Estabelecimento do contexto semântico em textos em linguagem natural

Iúri Chaer[1], Ricardo Luis de Azevedo da Rocha[1]

[1] Laboratório de Tecnologia Adaptativa - Escola Politécnica - Universidade de São Paulo (USP)

{iuri.chaer,luis.rocha}@poli.usp.br

Abstract. This work proposes a method for determining the semantic context (i.e. subject) within a segment of text written in a natural language – based in conditional probability, as posed in the Bayes theorem – to evaluate the relevancy relating each term in the text segment to every subject in a previously defined set.

Resumo. Neste trabalho será proposto um método de determinação do contexto semântico (i.e. assunto) de um trecho de texto em linguagem natural – usando o conceito de probabilidade condicional, de acordo com o teorema de Bayes – para avaliar a relevância dos termos do texto relativos a cada assunto de um conjunto pré-definido.

1. Introdução

Existe uma influência clara do conhecimento *a priori* do assunto sobre o qual um texto trata na maneira como as palavras são interpretadas [Schütze 1998]. Ao se deparar, em uma manchete de jornal, com os dizeres "Bolsa despenca", pode-se imaginar que:

a) o mercado de ações anda mal, ou
b) que está sendo reduzido o incentivo às pesquisas acadêmicas, ou ainda
c) que um recipiente sofreu uma queda.

E essas alternativas são apenas uma pequena amostra de todo o universo de possibilidades. O uso do contexto é essencial para decidir o significado da frase: por ter sido definido que o texto hipotético estaria em uma manchete de jornal já se pode notar, em leitores humanos, certa rejeição à alternativa c. Se, além disso, for especificado que o suposto jornal é especializado em finanças, parece natural esperar que se dê preferência à alternativa a.

Aceitando que o contexto é usado por seres humanos na interpretação semântica de textos, é natural imaginar que ele também seja relevante para sistemas computacionais com o mesmo objetivo. Esse é o motivo deste trabalho: estabelecer uma maneira de determinar os assuntos (doravante chamados domínios semânticos ou simplesmente domínios) a que se refere um texto em linguagem natural, que podem posteriormente ser usados como auxílio na sua interpretação.

A seção 2 deste artigo fala do trabalho descrito em [Magnini e Cavaglia 2000], a partir do qual foi proposto, em [Magnini et al. 2002a], um método para a associação de domínios a palavras modelando o espaço semântico como um espaço vetorial e relacionando manualmente sentidos a domínios. Esse método é explicado na seção 3.1.

 ISBN: 978-85-61228-18-7

Mais tarde, em [Gliozzo et al. 2004a] e [Gliozzo et al. 2004b], o trabalho foi continuado com a proposta de um novo método, baseado na teoria de Bayes em lugar de uma representação espacial, descrito na seção 3.2. Ambos os métodos foram testados na base de estudos de linguagens naturais WORDNET com resultados animadores, dando respaldo à idéia de que a informação do domínio contribui para a interpretação semântica correta. A seção 4 pretende continuar esses trabalhos: é a proposta deste trabalho, que consiste no uso da teoria de Recuperação de Informação para melhorar os resultados do processo de inferência de domínios. A seção 5 encerra o artigo com as conclusões do trabalho.

2. Definindo domínios semânticos

Em [Magnini e Cavaglia 2000], os autores descrevem um trabalho de categorização dos sentidos das palavras da WORDNET[1]. Foi proposto o uso do Dewey Decimal Classification[2] para rotular elementos (pares palavra-significado) de conjuntos de sinônimos. A intenção dos autores era enriquecer a WORDNET, inserindo uma hierarquia adicional para ajudar a diferenciar esses elementos – ou seja, decidir o significado pretendido para uma palavra ambígua.

O quadro 1 mostra um exemplo (extraído de [Magnini et al. 2002a]) de relacionamento palavra-domínio que representa bem a idéia e que pode ajudar a compreender melhor o formato geral da WORDNET: cada palavra tem os seus sentidos enumerados e descritos, formando pares representação-significado. O esforço dos pesquisadores em [Magnini e Cavaglia 2000] foi justamente para rotular esses pares com os domínios.

