Luppar: An Information Retrieval System for Closed Document Collections

Fabiano Tavares da Silva[1], **José Everardo Bessa Maia**[1]

[1]Universidade Estadual do Ceara - Fortaleza - CE - Brasil

Email: fabiano.tavares@aluno.uece.br, jose.maia@uece.br

Abstract. *This article presents Luppar, an Information Retrieval tool for closed collections of documents which uses a local distributional semantic model associated to each corpus. The system performs automatic query expansion using a combination of distributional semantic model and local context analysis and supports relevancy feedback. The performance of the system was evaluated in databases of different domains and presented results equal to or higher than those published in the literature.*

1. Introdução

A questão central em Recuperação de Informação (RI) é atender a necessidade de informação do usuário, apresentada através de uma consulta, retornando os documentos relevantes para esta consulta, caso eles existam. Os sistemas de recuperação de informação (SRI), em geral, possuem uma interface simples: um campo de entrada, para texto livre, para receber a consulta, e apresenta no retorno uma lista dos resultados considerados relevantes para essa consulta. Um exemplo são os motores de busca.

O processo nem sempre é eficaz, sendo que a ineficácia muitas vezes é causada pelo uso impreciso de palavras-chave na consulta. Na prática o usuário adiciona poucas palavras e com ambiguidade [Carpineto and Romano 2012]. Isso faz com que o usuário necessite refinar sua consulta com várias interações adicionando novas palavras ou então desista da busca. Uma abordagem comumente utilizada para resolver este problema é através da expansão de consulta [Xu and Croft 1996] onde se adiciona novas palavras ou frases aos termos da consulta para que o SRI recupere novos documentos e aproxime-se da necessidade do usuário. Os dois problemas comumente encontrados na consulta inicial são de especificidade ou de abrangência de significado dos termos utilizados.

Adicionar novos termos a uma consulta pode ser feita de forma manual pelo usuário, semiautomático ou automático. Na forma manual o usuário julga o que poderia melhorar sua consulta e reformula os termos da consulta. Na forma semiautomática o sistema assiste o usuário, por exemplo, sugerindo a adição de novos termos. Na expansão automática não há participação do usuário sendo a adição de novos termos realizada de forma implícita. Luppar realiza expansão automática de consulta.

A expansão de consulta (QE - *query expansion*) automática em um SRI pode ser separado em dois problemas: o mecanismo que dispara a expansão da consulta e o método de expansão da consulta. Em relação ao mecanismo que dispara a expansão da consulta, pode ocorrer em três modalidades, ou por quaisquer das suas combinações: expansão cega, baseada nos termos iniciais da consulta; expansão por pseudo realimentação de relevância, baseada nos documentos do topo recuperados pela consulta inicial; ou expansão

por realimentação de relevância explícita, na qual a interface do sistema oferece ao usuário informações iniciais sobre os documentos recuperados juntamente com links através dos quais o usuário pode indicar o(s) documento(s) mais relevante(s), os quais são utilizados no método de expansão.

O método geral de expansão consiste em construir ou utilizar um tesauro para adicionar novos termos, atribuir pesos ou recalculá-los para assim modificar a representação original. Esse processo pode ser visto no diagrama de fluxo da Figura 1. A consulta inicial do usuário é modificada utilizando termos de um tesauro, sendo essa consulta modificada a que de fato é submetida ao sistema de busca.

Um tesauro é uma lista de palavras com significados semelhantes ou afins. O algoritmo de construção do tesauro diferencia as abordagens utilizadas. Esses algoritmos podem ser caracterizados em três dimensões: pela abrangência (escopo) dos textos utilizados na construção do tesauro, pela noção de contexto adotada (quando se utiliza a hipótese de que o significado das palavras é dado pelo seu contexto de uso), e pela medida utilizada na avaliação da semelhança semântica entre as palavras.

Em relação à abrangência, um tesauro pode ser amplo, construído para todo o vocabulário e ocorrências conhecidas de uso das palavras de uma língua, ou pode ser restrito aos termos presentes em uma coleção particular de documentos. A noção de contexto refere-se à hipótese adotada sobre como se determina o significado das palavras na língua. Por exemplo, pode-se adotar a hipótese de que o significado de uma palavra pode ser inferido das palavras que ocorrem no seu entorno, sendo que o entorno adotado pode ser uma janela determinada, a sentença, todo o parágrafo ou mesmo o documento. A medida de semelhança, por sua vez pode variar sendo probabilística ou determinística, pode considerar ou não a distância entre as palavras, ou pode ainda utilizar elementos externos tais como uma ontologia de domínio [Bhogal et al. 2007].

