPSTER, que foi produzida como um resultado do TREC (Text REtrieval Conference) e do Workshop TIPSTER [5]. Este disco contém:

- 231000 documentos da língua inglesa, totalizando 2Gbytes de texto que abrangem uma ampla variedade de tópicos e estilos;
- 50 consultas prontas da língua inglesa;
- Julgamentos de relevância[8] para aproximadamente 78000 dos 11,55 milhões pares de possíveis julgamentos consulta-documento da coleção.

O subconjunto de pares para os quais existe um julgamento de relevância é chamado de *pool* e consiste da união dos 1000 documentos recuperados e melhor classificados[9], para cada consulta pronta, pelos 20 SRIs que participaram do TREC-1[10]. A cobertura de cada sistema é então aproximada, uma vez que qualquer documento que esteja fora do *pool* e que, portanto, não possua julgamento de relevância, é considerado como irrelevante. Esta técnica, conhecida como *pooling* é uma tentativa de estimar a cobertura para coleções de documentos que são muito grandes e que, por isso, torna impraticável a tarefa de avaliar manualmente a relevância de cada documento para cada consulta da coleção.

[8] Caso o documento seja relevante para a consulta, seu julgamento de relevância assume valor *1*, caso contrário, assume valor *0*.
[9] De acordo com ordem de relevância.
[10] Isto é, um documento é parte do *pool*, para uma dada consulta, se ele aparece entre os 1000 primeiros documentos recuperados por pelo menos um SRI.

A coleção TIPSTER caracterizou-se como ideal para nossa proposta devido às seguintes razões[11]: (1) suas consultas e julgamentos de relevância permitiram automatizar o processo de avaliação; (2) possui duas características cruciais no processo de avaliação de engenhos de busca da Web, que são a grande quantidade de documentos e sua ampla diversidade de domínio; (3) tem se tornado uma referência utilizada por muitos pesquisadores, o que permite comparar nossos resultados com outros trabalhos correlatos; e (4) suas consultas são estruturadas em muitos campos que descrevem a necessidade de informação dos usuários em vários níveis de detalhamento. Um exemplo de consulta é dado na Fig. 1. Entre os múltiplos campos pertencentes às consultas, foi escolhido o campo *title*, uma vez que o seu tamanho o torna mais próximo de consultas tipicamente submetidas aos engenhos de busca da Web.

3.2 Métricas de Avaliação

Métricas para medir a eficiência de um SRI são medidas usadas para calcular a habilidade destes sistemas em recuperar os documentos relevantes existentes em sua base de dados, e apenas eles, para uma dada necessidade de informação do usuário. Em nossos experimentos, foi medida a eficiência para cada estratégia de expansão adotada, seguindo as seguintes métricas:

1. Precisão no *pool* (*P*)
2. Cobertura no *pool* (*C*)
3. Medida-F no *pool* (*F*)
4. Precisão no *pool* limitada aos L primeiros documentos recuperados, com $L = 50, 40, 30, 20$ e 10 (P_L)

As fórmulas que definem estas métricas são apresentadas pelas equações 1, 2, 3 e 4. Todas as métricas são referenciadas como "no *pool*" porque: (1) para todas elas, todos os documentos que não pertencem ao *pool* foram considerados como irrelevantes; e (2) o *pool* considerado aqui é o que foi construído no TIPSTER pelos sistemas participantes do TREC-1, ou seja, o *pool* utilizado não foi especialmente construído para os experimentos apresentados neste artigo. A vantagem do uso deste *pool* é que com ele, pôde-se realizar a avaliação de forma amplamente automática. Por outro lado, sua desvantagem é o risco de acontecer uma subestimação da eficiência das estratégias avaliadas, no caso delas recuperarem documentos relevantes que não tenham sido recuperados por nenhum dos participantes do TREC-1 e que

[11] Apenas depois que os experimentos foram realizados, que a coleção de documentos Web foi disponibilizada pelo TREC.

