Pré-processamento de Textos para a Extração de Relações do Domínio de Organizações para o Português

Tiago Comassetto Froes, Renata Vieira[2] (orientador)

[1]*Faculdade de Engenharia, PUCRS,* [2]*Faculdade de Informática, PUCRS*

Resumo

A Extração de Relações (ER) a partir de textos é um dos principais desafios da área de Extração de Informação (EI), a qual busca identificar relações entre determinadas entidades. Nesse contexto, as tarefas de EI envolvem a identificação de Entidades Nomeadas (ENs), como nome de pessoas e de organizações, e a extração de relações entre estas entidades em textos em linguagem natural. Várias abordagens têm sido propostas para ER a partir de dados não estruturados, dentre elas destaca-se no aprendizado estatístico os modelos condicionais globais, chamados de *Conditional Random Fields* (CRF). A literatura tem apresentado o modelo CRF como uma boa alternativa, uma vez que têm sido aplicado eficientemente em diversas aplicações de Processamento de Linguagem Natural (PLN), incluindo recentemente a tarefa de ER. O trabalho proposto aqui relaciona-se ao tema da tese de Doutoramento de Sandra Collovini que trata a extração automática de relações entre entidades do domínio de Organizações para o Português baseado em CRF. Este trabalho apresenta as etapas de pré-processamento dos textos de entrada necessárias para a ER proposta. De forma resumida, a etapa de pré-processamento é constituída pelos seguintes passos: (i) seleção dos textos do domínio de Organizações, como artigos de notícias da atualidade que tratem assuntos relacionados ao tema de pesquisa, como negócios, incorporações, entre outros; (ii) preparação dos textos, como por exemplo, limpeza e conversão para diferentes codificações dos textos, anotação dos textos por um analisador sintático, aplicação de um sistema de Reconhecimento de Entidades Nomeadas (REN) para identificação e categorização das ENs contidas nos textos; (iii) construção das instâncias de relações, envolvendo a identificação do segmento do texto que descreve uma relação entre pares de ENs identificadas em (ii). Os textos com as anotações necessárias para a aplicação e avaliação do modelo CRF são resultado da etapa de pré-processamento.