Um dos tesauros mais conhecidos é Wordnet [Miller 1995] construído de forma manual e com características globais. Tem a vantagem de trazer informações léxicas o que resolvem problemas de ambiguidade em alguns casos. Já as desvantagens é que são genéricos, dessa forma não trazem ganhos em domínios específicos e trabalhosos para incluir novos termos [Ooi et al. 2015]. Mesmo sendo difícil de crescer esse tipo de tesauro vários são os trabalhos [Gong et al. 2005, Lu et al. 2015, Hsu et al. 2006] que utilizam essa abordagem para carregar pares de sinônimos com o objetivo de reformular as consultas.

Os tesauros construídos de forma automática são baseados na hipótese distribucional de [Harris 1954] que afirma que as palavras que são usadas e ocorrem nos mesmos contextos tendem a ter significados semelhantes. A partir dessa hipótese construiu-se teorias e métodos para representar e quantificar a similaridade entre itens de dados linguístico. Um modelo baseado nessa hipótese é chamado de Semântica Distribucional. Para criar uma representação, dois tipos de modelos são geralmente construídos: modelos de contagem ou modelos preditivos. Na abordagem de contagem, tradicionalmente, usa-se as estatísticas de co-ocorrência das palavras e assim cria-se vetores no espaço de palavras [Turney and Pantel 2010]. A alta dimensionalidade dessa representação é então reduzida, sendo a densificação realizada por decomposição em valores singulares (SVD) [Landauer and Dumais 1997] ou por análise de componentes principais (PCA)

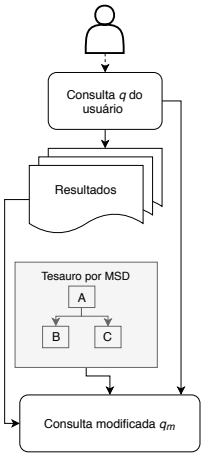

Figura 1. Diagrama de fluxo da expansão automática de uma consulta

[Lebret and Collobert 2015]. Já as abordagens preditivas utilizam as estatísticas dos termos para treinar redes neurais que criam vetores densos usados como representação dos termos [Mikolov et al. 2013a]. Para estas representações em geral são utilizadas uma das seguintes medidas de similaridade: Cosseno, medida de Lin [Lin 1998] ou coeficiente de Dice [Curran and Moens 2002].

Entende-se aqui uma coleção fechada de documentos como sendo um repositório eletrônico local de documentos, estável no horizonte de busca e processamento. Isso está em contraste com recuperação de informação na Web, em larga escala e de múltiplos proprietários e em constante evolução. Exemplos de coleções fechadas são uma biblioteca digital ou um repositório de informações clínicas de pacientes.

A proposta Luppar, descrita nesse artigo, é um modelo híbrido de expansão de consulta combinando propriedades de tesauros locais e globais. Um tesauro com características globais é construído, no qual cria-se um modelo distribucional semântico para os termos do vocabulário a partir de toda coleção. Isto é combinado com características locais de utilizar apenas o resultado do topo da consulta para restringir o contexto dos termos e aumentar a relevância dos documentos recuperados. O restante deste artigo está organizado em três seções. A Seção 2, Métodos, descreve os algpritmos utilizados em Luppar. A Seção 3, Resultados, dá detalhes da implementação e os procedimentos e resultados da avaliação. A seção seguinte é de conclusão.

2. Métodos

Esta seção descreve os conceitos e métodos utilizados na construção de Luppar. Luppar é constituído por uma combinação de Análise de Contetxo Local com Modelo Semântico Distribucional com uma medida de similaridade de textos associada.

