Linguagem Natural como Rede de Mundo Pequeno: analisando redes lexicais em um texto jornalístico extraído de um *corpus* de desastres naturais

Ariana Moura da Silva, Margarethe Born Steinberger-Elias

Mestrado em Engenharia da Informação - Universidade Federal do ABC (UFABC)
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas

{ariana.silva,mborn}@ufabc.edu.br

Abstract*. This paper aims to explore the small-world network concept (Watts & Strogatz, 1998) applied to natural language networks as described by Cancho & Solé, 2001. The search of an algorithm to generate patterns of word association conducted this exploratory study based on a single text in Portuguese. The long-term goal to be achieved is to find out how lexical-semantic associations behave in a limited domain (climate changes) and to generate an objective way to measure degrees of semantic similarity between words that co-occur in a text, assuming they can be treated as nodes of a network where co-occurrence measures and distance between nodes can function as similarity indicators to model lexical networks.*

Resumo. *Este trabalho inspira-se em Cancho & Solé (2001) sobre padrões de associação entre as palavras de uma língua a partir do conceito de redes de mundo pequeno (Watts & Strogatz, 1998). Calcular as conexões mais prováveis em âmbito tão amplo levou a este estudo exploratório baseado em um único texto em língua portuguesa, buscando verificar como se comportam associações léxico-semânticas em domínio limitado (mudanças climáticas). O objetivo é gerar medição objetiva do grau de similaridade semântica entre palavras que co-ocorrem em um texto, assumindo que elas podem ser tratadas como vértices de rede onde medidas de co-ocorrência e de distância entre nós sejam indicadores de similaridade, permitindo modelar redes lexicais.*

1. Introdução

O fenômeno do mundo pequeno baseia-se no principio ao qual qualquer nó de uma rede pode estar conectado a qualquer outro nó aleatoriamente em uma média de apenas alguns passos, independente do tamanho da rede (Watts & Strogatz, 1998).

Baseado nessa premissa pode-se utilizar o mesmo conceito aplicado por Watts em redes sociais em redes lexicais de palavras. Em redes sociais os indivíduos possuem *identidades sociais,* do mesmo modo, em redes lexicais pode-se dizer que as palavras possuem *papéis semânticos* relevantes que podem diferenciá-las uma das outras, bem como classificar as *similaridades semânticas*.

Na Ciência de Redes, o conceito de redes de mundo pequeno (Watts & Strogatz, 1998) revelou padrões de interação entre indivíduos observando dois métodos estatísticos: o agrupamento do coeficiente (número médio de ligações por palavra) e o comprimento do caminho (à distância em nós que liga duas palavras). Seguindo esta teoria e experimentação em redes de interações sociais, o problema de pesquisa estende-

se a partir do mesmo conceito, porém, seguindo a proposta de Steinberger (2010), a aplicação será testada e demonstrada utilizando redes de linguagem natural. Neste caso específico em estudo, uma rede lexical onde as palavras serão os nós, as arestas serão as ligações semânticas entre as palavras.

A distância entre os nós será representada por similaridade semântica de palavras dentro do mesmo domínio de desastres naturais. Por exemplo, se analisadas isoladamente, as palavras "terremoto" e "tempestades" podem ter significados diferentes. Associadas a um mesmo texto podem desempenhar papéis semânticos similares (Saussure, 1966), que podem ser categorizados e incluídos em tipologias de desastres naturais.

2. Metodologia

Como objeto de trabalho escolheu-se o texto "ONU: aumentam os desastres naturais relacionados às mudanças climáticas" (EFE, 2008), com 11 parágrafos, 552 palavras e 2.891 caracteres, foi dividido em 16 sentenças. As sentenças foram enumeradas obedecendo a uma ordem crescente e organizadas da seguinte forma: S1, S2, S3 e assim por diante. Onde as palavras que fazem parte da S1 são exatamente as palavras que encontram-se na primeira frase do texto, e as palavras que fazem parte da S2 são exatamente as palavras que encontram-se na segunda frase do texto e assim consecutivamente.

