MAPEANDO REDES LÉXICO-SEMÂNTICAS EM UM CORPUS SOBRE DESASTRES NATURAIS NO ESTADO DE SÃO PAULO

Thiago da Rocha Tedrus[1] e Margarethe Born Steinberger[1]

[1] Universidade Federal do ABC – UFABC - Santo André – SP

Mestrado em Engenharia da Informação

Área de concentração: Sistemas Inteligentes

Linha de Pesquisa: Inteligência Social

Pesquisa financiada com bolsa Capes iniciada em: 03/2012

Previsão de defesa da Dissertação: 03/2014

Este trabalho inscreve-se numa pesquisa mais ampla cujo objetivo geral é investigar métodos de recuperação de informação sobre desastres naturais em um corpus de textos jornalísticos em língua portuguesa. O objetivo específico deste trabalho é construir uma modelagem léxico-semântica no domínio dos desastres naturais capaz de classificar automaticamente as ocorrências quanto à região e ao tipo de ambiente onde tiveram lugar no estado de São Paulo, conforme categorias de Marcelino (2008). A hipótese de partida é a de que o mapeamento de redes léxico-semânticas a partir do vocabulário extraído dos textos noticiosos será capaz de permitir reconhecer e diferenciar, por exemplo, os desastres ocorridos por excesso de chuva e bloqueio das vias de escoamento natural da água que podem estar associados a entupimento de bueiros ou aqueles decorrentes da falta de absorção da água devido à impermeabilização do solo por asfalto ou concreto. A pesquisa irá mobilizar métodos e técnicas da Linguística Computacional (LC) e do Processamento de Línguas Naturais (PLN) que, com auxílio de recursos computacionais, permitem modelar representações semânticas de um dado domínio de conhecimento e recuperar automaticamente a informação (Lyons, 2009). Utilizando a teoria e especificações propostas pela Linguística de Corpus (Berber-Sardinha, 2004), foram coletadas 1.139 notícias referentes aos desastres naturais envolvendo água, ocorridos no estado de São Paulo, no período compreendido entre 01/01/2002 a 22/03/2012. Estas notícias foram disponibilizadas pelo jornal *Folha de S. Paulo*, em sua versão digital, disponível online. Os critérios para a escolha do corpus foram bem definidos para melhor caracterizar e delimitar a situação-limite do domínio abordado no estudo. De acordo com a orientação de Berber-Sardinha (2004), o corpus

foi armazenado em arquivos de extensão (txt), legíveis para computadores e ferramentas da análise linguística. Numa próxima etapa do trabalho, serão extraídos dados com o auxílio da ferramenta computacional *WordSmith*. Seguindo a proposta de Steinberger (2009), a modelagem linguística da informação presente no corpus pode ser baseada em processos de extração e modelagem de redes léxico-semânticas tomando como referência não só aspectos qualitativos, tais como as especificidades do léxico utilizado para veicular informação em corpora selecionados no domínio de interesse, como também aspectos quantitativos, tais como o mapeamento das probabilidades de co-ocorrência entre itens lexicais específicos. A análise das redes léxico-semânticas permite detectar, por exemplo, quais regiões do estado de São Paulo são mais provavelmente associadas a quais tipos de desastres e em que período temporal eles geralmente acontecem. Ou, também, quais zonas da cidade de São Paulo são mais provavelmente associadas a quais tipos de desastres e a quais condições típicas do meio urbano. Ou seja, com os recursos da modelagem linguística (léxico-semântica), é possível extrair correlações entre informações do corpus associadas através das redes, de modo a permitir construir diagnósticos, predizer comportamentos e definir caminhos para criar planos estratégicos capazes de lidar com os desastres estudados. O mapeamento de redes léxico-semânticas em um corpus de textos jornalísticos sobre desastres naturais no estado de São Paulo também poderá alimentar uma base de dados que permita recuperar automaticamente informação em situações de emergência.

REFERÊNCIAS BIBLIOGRÁFICAS

Berber-Sardinha, T. **Linguística de corpus**. Barueri, SP: Manole, 2004.

Lyons, J. **Linguagem e Linguística: uma introdução**. Tradução Marilda Winkler Averburg, Clarisse Sieckenius de Souza. Rio de Janeiro:LTC, 2009.

Marcelino, E. V. 2008. **Desastres Naturais e Geotecnologias: Conceitos Básicos**. Caderno Didático nº 1. INPE/CRS, Santa Maria, 2008.

Steinberger, Margarethe Born. **Modelagem linguística como recurso de análise em gestão de conhecimento**. UFABC, Santo André, SP, 2009: 15p.