2.1. Modelos de Semântica Distribucional

Esta subseção brevemente apresenta DSM (*Distributional Semantic Model*) e como ele é utilizado em Luppar. Uma característica fundamental de um tesauro é qualidade do grau de similaridade entre termos medida pela semelhança semântica. Os Modelos de Semântica Distribucional (DSM) são construídos utilizando a hipótese distribucional de [Harris 1954]. As palavras que possuem contexto semelhantes tendem a terem o mesmo

significado. O DSM é uma representação das palavras em espaços geométricos de palavras onde vetores expressam conceitos, e sua proximidade é uma medida semântica. Isso significa que as palavras são semanticamente semelhantes se os contextos (palavras vizinhas) nos quais aparecem são semelhantes e deve levar a que suas representações sejam próximos. Construção de um DSM envolve a definição de um modelo de distribuição, formado por uma quádrupla [Lowe 2001]: vetores de palavra e dimensão; uma função que leva em conta as co-ocorrências e como esses itens são representados no vetor final; uma função de similaridade definida sobre vetores; e eventualmente um mapeamento que transforma o espaço vetorial.

Dado um corpus e através de um processo de treinamento não supervisionado é possível alcançar uma representação das palavras em um espaço de contexto. A partir dessa representação de palavras e suas operações de similaridade é possível construir um tesauro semântico. Vários trabalho demonstram que espaços de palavras baseados em redes neurais (*Word Embeddings*) superam os modelos tradicionais baseados em contagem para calcular a similaridade da palavra [Mikolov et al. 2013a]. Para este trabalho especificamente foi utilizado o modelo preditivo distribucional *Word2Vec* de [Mikolov et al. 2013a].

Word2Vec é um método de *Word Embedding* utilizado para induzir modelos de espaço vetorial utilizando deep learning em redes neurais com modelos de linguagens [Mikolov et al. 2013a]. Baseia-se em uma rede neural simplificada com o número de entradas proporcional ao de palavras do vocabulário. A camada escondida realiza uma projeção linear com tantos nós quanto a dimensionalidade desejada do espaço vetorial. Esse espaço de características é projetado sobre uma camada de saída hierárquica *soft-max*. A rede é treinada em cada par de exemplo de entrada-saída por vez e, para cada par, a diferença entre a saída esperada e a real da rede é calculada. Os pesos da combinação linear da rede são posteriormente ajustados para diminuir o erro usando o procedimento de *back propagation*. Este procedimento é repetido para todos os pares de dados de treinamento, muitas vezes em várias passagens sobre todo o conjunto de dados de treinamento, até que a rede convirja e o erro não diminua mais [Bengio et al. 2003]. Este método vem motivando bons resultados pois mostrou produzir representações que preservam importantes características linguísticas [Mikolov et al. 2013b]. Na prática, [Levy and Goldberg 2014] demostraram que uma das principais vantagens *Word2Vec* reside na sua escalabilidade, permitindo o treinamento com até bilhões de palavras de texto de entrada em várias horas, diferenciando-se da maioria dos outros DSM.

2.2. Análise de Contexto Local com DSM

A expansão automática da consulta (sem interferência direta do usuário) pode ser baseada em métodos locais ou métodos globais. A *Análise Local* faz uso de n documentos inicialmente capturados pela primeira interação e, sem a participação do usuário, utiliza os termos mais próximos para expandir a consulta. Por outro lado, a abordagem de *Análise Global* faz uso de um tesauro externo, quer seja construído por especialista ou de forma automática. Já no trabalho de [Xu and Croft 1996], é proposto usar o primeiros resultados de uma consulta para construir uma representação por concorrência de conceitos (grupos de substantivos) e por similaridade destes com a consulta encontrar aqueles candidatos a serem agregados à expansão da consulta, ou seja, combinar análise local e global para QE. Em [Ermakova and Mothe 2016] este processo segue a mesmo raciocínio porém com

refinamento nos métodos envolvidos.

O algoritmo 1 apresenta o pseudo código do método ACL proposto:

Algorithm 1: Pseudocódigo de EC proposto com ACL e MSD

Data: query q, thesaurus th
Result: modified query q_m

```
1:  documents ← search_top_ranked(q);
2:  for each documents do
3:      passages ← window(document);
4:      for each passages do
5:          concepts ← find_concepts_in_context(passage);
6:      end for
7:  end for
8:  sort(concepts);
9:  for i ← 1 to N do
10:     m[i] ←sim(q, concepts[i], th);
11: end for
12: sort(m);
13: q_m ← q + m[1..n];
```

