Criando um corpus sobre desastres climáticos com apoio da ferramenta NLTK

Rafael Antonangelo Molina[1], Margarethe Born Steinberger-Elias[2]

[1,2] Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas (CECS) – Universidade Federal do ABC (UFABC)
CEP - 09.210-170 – Santo André – SP – Brasil

{rafael.molina, mborn}@ufabc.edu.br

Abstract. *This work is part of a broader research that explores information from a corpus of news about climate disasters and automatically recognizes, with the support of a tool for Natural Language Processing (NLP), words that denote the main actors involved and their actions in providing relief to victims. It starts with the hypothesis of Steinberger [2005] that news reports of disasters not only allow us to identify entities that participate in aid, but also provide conditions to characterize discursive networks associated with each type of event. This paper presents the stages of composition and description of a corpus about the earthquake in Haiti in 2010 with support from the Natural Language Toolkit (NLTK).*

Resumo. *Este trabalho integra uma ampla pesquisa que explora informações de um corpus de notícias sobre desastres climáticos e reconhece automaticamente, com apoio de uma ferramenta de Processamento de Linguagem Natural (PLN), palavras que denotem os atores envolvidos e suas principais ações na prestação de socorro às vítimas. Parte-se da hipótese de Steinberger [2005] de que relatos noticiosos de desastres não só permitem identificar as entidades que participam dos socorros, como também oferecem condições para caracterizar redes discursivas associadas a cada tipo de evento. Neste artigo, são apresentadas as etapas de composição e descrição de um corpus sobre o terremoto do Haiti em 2010 com apoio do pacote Natural Language Toolkit (NLTK).*

1. Introdução

O terremoto ocorrido no Haiti em 12 de janeiro de 2010 demonstra a necessidade de superação dos problemas apontados por um relatório do Programa das Nações Unidas para Desenvolvimento quanto à integração de formas de representação de informações em desastres naturais na América Latina [UNDP 2004]. Sistemas de provisão adequada de informação em situações de catástrofes associam-se a competências organizacionais através de conhecimento de natureza preditiva sobre ações e comportamentos. Ao mesmo tempo, formas de organização do trabalho improvisadas surgem à medida que as necessidades aparecem, revelando-se úteis e alternando-se com as formas mais estruturadas de ação que seguem um planejamento prévio. Busca-se neste *paper* relatar os primeiros passos para a composição de um corpus que permita a identificação de entidades assistenciais atuantes nestes eventos por meio de uma rede discursiva

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 194–198,
Cuiabá, MT, Brazil, October 24–26, 2011.

[Steinberger 2005]. O tratamento automático de informação em linguagem natural em domínio específico tem suas raízes lingüísticas, por exemplo, em Sinclair [1990] e, no Brasil, com a Lingüística de Corpus [Sardinha 2004]. Já a literatura sobre desastres climáticos não traz registros de trabalhos baseados em um corpus de relatos de desastres climáticos em Português tal como proposto aqui.

2. Método

PLN é um campo de pesquisa interdisciplinar que reúne competências da Lingüística e da Informática na aplicação de algoritmos de análise e geração de textos em um determinado idioma (língua natural) com apoio de ferramentas computacionais [Bird et. al. 2009]. A Lingüística Computacional é a parte da ciência lingüística que se preocupa com o tratamento computacional da linguagem natural [Steinberger 2010]. Dentre os métodos de Lingüística Computacional, Steinberger [2009] propõe a aplicação da Lingüística de Corpus ao estudo de relatos de desastres. A Lingüística de Corpus ocupa-se da coleta e exploração de corpus/corpora, ou conjuntos de dados lingüísticos textuais tratados com rigor e geralmente em grande escala, com o propósito de servirem para investigar uma língua ou variedade lingüística [Sardinha 2004]. Com o uso destes recursos é possível mapear linguisticamente domínios de conhecimento e modelos de uso desse conhecimento para fins específicos, uma modelagem lingüística que promova uma investigação única sobre associações e preferências de linguagem e mesmo sobre a indução de conhecimento por meio desta [Manning et. al. 2000].

O pacote NLTK realiza o processamento de linguagem natural em Python (linguagem de programação de excelente funcionalidade para processamento de dados lingüísticos) [Bird et. al. 2009]. NLTK foi concebido em 2001 como parte de um curso de Lingüística Computacional no Departamento de Ciência da Computação e Informação da Universidade da Pensilvânia [Bird et. al. 2009]. Devido a seu caráter de software aberto e gratuito, tem sido desenvolvido e ampliado com a ajuda de dezenas de colaboradores, pelo seu uso e concepção de módulos de análise lingüística [Bird et. al. 2009]. O estudo das funcionalidades deste pacote pela literatura de Bird et. al [2009] e de aplicações de PLN sobre análises de Manning et. al. [2000] permitiu as primeiras execuções com o corpus de notícias sobre o terremoto do Haiti.