<top>
<head> Tipster Topic Description
<num> Number: 086
<dom> Domain: Finance
<title> Topic: **Bank Failures**
<desc> Description: Document will identify recent bank failures within the United States, specifying name, location, assets, and federal or state authorities taking action.
<smry> Summary: Document will identify recent closings by the Federal Deposit Insurance Corporation of a failed commercial bank chartered within the U.S.
<narr> Narrative: A relevant document, in addition to giving specific identifying information on a failed bank (or banks), must be clear that the bank was chartered within the United State and that it is a bank, as opposed to another financial institution such as a savings and loan or credit union, and must specify the actions taken or being taken by the public authorities.
<con> Concept(s):
1. Bank failure
2. Federal Deposit Insurance Corporation, FDIC, Comptroller of the Currency, state banking authorities, state bank regulators
<fac> Factor(s):
<nat> Nationality: U.S.
<time> Time: current
</fac>
<def> Definition(s):
</top>

Figura 1. Exemplo de uma consulta da coleção TIPSTER

assim, não estejam no *pool*. Para a proposta de comparar várias estratégias de expansão de consultas, isto foram feitas, assim, estimativas conservativas, possivelmente subestimando os ganhos obtidos com a expansão das consultas.

$$P = \frac{RR}{RR + IR} \tag{1}$$

$$P_L = \frac{RR_L^K}{L} \tag{2}$$

$$C = \frac{RR}{RR + RN} \tag{3}$$

$$F = \frac{2 * P * C}{P + C} \tag{4}$$

Onde:

- RR = número de documentos relevantes recuperados;
- IR = número de documentos irrelevantes recuperados;
- RN = número de documentos relevantes não recuperados;
- L = limite do número de documentos recuperados considerados;
- K = número de documentos realmente recuperados[12]
- RR_L^K = número de documentos relevantes dentre os K primeiros recuperados, com $1 \leq K \leq L$

Desde que a cobertura pode ser facilmente melhorada com o decréscimo da precisão, e vice-versa, para avaliar os ganhos da eficiência com o uso da expansão da consulta, foi utilizada, também, a medida-F, que simboliza o comportamento harmônico das duas primeiras [11]. A medida-F cresce não apenas quando a precisão cresce com um valor fixo para a cobertura ou vice-versa, mas ela cresce também, quando os valores das outras duas se aproximam. Esta é uma característica importante e desejada para uma métrica de avaliação de SRIs, pois ela avalia tanto o comportamento da precisão, como o da cobertura, e mais ainda, o comprometimento de uma dada estratégia para com estas duas medidas. Além disto, para totalizar a precisão no *pool*, foi computada também a precisão para os *N* primeiros documentos recuperados pelo engenho. Esta é a métrica-chave para avaliar engenhos de busca da Web, desde que: (1) ela considera o *ranking*[13] dos documentos recuperados; e (2) a maioria dos usuários está interessada apenas em visitar os 10, 20 ou 30 primeiros documentos retornados pelos engenhos. Foi medida a precisão para os 10, 20, 30, 40 e 50 primeiros retornados. Note que estes resultados, limitados pelo *ranking* dos documentos, diferem daqueles que não consideram nenhum limiar: os primeiros (no caso, a precisão limitada) tornam a variação da precisão média entre duas estratégias igual à variação da cobertura média entre elas[14], fazendo com que estas duas métricas tornem-se redundantes, ao invés de complementares.

[12] Isto porque, em alguns casos, os engenhos podem recuperar uma quantidade de documentos menor que o limite L estabelecido. Por exemplo, pode-se estar calculando a precisão para os 50 (L) primeiros recuperados, enquanto que o engenho só recuperou 20 (K).

[13] Ordem em que as respostas são apresentadas ao usuário. Usualmente, este *ranking* é realizado em ordem decrescente de relevância, de forma que o primeiro documento apresentado (com *ranking* 1) seja mais relevante do que o segundo (*ranking* 2), e assim, sucessivamente.

[14] Isto é formalmente provado em [8].

3.3 Engenho de Busca

O clone do engenho de busca comercial que foi utilizado nos experimentos, indexa, recupera e ordena as respostas baseando-se apenas no seu conteúdo, sem fazer uso de qualquer forma de filtragem colaborativa[15]. Além do mais, ele aproxima o conteúdo do documento e da consulta através das formas de palavra que eles contêm, sem utilizar qualquer técnica de PLN[16]. Ele indexa documentos através de arquivos invertidos[17] que contêm para cada forma de palavra *P*: (1) sua freqüência no arquivo invertido; (2) a lista de URLs de todos documentos D_i nos quais *P* ocorre; (3) a lista de freqüências de *P* em cada D_i; (4) a lista de posições de *P* em cada D_i e (5) a lista de marcadores HTML associados a *P* em cada D_i. O engenho recupera documentos através de buscas *booleanas* realizadas no arquivo invertido. A ordenação das respostas obtidas é realizada heuristicamente considerando: a medida *TFIDF* (*Term Frequency Inverted Document Frequency*) [2], a posição do termo na consulta e no documento, e a proximidade dos termos da consulta dentro do documento. Desde que todas estas técnicas podem ser consideradas como padrões em RI, este engenho se constitui em um estudo de caso apropriado para ser utilizado na avaliação dos efeitos do recurso de conhecimento lingüístico incorporado em engenhos de busca da Web.

