Desambiguação Lexical de Sentido por meio de Medidas Topológicas de Textos

Ligia Spanó Nakano[1], Maria das Graças Volpe Nunes[2]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brasil

[2]Departamento de Ciências de Computação – Universidade de São Paulo
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brasil

liginhasn@gmail.com, gracan@icmc.usp.br

Abstract. *This paper describes a simple approach to the use of complex networks aimed at multilingual word sense disambiguation* (WSD). *The approach is based on English corpora containing texts from multiple sources with at least one occurrence of the word under review. The proposal is to verify whether there is any relationship between the measures of the complex network and the direction given by the analyzed word. In experiments with corpora containing only portions of text, a preliminary evaluation showed that certain measures of the sentence containing the word influenced those of the paragraph. Measures extracted from portions of text weren't enough, though, to discriminate the sense of the word it contained.*

Resumo. *Este artigo descreve uma abordagem sobre o uso de redes complexas aplicado a Desambiguação Lexical de Sentido* (DLS). *A abordagem é baseada em corpora em Inglês contendo textos obtidos de várias fontes com ao menos uma ocorrência da palavra analisada. A proposta consiste em verificar se existe alguma relação entre as medidas extraídas da rede complexa e o sentido assumido pela palavra analisada. Em experimentos com corpora formada por trechos de texto, uma análise preliminar mostrou que algumas medidas da sentença em que aparece a palavra influenciavam aquelas do parágrafo. Medidas extraídas dos trechos não foram suficientes, contudo, para descriminar o sentido da palavra que continham.*

1. Introdução

No atual contexto de globalização, a disseminação de informações é feita envolvendo-se várias línguas. A comunicação multilíngüe tornou-se, portanto, uma tarefa imperativa, que é muitas vezes auxiliada pela Tradução Automática (TA). Embora seja uma área de pesquisa antiga, a TA constitui uma aplicação bastante difícil de ser implementada computacionalmente. Um dos principais problemas apresentados pela TA é a ambigüidade lexical, ou seja, a dificuldade de escolha de uma palavra, na língua-alvo, para refletir o significado de outra, na língua-fonte, quando há várias opções de tradução. Assim, a desambiguação lexical de sentido (DLS) consiste na identificação do sentido mais adequado de uma palavra de conteúdo, dado o seu contexto, em função de uma lista pré-definida de sentidos, a qual pode variar de acordo com a aplicação.

Diversas abordagens têm sido propostas para a DLS, especialmente em contextos monolíngues e independentes de aplicação. Essas abordagens classificam-se em três principais grupos (Specia, 2007): abordagens baseadas em conhecimento lingüístico, manualmente ou semi-automaticamente especificado; abordagens baseadas em corpus, ou seja, baseadas em conhecimento superficial extraído de corpus de exemplos por meio de técnicas estatísticas ou de aprendizado de máquina, gerando modelos de desambiguação; e abordagens híbridas, que combinam características das outras duas abordagens para gerar modelos de desambiguação automaticamente a partir de corpus de exemplos e de conhecimento lingüístico / extralingüístico.

Este projeto busca investigar uma nova metodologia para a DLS através do uso de Redes Complexas (Costa et al., 2008). Representar um texto como uma Rede Complexa significa mapeá-lo como um grafo em que vértices representam as ocorrências lexicais de conteúdo (normalmente lematizadas, excluindo-se *stopwords*), e arestas ligam uma palavra a suas palavras vizinhas. Essas redes possuem algumas características topológicas não-triviais, distinguindo-se das redes comuns. A primeira delas é conhecida como *free-scale*, isto é, a rede tem poucos nós altamente conectados (grau alto) – chamados *hubs* – e muitos nós fracamente conectados (grau baixo). Além disso, novos nós adicionados tendem a se ligar aos *hubs*, fazendo com que a distância entre quaisquer dois nós seja relativamente pequena, dando-lhe a propriedade chamada *small-world*.