3. Inferência dos domínios de um trecho de texto

Em uma série de trabalhos do mesmo núcleo de autores [Magnini et al. 2002a], [Magnini et al. 2002b], [Gliozzo et al. 2004a], [Gliozzo et al. 2004b], foram propostos e testados dois métodos de determinação do domínio de textos em linguagem natural. O primeiro exige supervisão humana intensiva, sendo baseado em designações manuais de relacionamentos entre significados de palavras e domínios. O segundo é baseado no uso conjugado de probabilidades condicionais e do teorema de Bayes e de métodos estatísticos de tratamento de dados, exigindo menos interferência hu mana. As subseções seguintes descrevem esses dois métodos.

3.1. Método vetorial

Em [Magnini et al. 2002a] foi proposto um método de determinação do domínio de um texto usando associações, atribuídas por seres humanos, entre significados de palavras e domínios. Ele consiste no seguinte procedimento: cada palavra está associada a um vetor no espaço de domínios, e pode ter, em cada eixo, um comprimento 1 ou 0, indicando respectivamente ter ou não ter relevância para aquele domínio. O comprimento total é depois normalizado para representar a freqüência geral do uso

[1] A WORDNET é um banco de dados contendo informações léxicas e semânticas da língua inglesa.
[2] O Dewey Decimal Classification é um sistema de classificação biblioteconômica criado em 1876 e que ainda hoje é dos mais populares.

daquela palavra no sentido observado na base de dados de referência. Considera -se que um trecho em torno de uma palavra trata de um determinado domínio se as somas das contribuições das 2c palavras (sendo c um número natural ajustado empiricamente) que a cercam para o eixo desse domínio superam o valor médio (somado ao desvio padrão) encontrado na base de dados de referência para o domínio sob análise.

 ISBN: 978-85-61228-18-7

Quadro 1. Domínios relacionados à palavra *bank* (extraído da WORDNET apud [Gliozzo et al. 2004a] e mantido em inglês para manter coerência com a fonte original)

Identificador	Significado	Domínios
#1	depository financial institution, bank banking concern, banking company — (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home")	Economy
#2	bank — (sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank "; "he sat on the bank of the river and watched the currents")	Geography, Geology
#3	bank — (a supply or stock held in reserve for future use (especially in emergencies))	Economy
#4	bank, bank building — (a building in which the business of banking transacted; "the bank is on the corner of Nassau and Witherspoon")	Architecture , Economy
#5	bank — (an arrangement of similar objects in a row or in tiers; "he operated a bank of switches")	Factotum
#6	savings bank, coin bank, money box, bank –(a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty")	Economy
#7	bank — (a long ridge or pile; "a huge bank of earth")	Geography, Geology
#8	bank — (the funds held by a gambling house or the dealer in some gambling games; "he tried to break the bank at Monte Carlo")	Economy, Play
#9	bank, cant, camber — (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)	Architecture
#10	bank — (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank")	Transport

3.2. Método Bayesiano usando Mistura Gaussiana

Em [Gliozzo et al. 2004a] e [Gliozzo et al. 2004b] foi feita uma proposta utilizando probabilidades condicionais para reduzir a interferência humana e tentar melhorar os resultados conseguidos por meio do método vetorial descrito acima. A pontuação, estabelecida a partir da freqüência de um domínio para uma palavra em um texto, é dada por:

$$F(D_k, t, j) = \sum_{i=j-c}^{j+c} \left(R_{palavra}(D_k, w_i^t) . G(i, j, (0{,}5.c)^2) \right) \quad (1)$$

 ISBN: 978-85-61228-18-7

Onde:

F é a pontuação de freqüência do domínio;

D_k é o domínio de índice k;

t é o texto do corpus[3] sendo analisado;

j é a posição da palavra sob escrutínio;

c é a metade do comprimento da janela que se está usando em torno de cad a palavra para determinar o contexto;

$R_{palavra}(D,w)$ é a relação de relevância que associa o par palavra-significado w ao domínio D;

e $G(x, \mu, \sigma)$ é a densidade da distribuição normal de média μ e desvio σ no ponto x.