Os resultados referem-se à composição do corpus, executando-se a organização, os comandos para reconhecimento de textos e conversão de palavras em listas, distribuições de freqüência e única *string* (para a apresentação codificada para o Português). Promoveu-se a aplicação de filtros (comandos de eliminação de determinados padrões no corpus) em lista de palavras (contém todo o corpus), seguido de trabalhos com recortes das primeiras 100 expressões em ocorrência dentro de cada um dos diferentes filtros. Depois isto, categorizou-se as 100 primeiras ocorrências dentro dos dois filtros apresentados, com o fim de delimitar elementos com alto valor semântico no corpus e potenciais referências a entidades assistenciais. O comando de concordância, bem como a familiaridade com o conteúdo do corpus ajudaram a determinar as categorias com seus elementos. Através de comandos de contabilização de itens e contagem de expressões foi possível levantar dados de freqüência de aparecimento no texto dos itens lexicais que compõem *collocations*, calcular probabilidades e medidas estatísticas de validação de forma a demonstrar que os retornos obtidos representam associações não aleatórias dos itens, tais como a razão

Observado/Esperado (O/E), Informação Mútua (I) e o Escore T (T), além do Intervalo Médio (IM) de ocorrência de palavras no corpus estudado. O valor de O/E deve ser interpretado como quantas vezes um valor é maior que o esperado probabilisticamente, enquanto um valor de I maior que 3 e/ou de T maior que 2 indicam associações não aleatórias entre palavras [Sardinha 2004].

3. Resultados

O corpus adotado nesta pesquisa constitui-se de textos extraídos da Folha de S. Paulo no período de 12/01/2010 a 12/02/2011 para a busca "Haiti" [Folha de S. Paulo]. Foram levantados 842 textos noticiosos sobre o terremoto ocorrido no início de 2010, sendo que o corpo de texto foi salvo em formato txt (compatível com o pacote NLTK apresentado na seção 2) e os metadados de Identidade do Evento (ID), Identidade Numérica (Nº), Data, Título da Matéria, Subtítulo, Link, Instituição, Autoria, Seção, Local, Figura e Legenda foram organizados em uma planilha eletrônica. Além disto, foi realizada uma classificação dos textos para filtrar links que a busca retornou e que, no entanto, não eram pertinentes ou eram apenas parcialmente pertinentes para o tema pesquisado. Também foram filtrados textos com referência a outras catástrofes.

Procurou-se descrever o corpus quanto a características gerais: número total de ocorrências acumuladas (*tokens*), total de ocorrências exclusivas (*types*), expressões sem diferenciação de maiúscula (332.397), caracteres (2.244.234) e sentenças (19.802). A tabela 1 apresenta os filtros aplicados, o número total de ocorrências (em *types* e *tokens*), densidade lexical (*tokens/types*), porcentagem de ocorrências acumuladas das 100 primeiras expressões com relação ao todo e a quantidade de palavras acrescida à lista de 100 mais ocorrentes com relação ao filtro anterior.

Tabela 1. Dados de aplicações de filtros sobre 100 primeiras ocorrências em freqüência

Filtro	Nº de *tokens*	Nº de *types*	Densidade lexical	Representação do recorte sobre o total	Acréscimo de novas palavras
Nenhum	429135	34788	12,34	48,53%	-
Eliminando stopwords (Filtro1)	307559	34643	8,88	34,52%	39
(Filtro1) + tomando somente alfabéticos (Filtro 2)	234355	33456	7,00	21,24%	21
Maiúscula sem estar após "." (Filtro 3)	46712	8425	5,54	36,25%	-
(Filtro 3) + sem stopwords (Filtro 4)	42394	8328	5,09	34,32%	16

A aplicação do Filtro 2 permite a visualização de um perfil de itens lexicais que possuem alto conteúdo semântico (como substantivos e verbos, por exemplo). Levando em conta que dentro dos resultados deste filtro é que se encontram as informações buscadas, os resultados apontam que pouco mais que 1/5 destas (21,24%) se concentra nas 100 primeiras ocorrências, categorizadas em: localização, personalidades de atuação política, entidades, referências a mídia, referências temporais, desastre, ações e estados e outros itens que passaram pelo filtro por serem grafados com letra maiúscula ou não estarem inclusos no filtro de *stopwords*. Filtros que selecionam maiúsculas que não estejam após ponto deram retornos válidos em termos de identificação de entidades assistenciais. No Filtro 4 os resultados podem ser categorizados em localização, personalidades de atuação política, entidades, referências a mídia, referências temporais, ajuda e outros itens que passaram pelo filtro por possuírem letra maiúscula.

Na aplicação de comandos na busca por *collocations* (A+B) dentro do corpus sem a aplicação de nenhum filtro houve o retorno de 20 resultados (*default* do pacote NLTK). A tabela 2 apresenta os cinco primeiros retornos e suas validações, de forma a ilustrar os resultados obtidos nesta etapa.