O uso de redes complexas como forma de representação textual já trouxe resultados promissores na área de Processamento de Línguas Naturais. Em Tradução Automática, demonstrou-se correlação de certas medidas das traduções modeladas como redes com a fonte da tradução, se humana ou automática (Amâncio et al., 2008). Traduções feitas manualmente geraram medidas topológicas similares entre si e distintas daquelas geradas tradutores automáticos (Amâncio et al., 2008). Em (Antiqueira et al., 2006), os autores mostram que algumas medidas das redes podem agrupar textos de um mesmo autor, numa tarefa de *clustering*. Já na área de Sumarização Automática, (Antiqueira et al., 2008) mostram como as redes podem ser úteis para sumarizar textos.

A DLS bilíngüe, ou seja, para a TA, consiste em determinar qual é o significado, entre vários possíveis, da ocorrência da palavra do texto-fonte, na língua-alvo. Por exemplo, na tradução do inglês para o português da sentença *She asked me money*, a tradução do verbo *asked* deve ser *pediu*, e não *perguntou*. Neste trabalho pretende-se verificar se há correlação entre medidas da rede que modela um texto onde ocorre uma palavra ambígua com seu significado. Em outras palavras, queremos responder às seguintes questões: Para uma dada palavra ambígua, as medidas dos textos em que ela aparece com igual significado são semelhantes? Para textos cujas ocorrências têm significado distinto, as medidas também são distintas?

2. Experimentos e Resultados Preliminares

Os experimentos foram realizados de forma a verificar a correlação de sentido de palavras ambíguas da língua inglesa, objetivando sua aplicação, no futuro, para a TA inglês-português. Foi utilizado o córpus Senseval-3 (disponível em http://www.senseval.org/senseval3) para o primeiro experimentento, contando com 287 trechos de texto com sentido anotado para a primeira palavra pesquisada. Nos primeiros experimentos, o verbo *ask* foi escolhido como um estudo de caso devido aos seguintes

critérios: alto número de ocorrências no corpus; distribuição razoavelmente uniforme dos sentidos e; distribuição similar das ocorrências dos diversos sentidos nos exemplos.

Para o segundo experimento, foram adotados 165 textos na íntegra extraídos da Internet. Os textos são de caráter jornalístico e foram encontrados com o auxílio da ferramenta de busca Google através do serviço Google News. Detalhes relativos a quantidade média de nós e arestas gerados para os textos extraídos de cada córpus são encontrados na tabela 1.

Tabela 1. Características referentes às redes geradas que modelam os textos para cada córpus

Córpus	**Fonte**	**Número total de textos**	**Número médio de vértices por texto**	**Número médio de arestas por texto**
1	Senseval-3	287	36	40
2	Diversas - internet	165	198	300

O primeiro experimento consistiu em modelar cada trecho de texto extraído do corpus onde havia uma ou mais ocorrências da palavra avaliada, com um determinado sentido. Para cada fragmento de texto, calcularam-se as seguintes medidas da rede correspondente: grau de entrada, grau de saída, coeficiente de aglomeração, caminho mínimo e desvio da dinâmica dos componentes (veja definição dessas medidas em Antiqueira et al., 2008). Observou-se, conforme era esperado, que os pequenos trechos do corpus eram insuficientes (muito pequenos) para se obterem medidas estáveis das respectivas redes. Ou seja, as faixas de valores das medidas, para cada sentido, intersectam-se ou sebrepõem-se em grande parte, independentemente da medida escolhida, impossibilitando qualquer conclusão a respeito da correlação[1].

O segundo experimento, ainda em andamento, considera corpora de textos maiores e completos, extraídos da Web, e construídos da seguinte forma: considerando-se as vizinhanças da palavra ambígua no corpus Senseval-3 (até 2 palavras à esquerda e à direita), para cada sentido, foram buscados textos em que a palavra ambígua ocorre nessas mesmas vizinhanças, sob a hipótese de que o significado seja afetado pelo contexto. Antes, porém, é necessário garantir que as medidas do parágrafo e do texto completo são de fato influenciadas pela medida da sentença onde ocorre a palavra. As figuras 1 e 2 mostram as medidas correspondentes a 2 vizinhanças distintas (16 textos cada), uma para cada sentido de *ask*, correspondentes à sentença com a ocorrência, e ao parágrafo que a contém. É possível constatar comportamento semelhante das medidas de sentenças e parágrafos: a curva referente aos valores das medidas para o parágrafo acompanha razoavelmente o formato da curva formada pelos mesmos valores para a sentença. Vale notar que Grau e Caminho Mínimo são as medidas que mais mostraram correlação com características textuais nos trabalhos da literatura.