As relações de relevância entre os pares palavra-significado e os domínios são, evidentemente, o ponto crucial dessa equação, e são definidas como sendo:

$$R_{palavra}(D_k, w_i^t) = \frac{1}{sentidos(w)} \sum_{s \in sentidos(w)} R_{sinônimo}(D_i, s) \qquad (2)$$

Dado que $R_{sinônimo}(D,s)$ é a estimativa da relação entre um conceito e um domínio, estabelecendo-se que conceitos são descritos por conjuntos de sinônimos s.

Mas os autores discutem que, por causa do ruído nos dados gerado por palavras comuns muito ambíguas, a freqüência local calculada pela equação 1 é inadequada. Eles usam o teorema de Bayes, associado ao algoritmo de Maximização de Expectativas aplicado sobre o modelo da Mistura Gaussiana, para obter a relevância de um termo para um domínio. Em suma, a Maximização de Expectativas aplicada sobre o modelo da Mistura Gaussiana dá uma estimativa do valor que tem a maior chance de representar os dados empíricos medidos para as probabilidades condicionais das pontuações de freqüência em relação aos domínios. Com a aplicação desse artifício estatístico, os autores afirmam que não é necessária interferência humana no aprendizado do sistema, já que o ruído nos valores obtidos é removido automaticamente. É um método bem embasado e muito interessante para fazer um pós-processamento em dados colhidos em corpora de treinamento, mas certamente bastante complexo.

4. Método Proposto: Equação tfc·nfx

A proposta, neste trabalho, é usar a equação tfc·nfx descrita em [Salton 1988], baseada no teorema de Bayes mas empiricamente melhorada para tarefas de recuperação de informação. Testes nesse mesmo artigo mostram que a sua precisão e revocação são substancialmente melhores que as da equação probabilística original para tarefas de recuperação de informações. Ela define a relevância dos termos em um documento como sendo:

$$W_t = \frac{tf.log\frac{N}{n}}{\sqrt{\sum W_{t_i}^2}} \qquad (3)$$

Onde:

[3] Corpus, em latim, significa corpo. O termo, em estudos de lingüística, descreve um conjunto estruturado de textos usado como amostra em experimentos.

tf é a freqüência do termo no documento (o número de vezes que ele ocorre no documento);

N é o número total de documentos no corpus sendo examinado;

n é o número de documentos no corpus que contêm o termo;

e $\sqrt{\sum W_{t_i}^2}$ é um fator de normalização que consiste na raiz da soma dos quadrados dos pesos de todos os termos pertencentes a esse documento.

Usando essa fórmula, podemos atribuir a cada domínio valores de afinidade relacionados a cada termo encontrado em textos classificados como pertencentes ao domínio. Em lugar de identificar as palavras manualmente, identificar-se-ia trechos de texto e as afinidades (ou relevâncias) seriam obtidas naturalmente.

Essa métrica de relevância é usada com muito sucesso para buscas em textos simples e só foi abandonada nos dispositivos de busca da Internet após o surgimento de sistemas de busca como o do Google®, que adicionam à métrica informações de referência disponíveis no hipertexto [Brin e Page 1998]. A equação tfc·nfx é mostrada em [Salton 1988] ser substancialmente melhor que o método vetorial no qual [Magnini et al. 2002a] se baseia e também melhor que o método probabilístico simples baseado no teorema de Bayes. A proposta de [Gliozzo et al. 2004a], [Gliozzo et al. 2004b] é essencialmente uma aplicação desse último e provavelmente apresentaria precisão e revocação [4] melhores usando a equação 3.

