

# **Aquisição Automática de Conhecimento Discursivo para o Português do Brasil**

**Pedro Paulo Balage Filho, Thiago Alexandre Salgueiro Pardo**  
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos

## **1. Objetivos**

O uso de conhecimento discursivo, em seus vários níveis, é de grande valia para sistemas de Processamento de Língua Natural (PLN). Suas aplicações estão em diversas áreas, como, por exemplo, sistemas de diálogo; sumarizadores de texto; sistemas de auxílio à escrita; tradutores automáticos; etc. O sistema capaz de extrair o conhecimento discursivo subjacente a um texto é chamado analisador discursivo. Dentro deste contexto, objetiva-se neste trabalho investigar e desenvolver uma ferramenta para aquisição automática de conhecimento discursivo a partir de corpus. Tal conhecimento visa alimentar analisadores discursivos, possibilitando assim, a automatização do desenvolvimento/adaptação desses sistemas para textos de quaisquer gêneros e domínios.

## **2. Material e Métodos**

Com base em um corpus anotado retoricamente, conhecimentos de diversas naturezas serão pesquisados pela ferramenta: palavras e frases indicativas, marcadores discursivos, relações e marcas sintáticas, a semântica das palavras, estatísticas, etc. Para isso, diversas ferramentas de PLN subsidiarão o trabalho, como analisadores morfossintáticos, analisadores sintáticos e wordnets, recursos estes disponíveis para o português do Brasil. O conhecimento discursivo produzido será avaliado no contexto do analisador discursivo DiZer [1] que deverá demonstrar a utilidade da ferramenta produzida.

## **3. Resultados e Discussão**

Inicialmente foram estudados trabalhos semelhantes com o tema proposto. Os trabalhos mais relevantes se mostram em Anthony e Lashkia [2] e Pendar e Cotos [3], no qual são apresentadas ferramentas para o auxílio da escrita de artigos científicos através da identificação automática de sua estrutura

retórica. Estas ferramentas são capazes de aprender as estruturas retóricas de um texto previamente anotado de forma automática. De forma semelhante às ferramentas desenvolvidas nestes trabalhos, a ferramenta que está sendo desenvolvida utilizará uma base de aprendizado obtida de corpus e através dela induzirá informações que identifiquem as relações retóricas presentes. Na fase atual do projeto, o sistema está parcialmente funcional, sendo possível a análise das informações mais frequentes dentro de uma determinada relação retórica. O próximo passo será a adição de ferramentas de PLN para o português do Brasil, diferencial deste trabalho, e a avaliação com o analisador discursivo DiZer.

## **4. Conclusões**

O trabalho, ainda em fase de implementação, se mostra de grande importância. O estudo dos métodos de aquisição automática de conhecimento discursivo ainda é novo para a língua portuguesa, e dessa forma, uma ferramenta com esta capacidade representará um grande avanço na área.

## **5. Referências Bibliográficas**

- [1] Pardo, T.A.S. e Nunes, M.G.V. (2005). *Investigação e Desenvolvimento de Modelos Estatísticos para Análise Discursiva Automática*. Série de Relatórios Técnicos do ICMC-USP, no. 251. São Carlos-SP, Janeiro, 31p
- [2] Anthony, L.; Lashkia, G. V. (2003). *Mover: a machine learning tool to assist in the reading and writing of technical papers*. IEEE Transactions on Professional Communication 45 pp.185-193.
- [3] Pendar, N. And Cotos, E. (2008). In the *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, USA, pp 62–70.