No texto todas as palavras foram consideradas individualmente, ou seja, considerou-se apenas a morfologia de palavras isoladas. As únicas palavras morfologicamente compostas consideradas foram: Nações Unidas, Desastres Naturais e Catástrofes Naturais. Isto de seu, devido à relevância semântica das palavras para o estudo proposto. No que se diz respeito à sintática das palavras não foram alteradas nenhuma palavra, manteve-se a ordem original do texto extraído do *corpus*. Posteriormente será mostrado como as palavras foram organizadas por frequência, semântica, lembrando que mesmo realizando o agrupamento por similaridade semântica, as palavras não foram modificadas em sua ordem sintática no texto. A análise não se preocupou em verificar as variações dos morfemas gramaticais, morfemas lexicais, desambiguação, lematização, desinências nominais e verbais. Ou seja, as palavras foram analisadas em sua forma original encontradas no texto e apenas agrupadas por similaridade e semântica.

A segmentação das sentenças e agrupamentos semânticos das palavras foram realizados manualmente, por se tratar de um único texto, inseridas em colunas e referenciadas a qual grupo de sentença cada palavra iria pertencer. Para contagem da frequência de palavras foi utilizado o banco de dados Microsoft Access 2007 e a linguagem T-SQL. Para o desenvolvimento das redes e geração das figuras utilizou-se o software Microsoft Excel com a extensão do aplicativo NodeXL.

Utilizou-se um modelo misto formado por Métodos Estatísticos (análise quantitativa de correlação, o agrupamento do coeficiente, o comprimento do caminho (onde cada frase é representada por uma sentença / unidade de medida) (Cancho & Solé, 2001) e Métodos Semânticos (categorias de níveis (Rosch, 1975)).

2.1. Tratamento Estatístico: Criação da lista de frequência

Estatisticamente, analisando o conceito de freqüência pode-se constatar que a palavra "inundações" aparece no texto 5 vezes, sendo a palavra de maior frequência dentro do contexto de desastres naturais. Abaixo a figura 1 mostra as ligações entre as palavras, que podem ser chamadas de laços (*links*), e os pontos são os nós da rede.

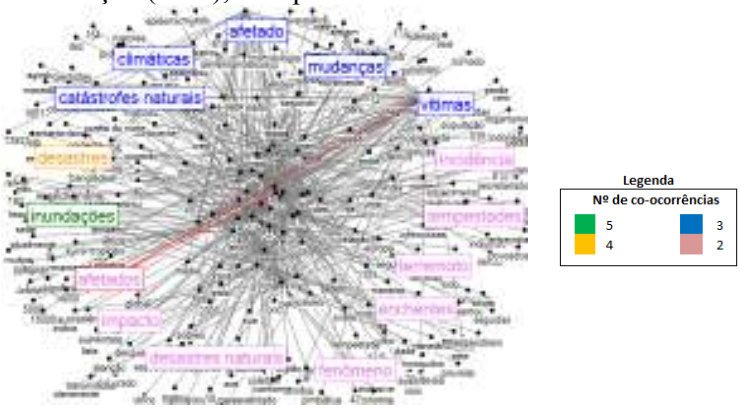

Figura 1. Rede de palavras coocorrentes.

2.2. Tratamento Semântico

O texto de desastres naturais abordado representa apenas uma amostra do *corpus* a ser analisado. Foram consideradas as palavras co-ocorrentes no tratamento estatístico mais a similaridade semântica que elas possuem com o tema do texto. A palavra "inundações" que teve uma freqüência de número 5 e aparece nas sentenças S2, S3, S4, S11 e S12 possui um grau de distribuição alto (Cancho & Solé, 2001), comparado com uma palavra de freqüência de número 2, por exemplo, "enchentes" que aparece na S1 e S13. Sendo palavras de mesma similaridade semântica poderiam ser relevantes dentro da mesma categoria, possibilitando então um maior grau de distribuição de frequência (Watts & Strogatz 1998) e diminuindo então a distância entre as sentenças.