Ainda há mais uma melhoria em relação aos métodos anteriores que merece atenção: o uso de janelas de tamanho fixo para a análise das palavras é extremamente artificial e não parece refletir como os seres humanos organizam suas idéias, somente a tendência estatística em textos. Uma proposta mais lógica seria calcular, ao longo do texto, curvas de domínio — textos coerentes não mudam de assunto abruptamente e isso é algo que se pode esperar de qualquer comunicação, usando linguagem natural, realizada com sucesso. O cálculo dos pontos pertencentes a essa curva deveria ser por meio de uma equação que incluísse um fator de amortecimento relacionado à distância, como, por exemplo:

$$R_{local}(t,j,D) = \sum_{i=0}^{comprimento(t)} \frac{W(t,j,D)}{1+(j-i)^2} \qquad (4)$$

A equação acima se refere à relevância local da palavra na posição j do texto t para o domínio D de $R_{local}(t,j,D)$ e o peso (calculado usando 3) de cada termo j do texto t em relação ao domínio D de W(t,j,D). Essa equação não tem base teórica nem foi testada empiricamente: é apenas uma possibilidade para calcular as tendências locais do texto, aproveitando as contribuições das palavras do contexto. Uma possível melhoria a ser adicionada nessa idéia é contabilizar, no termo de amortecimento, a presença de sinais gráficos como pontuação e quebras de parágrafo, que muitas vezes representam quebras no contexto semântico.

[4] No estudo de Recuperação de informação, para os resultados R de uma busca sobre uma base de dados D, define-se precisão como a proporção de documentos relevantes em R, e revocação como a proporção de documentos relevantes de D que estão presentes em R.

5. Conclusão

A evolução recente nos métodos de categorização de textos em linguagem natural seguiu o mesmo caminho que a dos sistemas de recuperação de informação, mas o passo seguinte que foi dado neste ainda não havia sido usado. Foi proposta, neste trabalho, a aplicação de uma versão da equação conhecida como tfc·nfx para calcular a pertinência de um texto a domínios. É necessário agora desenvolver um programa para verificar se os resultados usando tfc·nfx são realmente melhores, e para verificar o efeito das outras adaptações propostas na seção 4 – as mudanças relativas ao fator de amortecimento parecem particularmente promissoras.

Depois de verificada, através de testes, a eficácia dos métodos propostos e do aperfeiçoamento do fator de amortecimento, o plano é continuar o trabalho usando os desenvolvimentos das últimas décadas da disciplina da lingüística chamada Gramática Gerativa. Ela estuda métodos, baseados em gramáticas formais bem determinadas (e, portanto, diretamente transponíveis para linguagens de programação), para computar o papel semântico de cada elemento de uma frase, e as suas adaptações às diferentes linguagens naturais. Dado que as linguagens humanas não são agrupamentos de eventos aleatórios, mas reflexo da estrutura da mente humana, uma abordagem menos dependente de componentes probabilístico-estatísticos e mais estruturada parece promissora.

Referências

Brin, S. and Page, L. (1998) "The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems, 30(1-7) p. 107-117.

Gliozzo, A., Strapparava, C. and Dagan, I. (2004a) "Unsupervised and supervised exploitation of semantic domains in lexical disambiguation", Computer Speech and Language, 18(3) p. 275-299.

Gliozzo, A., Magnini, B. and Strapparava, C. (2004b) "Unsupervised domain relevance estimation for word sense disambiguation" In: Conference on Empirical Methods in Natural Language Processing.

Magnini, B. and Cavaglia, G. (2000) "Integrating subject field codes into WORDNET" In: LREC conference.

Magnini, B., Strapparava, C., Pezzulo, G. and Gliozzo, A. (2002a) "The role of domain information in word sense disambiguation", Natural Language Engineering, 8(3) p. 359-373.

Magnini, B., Strapparava, C., Pezzulo, G. and Gliozzo, A. (2002b) "Comparing ontology-based and corpus-based domain annotations in WORDNET", In: First International WORDNET Conference, p. 146–154.

Salton, G. and Buckley, C. (1988) "Term-weighting approaches in automatic text retrieval", Information Processing and Management: an International Journal, 24(5) p.513-523.

Schütze, H. (1988) "Automatic word sense discrimination", Computational Linguistics, 24(1) p.97-123.