3.4 Recurso Lingüístico

O recurso lingüístico utilizado para expansão da consulta foi o Princeton WordNet 1.6, que é um thesaurus[18] gratuito da língua inglesa. Esta versão contém 95600 formas de palavra indexadas hierarquicamente por 70100 sentidos de palavra, que se referenciam através de seis relações semânticas: antonímia, hiperonímia (relação *é-um-tipo-de*), três diferentes relações de meronímia (relação *parte-de*) e sinonímia. Neste trabalho, a avaliação foi limitada a duas relações: sinonímia e hiperonímia (em seu primeiro nível de profundidade, i.e., com profundidade igual a 1[19]), ambas entre substantivos. Este foco nos substantivos é motivado pelo fato de todas as 100 consultas mais freqüentemente submetidas ao engenho de busca comercial, usado em nossos experimentos, serem compostas apenas por substantivos. O WordNet contém aproximadamente 57000 substantivos organizados em aproximadamente 48800 significados. Algumas das relações semânticas do WordNet são ilustradas na Fig. 2.

O WordNet classifica seus sinônimos (e outras relações) de acordo com os *sentidos* de

[15] Ao contrário do Google, http://www.google.com
[16] Ao contrário do Ask Jeeves, http://www.askjeeves.com
[17] Um índice de texto composto de um vocabulário e uma lista de ocorrências associada a cada palavra pertencente ao vocabulário.
[18] Banco de dados léxico que relaciona semanticamente seus termos, de acordo com uma estrutura conceitual.
[19] Uma palavra pode possuir vários hiperônimos em diferentes níveis de profundidade, de acordo com a estrutura hierárquica do thesaurus. Por exemplo, *car* é um tipo de *motor vehicle* (profundidade 1), assim como pode ser considerado um tipo de *transport* (profundidade 4), ou ainda um tipo de *physical object* (profundidade 7), e assim sucessivamente até atingir o conceito de abstração mais alta na estrutura do thesaurus.

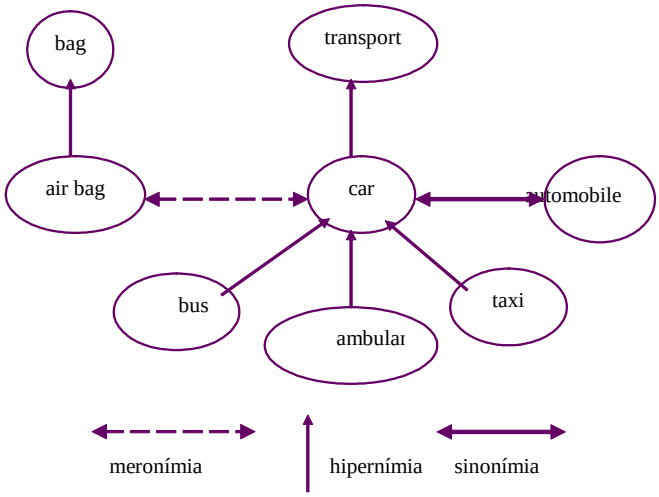

Figura 2. Exemplos de relacionamentos entre termos do thesaurus WordNet

cada palavra. Isto devido à ambigüidade dos termos. Assim, uma palavra que possui mais que um significado, possui também mais que um conjunto de sinônimos, sendo cada conjunto associado a um específico sentido para aquela palavra. Por exemplo, na Fig. 2, a palavra *car* possui vários sentidos, dentre eles: (1) *car, auto, automobile – motor vehicle, usually propelled by an internal combustion engine*; e (2) *car, elevator car – where passengers ride up and down*.