O algoritmo 1 toma como entrada a consulta original q, e th, o tesauro DSM previamente construído para coleção fechada de documentos da respectiva consulta. Na linha 1, recupera-se com a consulta original os documentos mais bem ranqueados. Nas linhas 2 a 8, recupera-se as n passagens mais bem ranqueadas usando a consulta original. Isto é conseguido quebrando os documentos inicialmente recuperado pela consulta em *passagens* (sentenças) e classificando as passagens como se fossem documentos. Nas linhas 9 a 11, para cada conceito (grupo de substantivos) nas passagens do topo dos resultados, calcula-se a similaridade $sim(q, c)$ entre toda a consulta q (e não os termos individuais da consulta) e o conceito, usando uma variante do TF-IDF. Nas linhas 12 e 13, finalmente, os m conceitos mais bem ranqueados, de acordo com $sim(q, c)$, são adicionados à consulta original q. Para cada conceito adicionado atribui-se um peso dado por $1 - 0, 9xi/m$, onde i é a posição do conceito no ranking de conceitos. Os termos na consulta original q podem ser enfatizados através da atribuição de um peso igual a 2 para cada um deles.

Este algoritmo para expansão de consulta com ACL que possui características locais e globais bem definidas. O melhor da análise local é poder utilizar os resultados do topo na qual assumimos serem os melhores resultados relacionados. Na análise global utilizamos o conceito de contexto e estruturas frasais sobre o conjunto local [Baeza-Yates and Ribeiro-Neto 2013]. Há dois pontos sensíveis nesse algoritmo de ACL 1. O primeiro na linha 5, que corresponde ao método de encontrar os conceitos (grupo de substantivos mais significativos no topo da consulta original) e o segundo a função de similaridade, na linha 10, onde é calculado o *score* entre cada conceito e a consulta. É justamente nesses dois aspectos que nossa abordagem se diferencia.

Primeiro, para extrair os conceitos é utilizada uma redução da representação dos documentos do topo do rank, no qual Croft chamou de passagem, uma estrutura de documento um pouco mais reduzida com um janela de tamanho W. Na nossa abordagem

essa janela é automaticamente criada utilizando-se de um período completo, até seu fechamento por uma pontuação. Assumimos que um significado está mais fechado a um período completo ou parágrafo. A representação do período continua sendo *bag-of-word*. Note que esta estratégia corresponde a uma janela de tamanho variável.

O segundo ponto sensível do algoritmo é a função de similaridadeque quantifica a correlação entre o conceito e o termo da consulta. O objetivo final é expandir e estimar a consulta original com uma probabilidade P_q para que assim ela possa ser utilizada no modelo de linguagem para recuperar documentos. No trabalho de [Xu and Croft 1996] é proposta uso da correlação simples combinado com uma variante não trivial do TF-IDF. Já o trabalho de [Ermakova and Mothe 2016] além de fazer uso de estatísticas sobre o documento, também se utiliza de informações dos termos que cercam os termos da consulta com uma fórmula de similaridade que usa a distância e a correlação. A proposta deste trabalho é tomar como base a similaridade proposta por Croft com a construção de passagens curtas do topo do rank para os conceitos, mas também, como no trabalho de Emakorva, encontrar palavras com significados que vão além da correlação entre os termos da consulta e os conceitos. A estratégia foi utilizar a hipótese distribucional na função f para determinar se o conceito merece ser um termo candidato a expansão com bom significado associado a consulta. A partir dos corpus definidos um processo off-line foi utilizado para construir os DSM. Esses modelos são baseados em contextos que possuem a operação de similaridade, e assim podem agregar uma estimativa de P_q associada aos contextos no quais os termos da consulta poderiam vir a aparecer nos documentos.

Em linhas gerais, o método consiste em utilizar as passagens para limitar o universo de possibilidades de termos que contextualizam a consulta e isto é uma das vantagens da abordagem local pois os resultados iniciais trazem muito da real intenção do usuário. A relevância da consulta agora deve ser refinada para documentos despercebidos por questões de sinonímia ou polissemia. Então dessas passagens são retirados os conceitos conforme ACL. Esses conceitos são então selecionados segundo o critério de similaridade semântica construído pelo modelo distribucional de maneira global. Essa segunda etapa faz total diferença em comparação ao modelo global por tesauro porque por mais que duas palavras tenham similaridade semântica e sempre apareçam no mesmo contexto é através da ACL que restringimos a relevância do novo termo para expansão.