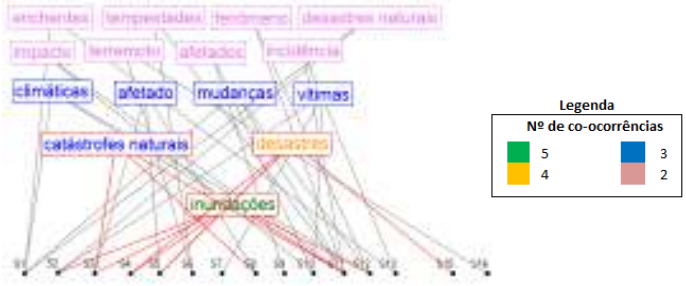

Figura 2. Distância entre as sentenças.

3. Resultados e Discussões

Utilizando os modelos de redes de mundo pequeno conjuntamente com medidas de grau de similaridade entre as palavras pode-se concluir que o grau de freqüência entre as palavras aumenta, fazendo com que a distância entre elas diminua. Porém, quanto mais distantes forem as palavras, maior será o comprimento do caminho, porém esta medida não está relacionada com a de similaridade semântica. Exemplo: a palavra composta "desastres naturais" possui uma freqüência 2 baixa e aparece apenas na S1 e S2. Já a

palavra “desastres” aparece na sentença S4, S5 e S15. Ou seja, as distâncias entre as palavras são maiores o que não irá influenciar em suas similaridades semânticas.

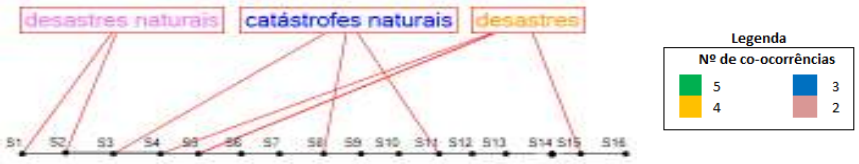

Figura 3. Distância entre as palavras de mesma similaridade semântica onde a frequência é maior que 2

A análise da Figura 3 mostra que as palavras não se repetem dentro de uma mesma unidade de medida. Elas possuem frequências 2, 3 e 4. Verifica-se que a expressão “desastres naturais” possui um grau 0 de distância entre as sentenças 1 e 2, grau 1 de distância com a expressão “catástrofes naturais” e grau 2 com a palavra “desastres”. Enquanto que “catástrofes naturais” possui grau 1 de distância tanto com “desastres naturais” como também com a palavra “desastres”, tendo como ponto de análise S3.

4. Conclusões

Os resultados do experimento mostraram que o comprimento do caminho entre as palavras e o grau de distância entre elas não esta relacionado diretamente com o conceito de similaridade semântica. Mas para a utilização de métodos estatísticos a utilização do grau de distância de mostra eficaz, provando o conceito de Redes de Mundo Pequeno, onde qualquer ponto da rede pode estar ligado a qualquer outro ponto da rede em pequenos passos. Quanto maior o grau de frequência das palavras de mesmo significado semântico menor será o comprimento do caminho entre elas.

5. Referências bibliográficas

Cancho, R. F. i., Solé, R. V. The small world of human language. Proceedings of the Royal Society of London. Series B, Biological Sciences 268, 2001. p.2261-2266.

Efe. ONU: Aumentam os desastres naturais relacionados às mudanças climáticas. Jornal O Globo, publicado em 18 Jan. 2008. Disponível em: <http://oglobo.globo.com/ciencia/mat/2008/01/18/onu_aumentam_os_desastres_naturais_relacionados_as_mudancas_climaticas-328104052.asp>. Acesso em 02 Out. 2011.

Rosch, E.H., Mervis, C.B. Family resemblances: Studies in the internal structure of categories. Cognitive Psychology 7, 1975. p.573-605.

Saussure, F.(1916/1966). “Cour de Linguistic Generale” (Course in General Linguistics), (W.Baskin, Trans.) New York: McGraw-Hill apud Echtner.

Steinberger, M. B. Estudo das Condições de Produção e Circulação de Relatos sobre Desastres e Catástrofes na América Latina. Anais Colóquio Internacional de Ciências da Comunicação Brasil-EUA, Caxias do Sul, RS, 2010.

Watts, D. J., Strogatz, S. H. Collective dynamics of small-world networks. Nature 393, 1998. p.440-442.

Watts, D. J. Seis Graus de Separação. A evolução da ciência de redes em uma era conectada [Tradução André Alonso Machado]. São Paulo: Leopardo, 2009.