Aprimoramento e Avaliação do Analisador Discursivo Automático DiZer para o Português do Brasil

Erick Galani Maziero, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

erickgm@grad.icmc.usp.br, taspardo@icmc.usp.br

1. Introdução

A tarefa denominada de Análise Discursiva Automática consiste em determinar automaticamente a estrutura discursiva subjacente a textos. Em geral, tal estrutura, em forma de árvore (em geral binária), relaciona as partes do texto (orações, sentenças ou parágrafos, por exemplo) por meio de relações discursivas, como causa-efeito e contraste, dentre várias outras. Assim, as folhas da árvore são os segmentos textuais e os nós internos são as relações que se estabelecem entre os segmentos textuais. A teoria discursiva mais difundida que dita como estruturar um texto é a RST [1].

O conhecimento discursivo é de grande valia para diversas aplicações de processamento da língua, como sumarização automática e perguntas e respostas. Neste cenário, diversos sistemas de análise automática segundo a RST têm surgido. Para a língua portuguesa, há somente um: o sistema DiZer [2]. Por ser uma primeira abordagem para o problema, o sistema apresenta várias limitações, como os métodos de segmentação textual empregados.

2. Objetivos

O objetivo deste trabalho é aprimorar o DiZer, visto que este detecta apenas orações ou sentenças do texto a ser analisado, de forma que alguns segmentos discursivos não são corretamente identificados. Assim, em uma primeira etapa, foi desenvolvido um novo módulo de detecção de segmentos discursivos, isto é, as partes do texto que se relacionam via relações discursivas.

Dado que a tarefa de avaliar estruturas retóricas é sujeita a subjetividades como delimitação dos segmentos discursivos e similaridade entre relações retóricas atribuídas, uma ferramenta de avaliação automática de sistemas de análise discursiva foi proposta, visto que, até então, a avaliação de sistemas era feita manualmente, não fornecendo um valor muito confiável, pois pode variar de pessoa para pessoa.

3. Material e Métodos

O método de segmentação incorporado ao DiZer caracteriza-se por usar informações morfossintáticas e sintáticas proveniente do PALAVRAS [3], um dos melhores analisadores sintáticos automáticos para o português. Foram definidas diversas regras de segmentação que ditam os limites dos segmentos discursivos. Para a definição das regras seguiu-se o manual de Carlson e Marcu [4].

Utilizando como base a metodologia de avaliação proposta por Marcu [5], foi implementada uma ferramenta web para avaliar as estruturas discursivas produzidas pelo DiZer, tomando-se como referência estruturas criadas manualmente. Desta forma as subjetividades já tratadas neste

documento foram eliminadas, pois todo o processo é feito de forma automática. No entanto alguns relaxamentos, que não prejudicam a avaliação, foram empregados: eliminação de stopwords, por exemplo. A avaliação tem como resultado medidas chamadas *recall* e *precision*, amplamente empregadas na avaliação de tarefas de Processamento de Línguas Naturais.

4. Resultados e Discussão

Com a ferramenta de avaliação automática foram avaliados os dados gerados pelo DiZer em sua primeira versão. Como próximo passo, será avaliado o sistema com seu novo módulo de segmentação. Espera-se uma melhor precisão das estruturas geradas com a nova segmentação por esta se aproximar mais da segmentação humana e permitir que o DiZer empregue melhor as relações retóricas.

5. Conclusões

Este trabalho visa a contribuir com tarefas de processamento da língua portuguesa que envolvam algum nível de conhecimento discursivo, por contar com a geração automática de estruturas retóricas do texto em tratamento. A ferramenta de avaliação automática permite uma avaliação uniforme e rápida de sistemas de análise retórica automática. Espera-se que, com isso, a área de pesquisa tratada neste artigo evolua mais rapidamente.

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] Pardo, T.A.S. and Nunes, M.G.V. (2006). DiZer – an Automatic Discourse Analyzer for Brazilian Portuguese. In the Proceedings of the V Best MSc Dissertation/PhD Thesis Contest – CTDIA. Ribeirão Preto-SP, Brazil. October 23.

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

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

[5] Marcu, Daniel (2000). The Rhetorical Parsing of Unrestricted Texts: A SurfaceBased Approach. Computational Linguistics, 26(3):395--448.