2.3. Expansão de Consulta Automática (AQE)

O método de recuperação em duas etapas proposto e implementado em Luppar faz uso de medidas de similaridade em dois momentos: na recuperação sem expansão inicial e na recuperação com a consulta expandida baseada nos documentos ranqueados no topo daqueles recuperados na primeira etapa.

É dada uma coleção D de documentos onde cada documento d está representado em um espaço vetorial de palavras de t termos, com os termos indexados formando um vocabulário $V = \{k_1, k_2, ...k_t\}$. Neste espaço, cada documento é representado por um vetor de pesos das palavras $di = \{w_{1,i}, w_{2,i}..., w_{t,i}\}$. Seja q a consulta representada como um pseudo documento também no mesmo espaço $q = \{w_{1,0}, w_{2,0}..., w_{t,0}\}$ sendo w_t o peso associado ao termo na consulta. A similaridade entre a consulta q e o documento d é expresso da seguinte forma:

$$sim(d, q) = \sum_{t \epsilon q \cap d} w_{t,d} \cdot w_{t,q} \, , \tag{1}$$

onde $w_{t,q}$ e $w_{t,d}$ são pesos calculados através do TF-IDF (frequência do termo pelo inverso da frequência nos documentos): $w_{i,j} = (1 + \log f_{i,j}) \times \log \frac{N}{n_i}$, se $f_{i,j} > 0$, senão 0.

A segunda etapa é a recuperação com a consulta expandida. O processo de AQE consiste em utilizar os termos de q para encontrar novos termos, sem a participação do usuário, que venham resolver problemas de desambiguação e que melhor discriminem os documentos. Dado inicialmente q que passará a ter novos termos e será denominada q'. Os novos termos compõem w_j em q'. Agora a similaridade é calculada conforme fórmula 1 mas com $sim(d, q')$. A escolha dos novos termos é realizada através de dois tesauros. Um tesauro global e estático *WordNet* [Miller 1995] e um outro construído automaticamente utilizando Modelo de Semântica Distribucional com representação *Word Embbeding* [Mikolov et al. 2013a].

No processo online a consulta possui a mesma representação dos documentos. O *Vector Space Model* (VSM) [Salton et al. 1975] e o probabilístico Okapi BM25 [Robertson and Zaragoza 2009] foram implementados e utilizados como modelos de busca em RI. Na ACL [Xu and Croft 1996] todo processo para expansão da consulta ocorre online e não depende de tesauro global. Entretanto, esta abordagem exige o uso do DSM, o que necessita de um treinamento prévio para construção do modelo baseado em contextos. No processo de busca o método tem como entrada a consulta original q que será expandida e qualificada com a probabilidade P_q para cada termo da consulta original e os termos a ela acrescentados. A partir de q realizamos uma consulta rápida com VSM e obtemos n documentos no topo do *rank*. Através desses documentos são criadas as passagens (estruturas menores e com significado). Neste trabalho uma passagem é um período fechado por uma pontuação. As passagens são ordenadas como se fosse pequenos documentos e também utilizando VSM e a consulta original. Dessas passagens selecionamos as m passagens mais bem ranqueadas e então extraímos os conceitos que servirão de candidatos a expansão. Então seleciona-se os melhores conceitos (maiores P_q) extraídos das passagens do topo dos documentos da consulta. A similaridade $sim(q, c)$ entre cada conceito c e a consulta original q é computada por:

$$sim(q, c) = \prod_{k_i \in q}^{c} (\delta + \frac{\log(f(c, k_i) \times IDF_c)}{\log n})^{IDF_i} , \quad (2)$$

onde k_i corresponde a cada termo de q. IDF$_i$ e IDF$_c$ são o inverso da frequência sobre termo da consulta i e sobre o conceito c respectivamente. O $F(c, ki) = w2v.sim(c, k_i)$ onde $w2v.sim$ é a função do Word2vect utilizada para medir a similaridade semântica através do coseno do vetor gerado pelo DSM. Em seguida são ranqueados os conceitos que estão mais próximos da consulta como um todo. Os m melhores conceitos são escolhidos para serem quantificados segundo sua importância. No trabalho [Ermakova and Mothe 2016], Ermakorva propõem penalizar os candidatos em vários aspectos (IDF, *score*, importância e POS) enquanto que essa abordagem manteve a penalizar com peso 2 para as palavras da consulta original e penalizar os m conceitos com $1 - 0.9xi/m$ [Xu and Croft 1996] e também com o IDF, tendo em vista que os outros scores pouco alteraram significativamente a precisão.