Tabela 2. Dados de aplicações sobre freqüências de *collocations*

Collocations	**f(A|B)**	**f(A)**	**f(B)**	**IM**	**O/E**	**I**	**T**
Porto Príncipe	382	406	386	1123,3901	1046,03	10,0307079	19,52614
São Paulo	331	440	393	1296,4804	821,4416	9,68201424	18,17126
Nações Unidas	98	107	99	4378,9286	3970,096	11,9549583	9,897001
dos EUA	231	1823	624	1857,7273	87,14343	6,44532002	15,02427
Estados Unidos	95	157	98	4517,2105	2649,67	11,3715971	9,743116

4. Discussão

A descrição do corpus permitiu ver sua distribuição, o que deu margem para entender que a abordagem já exposta de Steinberger [2005 e 2009] poderia ser útil para obter uma rede discursiva sobre o evento estudado. No trabalho com filtros, observa-se a manutenção de categorias com a mudança de filtro, porém a quantidade de expressões em cada categoria muda muito, apontando para a validade do uso dos filtros na identificação de determinados valores semânticos e, conseqüentemente, para a identificação de etiquetas semânticas que vão compor a rede pretendida. Isso dá margem ao uso de filtros na impressão em arquivo de etiquetas léxico-gramaticais (distribuição de freqüência) e semânticas (categorias). As palavras tendem a ter uma distribuição de freqüência mais uniforme após a aplicação dos filtros (como mostra a medida de representatividade percentual na tabela 1). Segundo a Lei de Zipf, palavras com freqüências mais baixas possuem significado mais bem definido que o contrário [Manning et. al. 2000]. Assim, a aplicação de filtros para determinar quais palavras possuem alto valor semântico é validada. Também o alto número de entidades resgatadas pelo filtro com este fim é um bom resultado (somente 11 de 100 elementos categorizados não correspondem a entidades segundo Bird et. al [2009]). O trabalho com *collocations* foi realizado de forma a definir discretamente (nós) o léxico que concentra propriedade e significado próprios (desde que validado quantitativamente). Se categorizadas, apontam para a persistência das categorias apresentadas no estudo de filtros e demonstram indícios de que estas categorias podem ser diferenciadas quantitativamente pelos índices de validação, como O/E < 800 para a categoria de "tempo e outros", por exemplo. Todas as *collocations* foram validadas pelos índices com este propósito e deverão ser tratadas, na etiquetagem, como elementos únicos.

5. Considerações Finais

O trabalho mostrou etapas de composição e descrição de um corpus de textos sobre o terremoto do Haiti com apoio do pacote NLTK. O uso do pacote foi bem sucedido, com retornos em freqüência de itens lexicais, *collocations*, aplicação de filtros para gerar listas com propriedades pré-determinadas léxico-gramaticalmente, manipulação do corpus para obter dados que permitissem validar estatisticamente retornos trazidos pelo pacote e análise de itens e combinações de itens em categorias. Numa próxima etapa serão executados comandos de etiquetagem permitindo aplicação de filtros para capturar itens lexicais segundo seu perfil morfossintático, gramatical e semântico.

6. Referências Bibliográficas

Bird, Steven; Klein, Ewan; Loper, Edward. (2009) "Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit". 479 p. 1ª Edição. Sebastopol, CA: O'Reilly.

Folha de S. Paulo. "Busca - Haiti". Disponível em: http://search.folha.com.br/search?q=haiti&site=jornal&sd=12%2F01%2F2010&ed=12%2F02%2F2011. Acessado em 05 de outubro de 2010.

Manning, Christopher D.; Schutze, Hinrich. (2000) "Natural Foundations of statistical natural language processing". Massachusetts Institute of Technology, 680 p., 2ª Edição. Cambridge, MA: The MIT Press.

Molina, Rafael Antonangelo; Steinberger, Margarethe Born. (2009) "Modelagem lingüística de informação em revistas técnicas setorializadas". Relatório Final de Iniciação Científica pelo Programa PIC/PIBIC. 145 p. Apresentação no II Simpósio de Iniciação Científica da Universidade Federal do ABC. Santo André, SP: UFABC.

Natural Language Toolkit - NLTK. Disponível em: http://www.nltk.org/. Acessado em 02 de fevereiro de 2010.

Sardinha, Tony Berber. (2004) "Lingüística de Corpus", p. 1- 45, 200-209. Barueri, SP: Manole.

Sinclair, John. (1991) "Corpus, Concordance, Collocation", 179 p. Oxford, OXON: Oxford Univ.Press.

Steinberger, Margarethe. (2005) "Discursos Geopolíticos da Mídia: jornalismo e imaginário internacional na América Latina". 310 p. São Paulo, SP: Cortez e Fapesp.

Steinberger, Margarethe. (2009) "Modelagem lingüística como recurso de análise em Gestão de Conhecimento", 15 p. Santo André, SP: UFABC.

Steinberger, Margarethe. (2010) "Estudo sobre as Condições de Produção de Relatos de Catástrofes e Desastres na América Latina". 16 p. Anais do IV Colóquio Brasil-EUA de Ciências da Comunicação. Caxias do Sul, RS: Intercom. Disponível em: http://www.intercom.org.br/papers/nacionais/2010/resumos/R5-3369-1.pdf. Acessado em 29 de setembro de 2010.

United Nations Development Programme - UNDP. (2004) "Reducing disaster risk: a challenge for development, a global report". UNDP Bureau for Crisis Prevention and Recovery, p. 43, 52. New York, NY: UNDP.