Mais precisamente, foram avaliadas as seguintes estratégias de expansão:

1. Nenhuma expansão (NE);
2. Expansão através de sinônimos para todos os sentidos de cada substantivo pertencente à consulta (TS);
3. Expansão através de sinônimos para os 3 mais freqüentes sentidos de cada substantivo pertencente à consulta (3S);
4. Expansão através de sinônimos para os 2 mais freqüentes sentidos de cada substantivo pertencente à consulta (2S);
5. Expansão através de sinônimos para o mais freqüente sentido de cada substantivo pertencente à consulta (1S);

6. Expansão através de sinônimos e hiperônimos (de profundidade 1) para os 2 mais freqüentes sentidos de cada substantivo pertencente à consulta (2SH);
7. Expansão através de sinônimos e hiperônimos (de profundidade 1) para o mais freqüente sentido de cada substantivo pertencente à consulta (1SH);

Em todas as estratégias, o processo de expansão é caracterizado pela substituição de cada palavra-chave (que seja um substantivo) pertencente à consulta original por uma disjunção *booleana* da própria palavra-chave e seus sinônimos e/ou hiperônimos fornecidos pelo WordNet. Note que, caso a consulta seja composta por mais de um termo, sua expansão resultará em uma conjunção de disjunções.

3.5 ExpanSyns

Com o intuito de automatizar o processo de expansão de consultas, assim como toda a avaliação empírica, foi desenvolvido o ExpanSyns, um sistema para suporte aos experimentos empíricos que medem os ganhos de várias estratégias de busca inteligente baseada no uso de recursos lingüísticos. Conforme ilustrado na Fig. 3, sua arquitetura consiste de três camadas: a camada de dados, a camada lógica e a camada de interface.

A camada de dados contém três tipos de fontes de dados: (1) a coleção de testes, onde estão presentes o conjunto de documentos, as consultas prontas e os julgamentos de relevância dos primeiros para as consultas; (2) a base de conhecimento utilizada para armazenar informações do recurso lingüístico a ser avaliado, no caso, o thesaurus WordNet; (3) o banco de dados relacional onde estão armazenados os dados sobre os experimentos.

A camada lógica consiste de seis componentes, um para cada etapa de processamento do experimento. A primeira etapa envolve o acesso às consultas e o seu pré-processamento a partir do uso do recurso lingüístico, isto é, engloba o processo de expansão de consultas a partir da coleção de testes e do WordNet. A segunda etapa envolve a tradução das consultas para linguagem de consulta do engenho de busca para que elas possam ser submetidas ao mesmo. A terceira etapa é a responsável por extrair, das páginas de resultados do engenho de busca, o conjunto de links para os documentos recuperados. Nesta etapa, é realizada uma filtragem do conteúdo das páginas de resultados, uma vez que são desconsideras informações como resumos do conteúdo dos links, *banners* de propaganda comercial e outros links que não sejam aqueles para os documentos recuperados pelo engenho. A quarta etapa realiza a avaliação de relevância dos documentos recuperados, que é verificada a partir dos dados contidos na coleção de testes. A quinta etapa envolve os cálculos dos valores das métricas utilizadas para o experimento. Finalmente, a sexta e última etapa é a responsável por armazenar no banco de dados, os valores obtidos pelo experimento.