Assim, na aplicação da abordagem dois macro processos são executados. O primeiro off-line para preprocessamento do corpus, construção do índice invertido, armazenamento das estatísticas de contagem e treinamento do DSM. Já o segundo processo

Tabela 1. Coleções de teste

Coleção	Assunto	Idioma	Núm. de Termos	Núm. de Docs	Núm. de Tópicos	Tam. med. Documento (termos)	Tam. med. Tópico (termos)	Tamanho Tóp. x Doc (%)
MED	Medicina	Inglês	9622	1033	30	167.2	23.8	14.2
LISA	Biblioteconomia	Inglês	13706	6004	35	97.5	66.1	67.7
NPL	Eng. Elétrica	Inglês	7861	11429	100	41.9	10.9	26

online é responsável por efetivamente expandir a consulta por ACL e assim realizar a busca fazendo uso dos modelos de linguagem.

O projeto admite gerenciar múltiplos corpus de documentos. Na primeira etapa para cada *corpus* é construído um índice invertido. No índice ficam armazenados algumas estatísticas, o modelo de linguagem e o dicionário dos termos. Os termos em *stopwords* são removidos. O processo de radicalização Porter [Porter 1980] é aplicado. A frequência e a distância são armazenados no índice invertido. No processo de construção da representação dos documento utiliza-se o modelo unigram (*bag-of-word*) com fator de peso TF-IDF para a relação entre termos e documentos. O TF-IDF utiliza a normalização de frequência $L2$ [Amati and Van Rijsbergen 2002]. Com o modelo unigram completo iniciamos a etapa de treinamento para construção do DSM. Utilizamos um modelo de linguagem predito (*Word Embedding*). Especificamente utilizamos a implementação *Word2Vect* [Mikolov et al. 2013a] com arquitetura CBOW (*Continuos Bag-of-Word*). A dimensionalidade do vetor utilizada foi tamanho de 300. Janela de tamanho $W = 5$ para os contextos. Com todas essas etapas cumpridas o sistema está pronto para receber uma consulta.

3. Resultados e Discussão

Esta seção apresenta os conjuntos de dados e as figuras de mérito utilizadas na avaliação. E os resultados.

3.1. Dados e Avaliação

A avaliação fez uso de três conjuntos de dados já comumente usados como referência para avaliação em recuperação de informação. As três coleções de teste foram elaboradas por Ed Fox na Virginia Polytechnic Institute and State University[One 1990]. A tabela 1 mostra as características dessas coleções.

Todas as coleções são formadas por três arquivos. Um arquivo com os documentos, o segundo com as consultas e o terceiro a identificação da consulta e quais documentos foram julgados relevantes pela consulta. Estes documentos são entradas para os algoritmos propostos e sua execução produziu um quarto documento com os julgamentos automático de arquivos recuperados. Essa organização conhecida no paradigma *Cranfield* permitiu avaliar e comparar o par necessidade de informação-documento.

Os dados MED são documentos curtos mas com consultas de precisão alta mesmo sem expansão de consultas e com consultas curtas. O LISA são documentos com precisão baixa e com consultas longas que vão de oposto a necessidade de expandir consulta. O NPL com tamanho curto dos documentos, mas com uma quantidade bem maior de documentos, com precisão mediana e com muitas consultas variando entre longas e curtas.

Dessa forma foi possível traçar os resultados em cima desses três cenários e verificar qual melhor resposta para abordagem.

Para avaliar o desempenho da expansão de consulta ACL utilizando DSM foram implementados outros três métodos para comparação. A expansão de consulta utilizando a WordNet [Miller 1995], isto é, um tesauro externo global, a própria ACL com características locais [Xu and Croft 1996], mas sem empregar o DSM e a consulta original, sem expansão, tomada como *baseline*. As métricas utilizadas são as mesma da conferência TREC [Hashemi et al. 2016]. Inclusive o mesmo software chamado de *trec_eval* em sua última versão 9.0. Das onze métricas produzidas, quatro foram escolhidas para este trabalho:

- **MAP:** precisão média sobre todas as consultas;
- **Bpref:** calcula uma relação de preferência, ou seja, se os documentos julgados relevantes são recuperados antes daqueles julgados irrelevantes;
- **Reciprocal Rank:** precisão média em relação ao primeiros resultados.
- **A curva Recall-Precision:** 11 pontos interpolando a precisão média (em 0%, 10%,...., 100% do *recall*) que permite desenhar a curva de cobertura e precisão permitindo perceber que à medida que os documentos mais relevantes são recuperados (o *recall* aumenta) e enquanto documentos irrelevantes são recuperados (a precisão diminui).

