Identificação automática de segmentos discursivos por meio de conhecimento morfossintático

Erick Galani Maziero, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes

Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

erickgm@grad.icmc.usp.br, {taspardo, gracan}@icmc.usp.br

1. Introdução

Visando a um melhor desempenho dos métodos de Análise Discursiva Automática (ADA, uma tarefa da área de Processamento de Línguas Naturais, a qual serve de base para diversas outras aplicações), vários trabalhos abordam o problema de segmentação textual para obtenção de segmentos que encerrem em si idéias básicas do texto sobre o qual será feita a análise discursiva.

Em ADA, pressupõe-se que um texto contém uma estrutura bem elaborada, onde cada uma de suas partes não é desconexa ou desordenada, mas segue padrões, de acordo com os objetivos do compositor do texto. Assim, trabalha-se para produzir uma estrutura, muitas vezes chamada retórica, no formato de uma árvore, que reflita a organização textual. Isso é exatamente o que prescreve a RST (*Rhetorical Structure Theory*) [1], umas das teorias discursivas mais difundidas atualmente e foco deste trabalho.

Há vários métodos para se obter os segmentos de um texto: os baseados em expressões regulares, como apresentado em [2]; há os que utilizam análise numérica e estatística, como o sistema *TextTiling* [3]; e os métodos que se utilizam de Aprendizado de Máquina, uma subárea da Inteligência Artificial, como o sistema *Satz* [4].

2. Objetivos

O método aqui proposto caracteriza-se por propor e utilizar regras baseadas em informação morfossintática provenientes do parser PALAVRAS [5], um dos melhores analisadores sintáticos automáticos para o português do Brasil. Um exemplo de regra é: "segmenta-se um texto nos pontos em que há conjunções coordenativas".

Neste trabalho, visa-se à obtenção de segmentos para posterior incorporação ao analisador discursivo automático DiZer (*DIscourse analyZER for Brazilian Portuguese*) [6], o qual é baseado na RST. Sabe-se que a sintaxe tem um papel importante no processo de segmentação, como pode ser evidenciado em [7].

3. Resultados

O método de segmentar texto com vistas à análise retórica automática produziu os resultados esperados, ou seja, os segmentos de texto gerados contêm uma idéia básica do texto, diferentemente de segmentar o texto em sentenças, as quais podem conter em si mais de uma idéia básica.

Por exemplo, a Figura 1 abaixo exibe uma sentença com duas idéias básicas: "João foi à escola" e "ele não estava bem", evidenciando que uma sentença não contém apenas uma idéia.

João foi à escola embora não estivesse bem.

Figura 1: Sentença com duas idéias básicas.

Como esperado, o método apresentado segmenta o texto da figura em duas partes:

1: João foi à escola

2: embora não estivesse bem.

4. Conclusões

Este trabalho consiste em um pré-processamento do texto para a tarefa de análise retórica, e pode ser usado como base tanto para a tarefa manual quanto automática, o que produz um melhor acordo entre diversas análises de um mesmo texto.

Como trabalhos futuros, pretende-se incorporar o conjunto de regras desenvolvido ao sistema DiZer, já citado anteriormente.

5. Agradecimentos

Às agências de fomento à pesquisa FAPESP, CAPES e CNPq.

6. Referências Bibliográficas

[1] Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

[2] Walker, D.J.; Clements, D.E.; Darwin, M. and Amtrup, J.W. (2001). Sentence Boundary Detection: A Comparison of Paradigms for Improving MT Quality. In the *Proceedings of the 8th Machine Translation Summit*.

[3] Palmer, D. D. and Hearst, M.A. (1994). Adaptive Sentence Boundary Disambiguation. In the *Proceedings of the Conference on Applied Natural Language Processing*. Stuttgart, Germany.

[4] Palmer, D. D. and Hearst, M.A. (1997). Adaptive Multilingual Sentence Boundary Disambiguation. *Computational Linguistics*, Vol. 23, No. 2, pp. 241-267.

[5] Eckhard B. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

[6] Pardo, T.A.S. (2005). *Métodos para Análise Discursiva Automática*. Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Junho, 211p.

[7] Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual*. ISI Technical Report ISI-TR-545.