A camada de interface é a responsável por permitir que um usuário controlador do

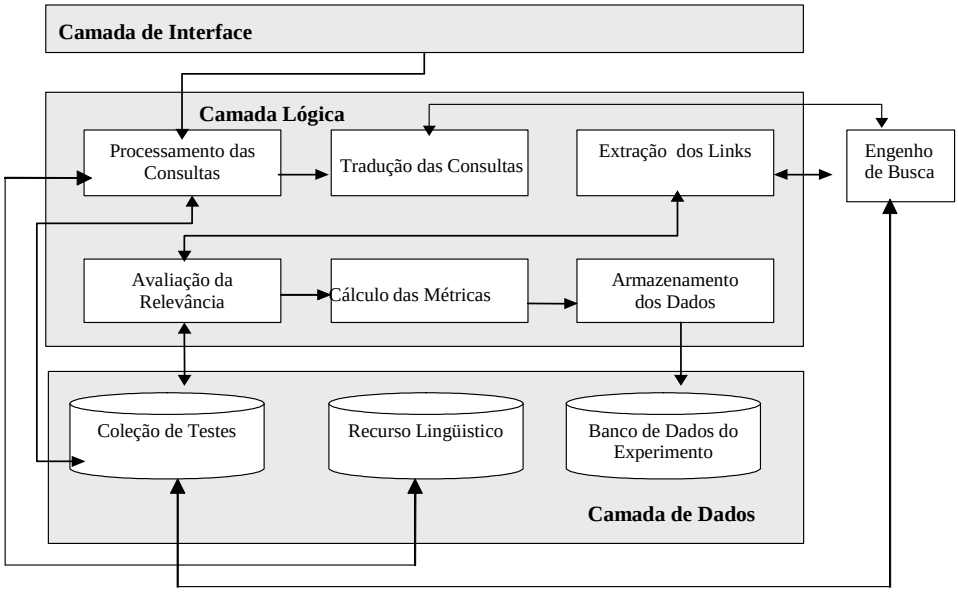

Figura 3. Arquitetura do ExpanSyns

experimento possa configurá-lo e iniciá-lo, conforme alguns parâmetros, tais como o número de consultas que serão submetidas ao engenho ou quais as estratégias de expansão que serão avaliadas.

3.6 Resultados e Discussão

Os resultados dos experimentos são sumarizados nas tabelas 1 e 2. Ambas têm uma célula para cada par métrica-estratégia. Em cada célula, a primeira linha contém o valor médio da métrica aplicada às consultas do experimento, enquanto as segunda e terceira linhas contêm as variações relativas e absolutas com relação ao caso base de nenhuma expansão (NE), respectivamente.

Um resultado chave mostrado na tabela 1 é que todas as estratégias de expansão consideradas melhoram a eficiência total (medida-F), uma vez que elas alcançam um maior aumento para o valor da cobertura do que um decréscimo para o valor da precisão (em termos relativos). Assim, a expansão de consultas através da sinonímia e/ou hiperonímia funciona. Um outro resultado chave na tabela 1 é que a estratégia mais efetiva para melhorar a eficiên-

cia total é limitar a expansão de *sinônimos* àqueles associados com o mais freqüente sentido de cada substantivo pertencente à consulta, elevando a medida-F em +37.28% sobre o caso base de nenhuma expansão.

Tabela 1. Precisão, cobertura e medida-F sem limites

Agents	Use case	Decomposition in actions
Librarian	Loan	a1: To identify an adherent a2: Count adherent (To check right loan for member) a3: To seek an exemplary a4: To handle an exemplary (to validate the output of an exemplary) a5: To handle an adherent (to indicate the loan by adherent)
	Reservation	a1: To identify an adherent a2: Count adherent (To check right loan for member) a3: To seek an exemplary a6: To reserve an exemplary
	Restitution	a1: To identify an adherent a3: To seek an exemplary a4: To handle an exemplary (to validate the input of exemplary) a5: To handle an adherent (to indicate the return of an exemplary)
	Loan if counts blocked	a1: To identify an adherent a2: Count adherent (To check right loan for member)
	Identification member	a1: To identify an adherent
adherents responsible	New adherent	a2: Count adherent (to Add adherent)
	Litigation	a1: To identify an adherent a2: Count adherent (Blocked count adherent) a7: To inform an adherent
Exemplaries responsible	Exemplary addition	a3: To seek an exemplary a8: To add an exemplary
	Exemplary deletion	a3: To seek an exemplary a9: To remove the damaged exemplaries

Um terceiro importante resultado é que a cobertura pôde ser aumentada em até 72.4%, com relação ao caso base, quando da aplicação da expansão através de *ambos*, sinônimos e hiperônimos do 1.^{o} nível de profundidade para os dois primeiros sentidos (2SH) de cada substantivo da consulta.

O último resultado notório da tabela 1 é que em termos absolutos, mesmo a melhor

estratégia de expansão da consulta resulta em apenas 2.51% de melhoramento na medida-F, alcançando 9.3% sobre 6.7% para o melhor caso. A melhor estratégia recupera ainda 11% de todos os documentos relevantes juntamente com 77.5% dos irrelevantes deles. Isto nos leva a concluir que, mesmo a expansão automática de consultas através de sinônimos, quando realizada de forma isolada, está longe de tornar os engenhos de busca próximos de SRIs verdadeiramente eficientes (com valores próximos de *1* para a precisão e a cobertura).