3.2. Resultados

As Tabelas 3, 2 e 4 apresentam os resultados dos experimentos para os conjuntos de dados MED, LISA e NPL. Cada tabela mostra quatro colunas de resultados: *baseline*, wordnet, ACL, e ACL-DSM. A coluna *baseline* refere-se à recuperação de informações por uma consulta sem expansão. Os resultados na coluna wordnet referem-se a expansão de consulta com o tesauro wordnet. A coluna ACL registra os resultados onde a expansão da consulta é baseada no contexto local utilizando sinônimos do Wordnet. finalmente, a coluna ACL-DSM aplica a representação semântica distribucional (word embedding) combinada com Análise de Contexto Local para construir um dicionário de sinônimos e usa este dicionário de sinônimos na expansão da consulta. Os experimentos foram realizados utilizando os modelos VSM e BM25.

Note que os resultados para ACL-DSM são consistentemente superiores para as três bases, para os dois modelos e os três índices de desempenho. Em apenas um caso, a métrica *map*, no modelo BM25 sobre a base de dados LISA o índice foi ligeiramente menor. As curvas *Recall-Precision* nas Figuras 2, 3 e 4, uma das principais métricas de comparação de performance por apresentarem uma comparação não pontual, confirmam esses resultados. Em todas elas, a curva para BM25-ACL-DSM está por cima das demais. Note ainda que *MAP* é uma medida aproximada da área sob a curva recall-precision (RP-AUC) não interpolada e confirma essas conclusões.

4. Conclusão

Luppar é um Sistema de Recuperação de Informação (SRI) concebido e implementado com vista ao uso corporativo, ou seja, para coleções fechadas de documentos (não para web). A aplicação tira proveito disso incluindo um tesauro semântico baseado em *word embedding* construído apenas com os documentos das coleções alvos. Essa abordagem

Tabela 2. Resultados para coleção LISA

Modelos	Métricas	*Baseline*	*WordNet*	ACL	ACL-DSM
VSM	map	0,2641	0,2034	0,2475	0,2602
	bpref	0,9981	1,0	1,0	0,9981
	recip_rank	0,5184	0,4618	0,5006	0,5038
BM25	map	0,3495	0,2520	0,3577	0,3627
	bpref	0,9981	1,0	1,0	0,9981
	recip_rank	0,6459	0.5085	0,6400	0,6693

Tabela 3. Resultados para coleção MED

Modelos	Métricas	*Baseline*	*WordNet*	ACL	ACL-DSM
VSM	map	0,5142	0,4949	0,5255	0,5348
	bpref	0,8985	0,9318	0,9418	0,9406
	recip_rank	0,8537	0,7726	0,8889	0,8889
BM25	map	0,5033	0,4873	0,5262	0,5459
	bpref	0,8985	0,9318	0,9660	0,9712
	recip_rank	0,8992	0,8294	0,8253	0,8944

Tabela 4. Resultados para coleção NPL

Modelos	Métricas	*Baseline*	*WordNet*	ACL	ACL-DSM
VSM	map	0,1886	0,1428	0,1968	0,2282
	bpref	0,9767	0,9886	0,9878	0,9333
	recip_rank	0,4437	0,3583	0,5014	0,4267
BM25	map	0,2124	0,1756	0,2640	0,2580
	bpref	0,9766	0,9819	0,9891	0,9815
	recip_rank	0,5987	0,5432	0,6142	0,6373

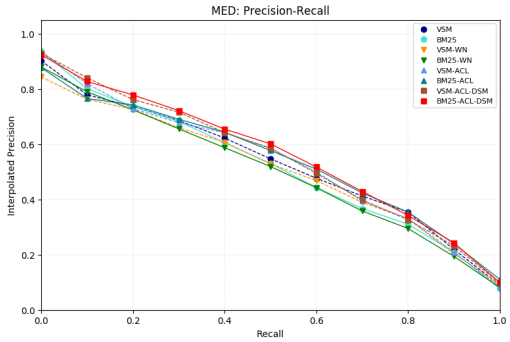

Figura 2. Precision x Recall para coleção MED.

