Construção de hierarquia de temas e subtemas de texto

Marco Gonzalez, Ana Martins, Fernanda Barão, Marcéu Leite, Vera L. S. de Lima

PUCRS - Faculdade de Informática
Av.Ipiranga, 6681 – Prédio 16 - PPGCC
90619-900 Porto Alegre, Brasil
{gonzalez, vera} @inf.pucrs.br

Abstract. *This paper presents a contribution to identification of themes and subthemes of Portuguese texts. We build hierarchical structures with terms extracted from a text, trying to define the order of their relative importance and the way how they group. We use techniques, such as stemming, extraction of lexical relations, and frequency weighting; as well as techniques from algorithms and data structures, such as construction of maximum spanning trees.*

Resumo. *Este artigo apresenta uma contribuição para a identificação dos temas e subtemas de um texto. Para tanto, são construídas estruturas hierárquicas com os termos extraídos do texto analisado, procurando definir a ordem de importância relativa dos mesmos e como se agrupam. São utilizadas técnicas de processamento da linguagem natural como stemming, captura de relações lexicais e cálculo de freqüência de ocorrência de termos; assim como técnicas de algoritmos e estrutura de dados, como construção de árvores geradoras máximas.*

Palavras-chave*: processamento da linguagem natural, identificação de temas e subtemas.*

1. Introdução

A expressão “representação de texto” é usada para abordar questões de representação do conteúdo de textos visando processamento automático. Nesse sentido, a aplicação de técnicas de processamento da linguagem natural visa obter a transformação das estruturas lingüísticas do texto em representações de seus possíveis temas.

Um tema, neste artigo, é definido como uma proposição tratada ou discutida em um texto. Nosso objetivo, além da identificação de temas e subtemas, consiste na definição da hierarquia dos mesmos em um texto. Após serem capturadas as relações hierárquicas na forma de um grafo valorado, são construídas árvores geradoras máximas a partir dele. Para tanto, são selecionados os termos, em ordem decrescente de seus pesos, e as relações mais freqüentes. As árvores assim construídas representam os temas principais e mostram a hierarquia entre os subtemas.

Este artigo se organiza nas seguintes seções: a seção 2 descreve a estratégia adotada e explica a construção das árvores geradoras máximas; a seção 3 apresenta um estudo de caso; a seção 4 tece algumas considerações sobre o trabalho realizado.

2. Estratégia adotada

Na Figura 1 é apresentada a estratégia adotada para a construção da hierarquia de temas e subtemas de um texto. A partir do texto, é obtida uma floresta **T** constituída de árvores onde é estabelecida a hierarquia. Floresta é um conjunto de estruturas de

dados hierárquicas (árvores) não conectadas entre si, ou seja, é um grafo sem ciclos (sem caminhos que permitam partir de um nó e chegar a ele mesmo após percorrer outros nós) [Goodrich2002].

Os termos, na árvore, são representados na forma de *stems* (parte essencial do item lexical, semelhante ao radical, obtida por processo de normalização morfológica, denominado *stemming*), pois o conceito que o tema ou o subtema identifica pode ter origem em itens lexicais de categorias morfológicas distintas no texto.

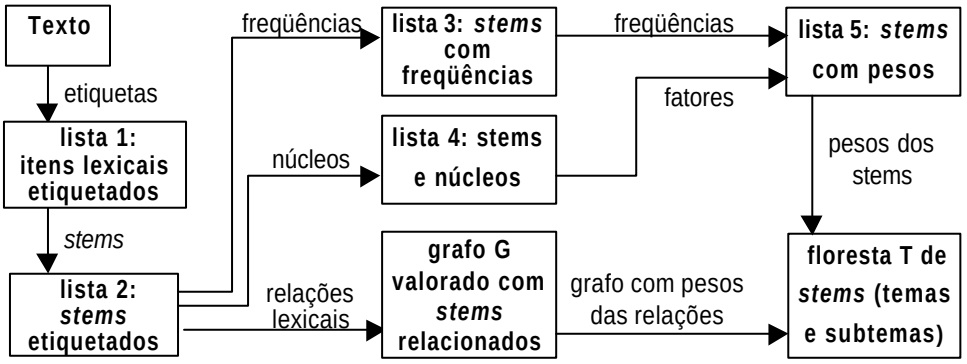

Figura 1. Estratégia de construção da hierarquia de temas e subtemas

A obtenção dos temas e subtemas é realizada através dos procedimentos de etiquetagem, normalização, identificação de relações lexicais, definição de peso dos termos e construção da floresta T, expostos a seguir, sempre relacionados à Figura 1.

A etiquetagem morfológica produz a lista **1** (lista de itens lexicais etiquetados, na ordem em que ocorrem no texto). A normalização (*stemming* e cálculo da freqüência de ocorrência dos *stems*) produz a lista **2** (*stems* etiquetados, na ordem do texto) e a lista **3** (*stems*, em ordem alfabética e sem duplicatas, com freqüência de ocorrência).