Tabela 2. Precisão limitada

	Top 10	Top 20	Top 30	Top 40	Top 50
NE	0.300	0.267	0.228	0.206	0.184
TS	0.241 -24.5% -5.9%	0.218 -22.5% -4.9%	0.211 -8.06% -1.7%	0.203 -1.48% -0.3%	0.187 +1.63% +0.3%
3S	0.237 -26.6% -6.3%	0.230 -16.1% -3.7%	0.223 -2.24% -0.5%	0.219 +6.31% +1.3%	0.203 +10.3% +1.9%
2S	0.263 -12.3% -3.7%	0.239 -10.5% -2.8%	0.237 +3.95% +0.9%	0.227 +10.2% +2.1%	0.232 +26.1% +4.8%
1S	0.296 -1.35% -0.4%	0.270 +1.12% +0.3%	0.256 +12.3% +2.8%	0.242 +17.5% +3.6%	**0.252** **+37.0%** **+6.8%**
2SH	0.196 -53.1% -10.4%	0.193 -38.3% -7.4%	0.196 -16.3% -3.2%	0.209 +1.46% +0.3%	0.170 -8.24% -1.4%
1SH	0.222 -35.1% -7.8%	0.239 -11.7% -2.8%	0.222 -2.7% -0.6%	0.206 +0.0% +0.0%	0.188 +2.17% +0.4%

Talvez o resultado mais animador sobre os experimentos realizados esteja na tabela 2: inesperadamente, a expansão de consultas baseada em sinônimos e/ou hiperônimos melhora a *precisão para os primeiros documentos recuperados*. Este ganho na precisão provavelmente é fruto de uma sinergia entre um aumento na cobertura alcançada e o uso de uma heurística de *ranking* eficiente. Como a cobertura é melhorada, mais documentos relevantes são recuperados. Com um algoritmo de *ranking* efetivo, alguns destes documentos superam alguns dos documentos irrelevantes que se encontram na lista dos primeiros retornados para os usuários, melhorando, desta forma, a *precisão limitada* para esta lista. As estratégias que resultam nestes ganhos de precisão limitada são sublinhadas e destacadas em fundo cinza na tabela 2. Outro ponto interessante é que mesmo para o caso do uso da precisão limitada, assim como no caso de uso de métricas não limitadas, a melhor estratégia é a expansão com os sinônimos

do sentido mais comum. Esta estratégia melhora a precisão limitada para os 20, 30, 40 e 50 primeiros documentos recuperados, resultando em um alto ganho de precisão de 37% para os primeiros 50 documentos retornados. Estes resultados confirmam, empiricamente, a qualidade do *ranking* dos sentidos das palavras empregado no WordNet.

4 Trabalhos Relacionados

Vários trabalhos prévios [1, 4, 6, 10, 12, 13, 14, 15] têm avaliado o uso do WordNet como recurso de conhecimento lingüístico utilizado para melhorar a eficiência de SRIs. Outros ainda, por exemplo [7], têm avaliado o uso de outros thesaurus para o mesmo propósito.

Aqui, tem-se espaço apenas para comparar nosso trabalho com os três [1, 12, 15] mais diretamente relacionados, uma vez que eles também:

- Avaliam o uso do conjunto de sinônimos e de outras relações semânticas do WordNet como *únicas* fontes de conhecimento lingüístico para o melhoramento da eficiência de um SRI;
- Avaliam o uso de conjuntos de sinônimos apenas para expandir vetores de termos durante o processamento da consulta do usuário[20];
- Executam comparação consulta-documento diretamente no vetor de representação destes;

Assim como o nosso trabalho, [12, 15] avaliam a expansão de consultas com conjuntos de sinônimos. Todavia, eles diferem de nós em alguns aspectos importantes. Primeiramente, eles medem o impacto de usar-se conjuntamente a expansão de termos da consulta através de sinônimos e a eliminação de ambigüidade dos substantivos da consulta, enquanto nosso trabalho mede o impacto *individual* da expansão dos substantivos da consulta via sinonímia[21]. Além do mais, estes experimentos não retratam um cenário de um engenho de busca típico da Web por duas razões: (1) a eliminação de ambigüidade é realizada manualmente por especialistas humanos; e (2) ora os documentos (no caso de [12]), ora as consultas (no caso de [15]), possuem tamanhos que não retratam a realidade da Web, não se caracterizando assim, como amostras representativas de documentos e consultas típicas da Web.

[20] Alguns trabalhos utilizam-se das relações semânticas contidas no WordNet durante o processo de indexação de documentos, também.

[21] Todavia, alguém poderia considerar que limitar a expansão para sinônimos e hiperônimos pertencentes a um, dois ou três sentidos mais freqüentes listados pelo WordNet, constitui-se em alguma forma "grosseira" de tratamento de ambigüidade.