[1] Gráficos omitidos por limitação de espaço.

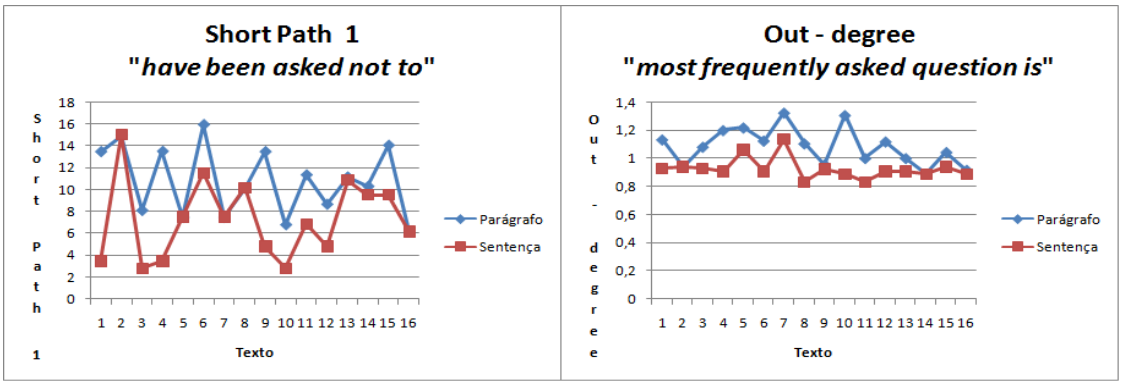

Figura 1: Valores médios da medida Caminho Mínimo da sentença e do parágrafo, para a vizinhança "*have been asked not to*" – sentido 1
Figura 2: Valores médios da medida Grau de Saída da sentença e do parágrafo, para a vizinhança "*most frequently asked question is*" – sentido 2

Atualmente, estão sendo compilados os corpora e computadas as medidas de seus textos para cada um dos diferentes sentidos de *ask*, de modo que em breve poderemos avaliar a existência de correlação entre medidas da rede e semântica lexical.

4. Conclusões e Trabalhos Futuros

Enquanto que o primeiro experimento permitiu descartar a hipótese de que medidas topológicas de fragmentos de texto possibilitam prever ou discriminar o sentido da ocorrência da palavra ambígua, os primeiros resultados do segundo experimento foram mais animadores. Os próximos passos incluem a extração de medidas para cada um dos corpora de diferente sentido, e a análise sobre os intervalos de valores para diferentes medidas, visando correlacionar alguma(s) delas com o sentido da palavra que originou os corpora.

5. Referências

Amancio, D. , Antiqueira, Lucas , Pardo, Thiago A. S. , Costa, Luciano F , Oliveria Jr, Osvaldo N. and Nunes, M. G. V. (2008) Complex networks analysis of manual and machine translations. International Journal of Modern Physics C, 19(4) April 2008.

Antiqueira, L ; Oliveira Jr, O.N; Costa, L.F; Nunes, M. G. V. (2008) A Complex Network Approach to Text Summarization. Information Sciences, p. INS 8125-52, 2008.

Antiqueira, L ; Pardo, T.S.; Nunes, M. G. V.; Oliveira Jr, O.N; Costa, L.F; (2006) Some issues on complex networks for author characterization. Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Information and Human Language Technology (TIL'2006), Ribeirão Preto, Brazil, October 23–28, 2006. CD-ROM. ISBN 85-87837-11-7

Costa, L. F.; Oliveira Jr., O. N.; Travieso, G.; Rodrigues, F. A.; Villas Boas, P. R.; Antiqueira, L.; Viana, M. P.; Rocha, L. E. C. Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Physics and Society, 2008.

Specia, L. (2007) Uma Abordagem Híbrida Relacional na Desambiguação Lexical de Sentido na Tradução Automática. Tese de Doutorado, 244p. ICMC-USP.