A identificação de relações lexicais da lista **2** é feita através da ferramenta Rellex (http://www.inf.pucrs.br/~gonzalez/ri/rellex). São produzidos o grafo valorado **G** (*stems* como nós e relações lexicais como arcos) e a lista **4** (atualização da lista **2**, com núcleos de sentenças: sujeito, verbo principal e objeto direto). Os pesos dos termos são ponderados conforme sua ocorrência como núcleos: sujeito (fator 9) e objeto direto (fator 4). O peso atribuído nesses casos se encontra ainda em análise.

```
Enquanto houver stems na lista 5 faça
      Selecionar, na lista 5, o stem S de maior peso ainda não selecionado
      Selecionar, em G, a relação R (onde S ocorre) de maior peso ainda não
            selecionada
            // R pode ser (S,S') ou (S',S), onde S tem peso maior que S' na
                  lista 5, senão S' já teria sido selecionado e R também.
                  Logo, em T, S é superior a S'
      Incluir em T os nós de R da seguinte maneira:
            Se S já está em T e S' não, então incluir S' como filho de S
            Se nenhum dos nós de R está em T, então incluir S e S' em uma
                  nova árvore em T
```

Algoritmo 1. Construção da floresta T

Finalmente, é construída a floresta **T** de árvores geradoras máximas (ver Algoritmo 1) a partir de subgrafos do grafo **G** construído anteriormente. Uma árvore geradora **A**, construída a partir de um grafo **G**, é aquela que contém todos os nós de **G** e apenas os arcos de **G** que não produzem ciclos em **A**. Uma árvore geradora é dita

máxima quando descarta os caminhos com pesos menores [Tenenbaum1995]. Um subgrafo **S** de um grafo **G** é um grafo cujos nós e arcos também pertencem a **G** [Goodrich2002]. A construção da floresta **T** de temas e subtemas utiliza o grafo **G** e a lista **5** com os pesos dos *stems* encontrados no texto. O Algoritmo 2 trata dos subtemas que são hierarquicamente inferiores a mais de um tema, completando a construção de **T**.

```
Para cada nó N de cada árvore de T faça
     Se N não é folha e ocorre mais de uma vez nesta árvore então
          N passa a ser folha nesta árvore e
          N e seus filhos são inseridos em nova árvore
```

Algoritmo 2. Tratamento de subtemas hierarquicamente inferiores a mais de um tema

3. Estudo de caso

A seguir, é apresentado um exemplo de parte de uma floresta de temas e subtemas obtida a partir de um texto. A Figura 2 apresenta o texto analisado (extraído do corpus Folhanot, gerado pelo Núcleo Interistitucional de Lingüística Computacional da USP – São Carlos, no âmbito do projeto AC/DC [Santos2000]). A Figura 3 apresenta as dez primeiras árvores obtidas a partir deste texto. Ao lado de cada nó das árvores aparece, entre parênteses, o peso do *stem*.

Enduro realiza duas provas em São Paulo. As quatro últimas vagas na equipe brasileira de enduro equestre para disputar o campeonato mundial na Holanda, em agosto, serão escolhidas dia 9 de abril. A seletiva, promovida pela Verdes Eventos, ocorrerá no hotel fazenda Dona Carolina, no km 39,5 da rodovia Bragança Paulista, em Itatiba, interior de São Paulo. Esta prova de velocidade livre terá 140 quilômetros, o mais longo percurso a ser disputado até hoje no Brasil. Outra prova do gênero é o "Grand Prix Frutilly de Enduro a Cavalo Infantil", promoção da Copercom que tem como objetivo incentivar nas crianças o gosto pelo esporte. A competição terá duas etapas: a primeira será realizada no dia 22 de maio no parque ecológico de Campinas e a segunda dia 26 de maio no terminal turístico de Águas de São Pedro (SP). Serão aceitas crianças com idade entre 5 e 12 anos, em categoria exclusiva para duplas, de velocidade controlada.