Trabalhando com recuperação de imagens baseada nos nomes das mesmas, [12] elaborou uma coleção de imagens e consultas para realizar seus experimentos. O tamanho médio dos documentos, i.e., dos nomes das imagens era de 5.89 palavras, que é um número bem abaixo do número médio de palavras contidas nas páginas HTML encontradas na Web.

[15] utilizou também a coleção de testes TIPSTER[22] e fez uso de *diferentes campos* das consultas[23] desta coleção. Ela realizou três diferentes simulações e, em todas, utilizou consultas com tamanho muito grande, não retratando assim, a realidade das consultas tipicamente submetidas aos engenhos de busca da Web. Na primeira simulação, foram utilizados todos os campos e sub-campos das consultas do TIPSTER, resultando em consultas com tamanho médio de 52.5 palavras. Na segunda, foram selecionados apenas os campos *summary* e *concept*, resultando em consultas com tamanho médio de 29.2 palavras. E, até na terceira simulação, onde foi escolhido apenas o campo *summary*, a consulta apresentou um tamanho médio grande, de 11 palavras. Em contraste, nós utilizamos o campo ***title***, o mais curto da coleção, e, ainda filtramos aquelas consultas que possuíam um tamanho maior que quatro palavras para este campo.

Dos trabalhos prévios, o mais similar com o nosso é o [1]. Eles mediram a contribuição individual da expansão dos substantivos e *verbos* pertencentes às consultas, através de várias relações semânticas do WordNet. Eles também não realizaram tratamento de ambigüidade. Assim como [12], eles se concentraram na recuperação de imagens, mas ao invés de realizar casamentos com os nomes das imagens, eles realizaram comparações baseadas em meta-dados associados à cada imagem. Devido a esta ênfase, [1] também executou seus experimentos sobre uma base de documentos especialmente elaborada para seu trabalho. Seus documentos, imagens e meta-dados, além de serem de tamanho pequeno (em comparação ao tamanho dos documentos encontrados na Web), são semanticamente estruturados, diferenciando-se assim de páginas HTML, que não possuem estruturação semântica imposta ao seu conteúdo. Por isto, o cenário de [1] também se revelou como não característico do cenário encontrado em um típico engenho de busca da Web.

Em suma, apesar da idéia da expansão de consultas baseada no WordNet para melhorar a eficiência de SRIs não ser nova, nós acreditamos que nosso experimento é o primeiro a avaliar empiricamente o impacto individual de tal expansão em um cenário que retrata a realidade dos engenhos de busca da Web, caracterizado por apresentar: (1) uma grande coleção de documentos não-estruturados, longos e heterogêneos; e (2) um conjunto de consultas curtas e compostas apenas por substantivos.

Em termos de resultados, [15] relatou que a expansão de consultas através dos conjuntos de sinônimos do WordNet realmente diminuiu a eficiência da recuperação, inclusive quando os sinônimos foram selecionados manualmente (eliminando a ambigüidade). Em

[22] Incluindo os discos 1 e 2 do TREC.
[23] Um exemplo de tais consultas e campos foi ilustrado na Fig. 1.

contraste, [1, 12] relataram resultados diferentes, afirmando que a eficiência da recuperação foi melhorada com a expansão, inclusive, no caso de [1], quando não houve nenhum tratamento de ambigüidade dos termos.

Nossos resultados positivos aparecem assim, como um estudo capaz de justificar os resultados contraditórios dos trabalhos prévios: eles sugerem que o fator que impediu que a expansão das consultas resultasse em qualquer benefício nos experimentos de [15] não foi a grande coleção de documentos longos e heterogêneos, mas as longas consultas utilizadas.

5 Conclusão e Trabalhos Futuros

Foi apresentado, neste artigo, um experimento que empiricamente mediu o impacto em RI do processo de expandir semanticamente, e de forma automática, as palavras-chave de consultas submetidas à engenhos de busca da Web. A tarefa de expansão foi auxiliada por um thesaurus que forneceu as palavras semanticamente relacionadas com os termos originais da consulta. Os experimentos comprovaram que, de uma forma geral, tal expansão sempre melhora a eficiência média sobre o conjunto total de documentos recuperados. Foi argumentado também que a estratégia de expansão que traz melhores ganhos é aquela que restringe a expansão somente aos sinônimos associados com o sentido mais comum da palavra-chave original. Esta estratégia foi a melhor tanto para o conjunto total de documentos recuperados (melhorando a medida-F em 37%), como para os conjuntos limitados aos 20, 30, 40 e 50 primeiros documentos recuperados (melhorando a precisão em 1%, 12%, 17% e 37%, respectivamente).

