VALIDAÇÃO DE EXTRAÇÃO SEMI-AUTOMÁTICA DE RELAÇÕES SEMÂNTICAS

Aluno: Andrea da Fonseca Barreto
Orientador: Violeta Quental

Introdução

Esse projeto teve por proposta a Validação, Avaliação e Revisão de Relações semânticas no corpus AC/DC (VARRA), de termos extraídos e classificados de forma automática pelo projeto PAPEL (www.linguateca.pt/PAPEL).

O VARRA é um sistema que foi desenvolvido com o objetivo principal de contribuir na avaliação (ou validação) manual detalhada de relações semânticas entre pares de palavras, tendo por fundamento as ocorrências de tais pares em contextos autênticos expressos por frases do corpus do projeto AC/DC (http://www.linguateca.pt/acdc). Procurou-se com isso construir um suporte confiável de considerações sobre determinadas relações semânticas, mais semelhantes com a valoração humana de falantes nativos do português, por se acreditar que mesmo apresentando um alto custo de elaboração (quanto ao tempo e à mão-de-obra) e da possibilidade de variação valorativa entre os sujeitos avaliadores, o julgamento humano é o meio mais confiável de se "avaliar a qualidade de um recurso construído de forma automática" (Freitas, Santos, Oliveira & Quental, 2010).

"A análise e etiquetagem semântica entre termos de textos revela-se especialmente importante nas tarefas de processamento de linguagem natural que envolvem o uso de ontologias e taxonomias e a resolução de conferência, como a sumarização automática, a recuperação e mineração de informação textual, a tradução automática." (Quental, 2010).

Objetivos

Esse projeto teve por objetivo avaliar as relações semânticas obtidas na ontologia do PAPEL, extraída semi-automaticamente de dicionário. Neste sentido, a bolsista foi estimulada a: i) apresentar sugestões de correção e adequação com relação à terminologia utilizada para caracterizar as relações em estudo, principalmente no tocante às relações que se apresentavam demasiadamente longas (*causador_de_algo_com_propriedade* e *propriedade_de_algo_que_causa*) e, por isso, poderiam ser alvo de má interpretação por parte dos varredores (avaliadores); ii) analisar, em um teste piloto, a pertinência da interface que seria aplicada aos dossiês de avaliação, inclusive com relação à seleção vocabular das alternativas que deveriam ser escolhidas pelos varredores para validar ou não as relações; iii) funcionar como um dos varredores, analisando as relações nos dossiês entregues pela equipe do VARRA. Pretendeu-se, também, "obter julgamentos de falantes nativos mais precisos quanto às relações semânticas em questão, buscando validá-las a partir do uso das palavras em contextos autênticos, representados por frases dos corpora do projeto AC/DC." (Quental, 2010).

Metodologia

A metodologia de trabalho utilizou as relações entre palavras apresentadas pelo PAPEL (Palavras Associadas Porto Editora – Linguateca – uma rede lexical pública para o português e acessível eletronicamente através do link já acima mencionado), através de interface pronta para testagem contendo as relações a serem validadas, as frases – exemplos, coluna para julgamento e coluna para comentários. Desta forma foram criados o que a equipe do VARRA

convencionou denominar de "dossiês", que se caracteriza por ser o documento eletrônico onde o varredor é convidado a fazer suas observações e análises da ocorrência das relações entre pares de palavras previamente selecionadas em contextos autênticos, baseando-se nas alternativas de validação previstas e explicitadas nas "Instruções para validação de relações semânticas entre palavras usando o VARRA", entregue pela equipe do VARRA aos varredores.Vale ressaltar que, como etapa inicial de todo o processo, as relações a serem apreciadas nos dossiês são, primeiramente, entregues aos varredores em forma de simples lista para uma primeira avaliação sem contexto, baseada apenas na intuição do avaliador. O objetivo desta avaliação prévia é de se verificar possíveis erros de relações (possibilidade existente, devido ao fato de terem sido extraídas de forma automática), e de meio de comparação entre as intuições significativas ocorridas com as relações dentro e fora de contexto.

Conclusão

Em um primeiro teste com o VARRA, dez alunos de graduação do curso de Letras da PUC-Rio responderam os dossiês distribuídos pela equipe de Lingüística Computacional, contendo, cada um 200 instâncias de relações, perfazendo um total de 5243 julgamentos.

Uma análise preliminar mostrou a necessidade de se levar em conta o fato de haver possíveis desajustes entre os corpora e os varredores, com relação a determinados julgamentos, devido ao fato de serem falantes nativos de português brasileiro e/ou português europeu.

Outra conclusão preliminar que pôde ser obtida foi que grande parte das relações que figuravam como relações de hiperonímia, estabeleciam, na realidade, relações de sinonímia. Observou-se que isto ocorre devido ao fato dos termos hiperônimos serem utilizados para conferir coesão textual em uma relação anafórica e, assim, hiperônimos funcionam como sinônimos (Freitas, Santos, Oliveira & Quental, 2010).

Referências:

1.FREITAS, Cláudia, Diana Santos, Hugo Gonçalo Oliveira & Violeta Quental. *VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC*. ELC2010 (2010). Disponível em www.linguateca.pt/Diana/download/resFreitasetalELC2010.pdf

2.FREITAS, Cláudia. Instruções para a validação de relações semânticas entre pares de palavras usando o VARRA – **V**alidação, **A**valiação e **R**evisão de **R**elações semânticas no AC/DC – versão 1.1, 18 de Dezembro de 2009.
Disponível em www.linguateca.pt/acesso/InstrucoesVARRA.pdf.

3.QUENTAL, Violeta. *Projeto Validação de extração semi-automática de relações semânticas* (PIBIC-CNPq/PUC-Rio), 2010.