Figura 2. Texto analisado

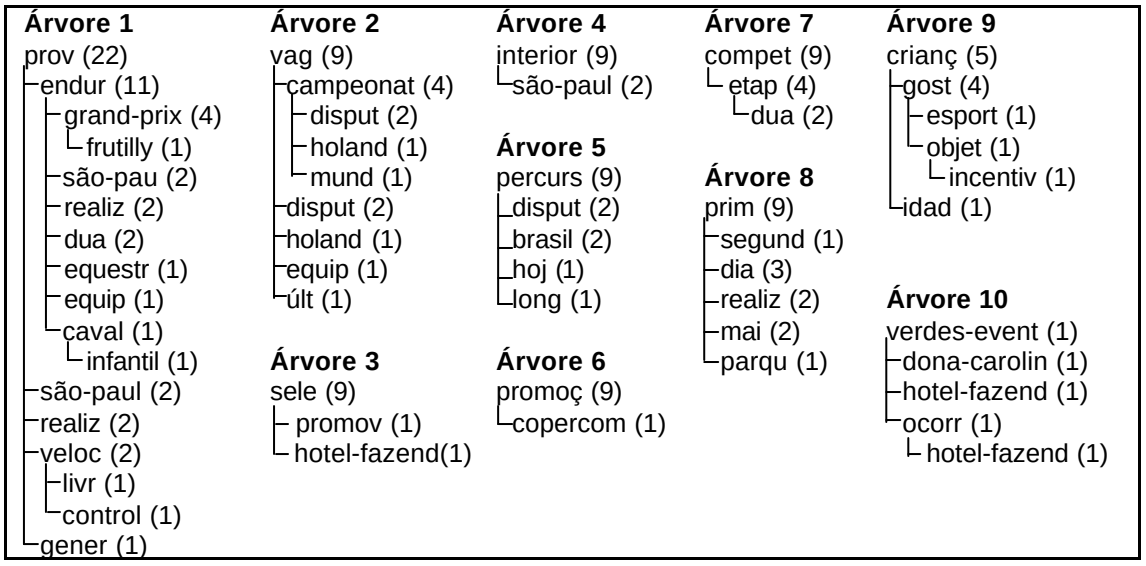

Figura 3. Parte da floresta de temas e subtemas

Na árvore 1 está representado o assunto “provas de enduro”, que seria considerado, nesta abordagem, o tema principal, dividido hierarquicamente em seus subtemas. Nas árvores restantes temos informações sobre estas provas, agrupando assuntos correlatos em cada estrutura. Nas árvores 2 e 9, temos as razões da realização das provas (respectivamente, as “vagas para o campeonato disputado na Holanda” e “crianças com gosto pelo esporte”). Nas árvores 3, 4, 5, 7 e 8, há informações sobre características e local das provas. E nas árvores 6 e 10, temos informações sobre os promotores dos eventos.

4. Considerações finais

A representação de textos através de estruturas hierárquicas tem aplicação na recuperação de informação, contribuindo para a indexação de documentos; na sumarização, dando bons indícios para a seleção das partes mais relevantes do texto original; e na categorização de textos, pelas semelhanças e diferenças que podem ser detectadas entre as florestas construídas.

Em http://www.inf.pucrs.br/~gonzalez/ri/tema pode ser encontrado um protótipo de uma ferramenta para a construção de hierarquia de temas e subtemas, que adota a abordagem descrita aqui. A pré-avaliação deste protótipo foi realizada através de análise visual, por observadores humanos, das estruturas construídas. Tal pré-avaliação, embora positiva, foi considerada subjetiva, não permitindo o cálculo de percentuais de precisão dos resultados obtidos. Como trabalho futuro, projetado também como forma de avaliar uma aplicação desta abordagem, pretendemos integrá-la a um sistema de recuperação de informação, mais especificamente, na etapa de indexação de documentos.

Referências bibliográficas

[Contreras2001] CONTRERAS, H.Y.; DÁVILA, J. Procesamiento del Lenguaje Natural basado en una “Gramática de Estilos” para el Idioma Español. CLEI’2001, Mérida, 2001. CD-ROM.

[Goodrich2002] GOODRICH, M.; TAMASSIA, R. *Algorithm Design – Foundation, Analysis, and Internet Examples*. New York: John Wiley & Sons, 2002. 708 p.

[Kim2000] KIM, M.; LU, F.; RAGHAVAN, V. V. Automatic Construction of Rule-based Trees for Conceptual Retrieval. 7th International Symposium on String Processing and Information Retrieval (SPIRE), 2000.

[Loukachevitch2000] LOUKACHEVITCH, N. V.; Dobrov, B. V. Thesaurus-Based Structural Thematic Summary in Multilingual Information SyXs. Machine Translation Review, N. 11, dezembro de 2000. p.10-20.

[Santos2000] SANTOS, D.; BICK, E. Providing Internet Access to Portuguese Corpora: The AC/DC Project. In: Gavriladou, M.; Carayannis, G.; Markantonatou, S.; Piperidis, S.; Stainhaouer, G. (editores). Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000, Atenas, 2000. p.205-210.

[Shatkay2000] SHATKAY, H.; WILBUR, J. Finding Themes in Medline Documents: Probabilistic Similarity Search, IEEE , Advances in Digital Libraries, 2000.

[Tenenbaum1995] TENENBAUM, A. M.; LANGSAM, Y.; AUGESTEIN, M. J. *Estruturas de dados usando C*. São Paulo: Editora McGraw-Hill, 1995. 826 p.