Este é o primeiro experimento que avalia amplamente o impacto individual de uma técnica de expansão automática de consultas baseada no WordNet e aplicada em um cenário de experimentos similar ao mundo real dos engenhos de busca da Web. Os resultados são encorajadores sobre o real potencial de técnicas de PLN incorporadas nestes engenhos. Eles também sugerem a necessidade de realização de muitos outros experimentos similares para que se possa identificar, de forma progressiva, as reais potencialidades de cada diferente técnica de PLN. Estes resultados reforçam, ainda, a importância de um thesaurus no processo de recuperação de informação. Contudo, os experimentos foram realizados utilizando um thesaurus e uma coleção de testes, ambos, da língua inglesa. Já existe um WordNet para outras línguas (como o francês, o italiano, o alemão, dentre outras línguas européias). Trata-se do EuroWordNet [16]. No entanto, não existe ainda um WordNet para a língua portuguesa, apesar de já existirem esforços nesta direção, como [3, 9]. Estas iniciativas são por demais importantes para que a expansão de consultas baseada em thesaurus possa ser aplicada também a termos da língua portuguesa.

Como trabalho futuro, nós pretendemos: (1) realizar a expansão através de variantes morfológicas dos termos da consulta; e (2) criar uma heurística, baseada nas relações do

WordNet, para realizar tratamento de ambigüidade entre estas palavras. Nós também planejamos a realização de experimentos que identifiquem o custo computacional do processo de expansão de consultas, identificando assim, o seu custo-benefício.

Referências

[1] Y. Aslandogan, C. Thier, C. T. Yu, J. Zou, and N. Rishe. Using semantic contents and wordnet in image retrieval. In *Proc. 20th Annual International ACM-SIGIR Conference*, pages 286–295, Belkin, N. J., 1997.

[2] Baeza-Yates and Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] B. C. Dias-da Silva, H. R. Moraes, and M. F. Oliveira. Groundwork for the development of the brazilian portuguese wordnet. *Lecture Notes in Artificial Intelligence*, 2389: Advances in Natural Language Processing:189–196, 2002.

[4] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proc. Coling-ACL'98 Workshop on Usage of WordNet for Natural Language Processing*, Montreal, Canada, 1998.

[5] D. Harman. Overview of the first text retrieval conference. In *Proc. 16th Annual International ACM/SIGIR Conference*, pages 36–48, Pittsburgh, PA, 1993.

[6] H. Jing and E. Tzoukermann. Information retrieval based on context distance and morphology. In *Proc. 22th Annual International ACM-SIGIR Conference*, pages 90–96, Berkeley, California, 1999.

[7] J. Kekäläinen and K. Järvelin. The impact of query structure and query expansion on retrieval performance. In *Proc. 21st Annual International ACM-SIGIR Conference*, pages 130–137, Melbourne, Australia, 1998.

[8] R. Krovetz. Viewing morphology as an inference process. In *Proc. 16th Annual International ACM/SIGIR Conference*, pages 191–203, Pittsburgh, PA, 1993.

[9] P. Marrafa. Wordnet do português: uma base de dados de conhecimento linguístico, 2001. Lisboa, Instituto Camões, 2001.

[10] R. Richardson and A. F. Smeaton. Using wordnet in a knowledge-based approach to information retrieval, 1994. School of Computer Applications Working Paper CA-0395, Dublin City University, 1994.

[11] V. Rijsbergen. *Information Retrieval*. ButterWorths, London, 2nd edition, 1979.

[12] A. F. Smeaton and I. Quigley. Experiments on using semantic distances between words in image caption retrieval. In *Proc. 19th International Conference on Research and Development in Information Retrieval*, pages 174–180, Zurich, Switzerland, 1996.

[13] M. A. Stairmand. Textual context analysis for information retrieval. In *Proc. 20th Annual International ACM-SIGIR Conference*, pages 140–147, Philadelphia, PA, 1997.

[14] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proc. 16th Annual International ACM/SIGIR Conference*, pages 171–180, Pittsburgh, PA, 1993.

[15] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proc. 17th Annual International ACM/SIGIR Conference*, pages 61–69, Dublin, Ireland, 1994.

[16] P. Vossen. *EuroWordNet - A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic publishers, 1998.