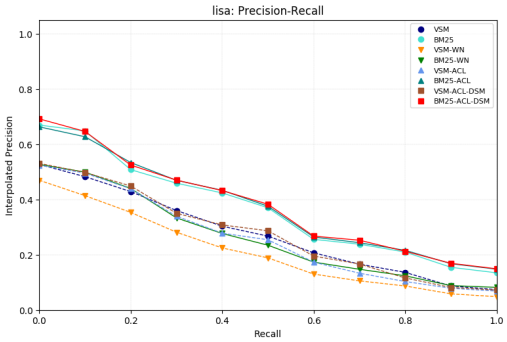

Figura 3. Precision x Recall para coleção LISA.

evita que consultas sejam expandidas com termos que, embora sejam significativos na lingua, são inexistente nas coleções em foco. *Word embedding* restrito ao corpus combinado com Análise de Contenxo Local (ACL-DSM) completam a proposta em Luppar.

O trabalho utilizou critérios e métodos da conferência TREC para avaliar a proposta. Os resultados das Tabelas 3, 2 e 4 mostram que os métodos de RI de Luppar são satisfatórios e compatíveis com o estado da arte. Os experimentos foram construídos de forma a revelar o ganho de eficácia da combinação ACL-DSM em relação a cada método individual, ACL ou DSM, aplicados isoladamente. A performance da consulta sem expansão foi incluída com método *baseline* para controle dos experimentos.

Uma evolução natural deste trabalho é extendê-lo para recuperação de informação na web. Luppar para busca na web é um projeto em andamento.

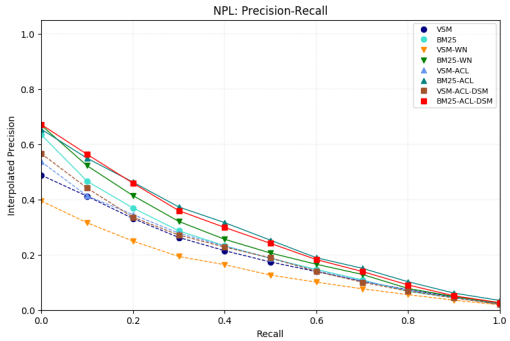

Figura 4. Precision x Recall para coleção NPL.

Referências

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Baeza-Yates, R. and Ribeiro-Neto, B. (2013). *Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bhogal, J., Macfarlane, A., and Smith, P. (2007). A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50.

Curran, J. R. and Moens, M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 59–66. Association for Computational Linguistics.

Ermakova, L. and Mothe, J. (2016). Query Expansion by Local Context Analysis. *Coria*, pages 1–16.

Gong, Z., Cheang, C. W., and Hou, U. L. (2005). Web query expansion by wordnet. In *International Conference on Database and Expert Systems Applications*, pages 166–175. Springer.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hashemi, S. H., Clarke, C. L., Kamps, J., Kiseleva, J., and Voorhees, E. M. (2016). Overview of the trec 2016 contextual suggestion track. In *Proceedings of TREC*, volume 2016.

Hsu, M.-H., Tsai, M.-F., and Chen, H.-H. (2006). Query expansion with conceptnet and wordnet: An intrinsic comparison. In *Asia Information Retrieval Symposium*, pages 1–13. Springer.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lebret, R. and Collobert, R. (2015). Rehabilitation of count-based models for word vector representations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 417–429. Springer.

Levy, O. and Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

Lowe, W. (2001). Towards a theory of semantic space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Lu, M., Sun, X., Wang, S., Lo, D., and Duan, Y. (2015). Query expansion via wordnet for effective code search. In *Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on*, pages 545–549. IEEE.

Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

One, V. D. (1990). Cd-rom from virginia polytechnic institute and state university. *Blacksburg, VA*.

Ooi, J., Ma, X., Qin, H., and Liew, S. C. (2015). A survey of query expansion, query suggestion and query refinement techniques. *2015 4th International Conference on Software Engineering and Computer Systems, ICSECS 2015: Virtuous Software Solutions for Big Data*, pages 112–117.

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

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.