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).

Relações semânticas entre pares de palavras extraídas automaticamente de corpus, geralmente apresentam resultados lacunosos, porém, com a crescente importância da construção e manutenção de léxicos computacionais capazes de utilizar informação semântica de forma adequada, é que surge o VARRA: um sistema 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). Assim, procurou-se 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

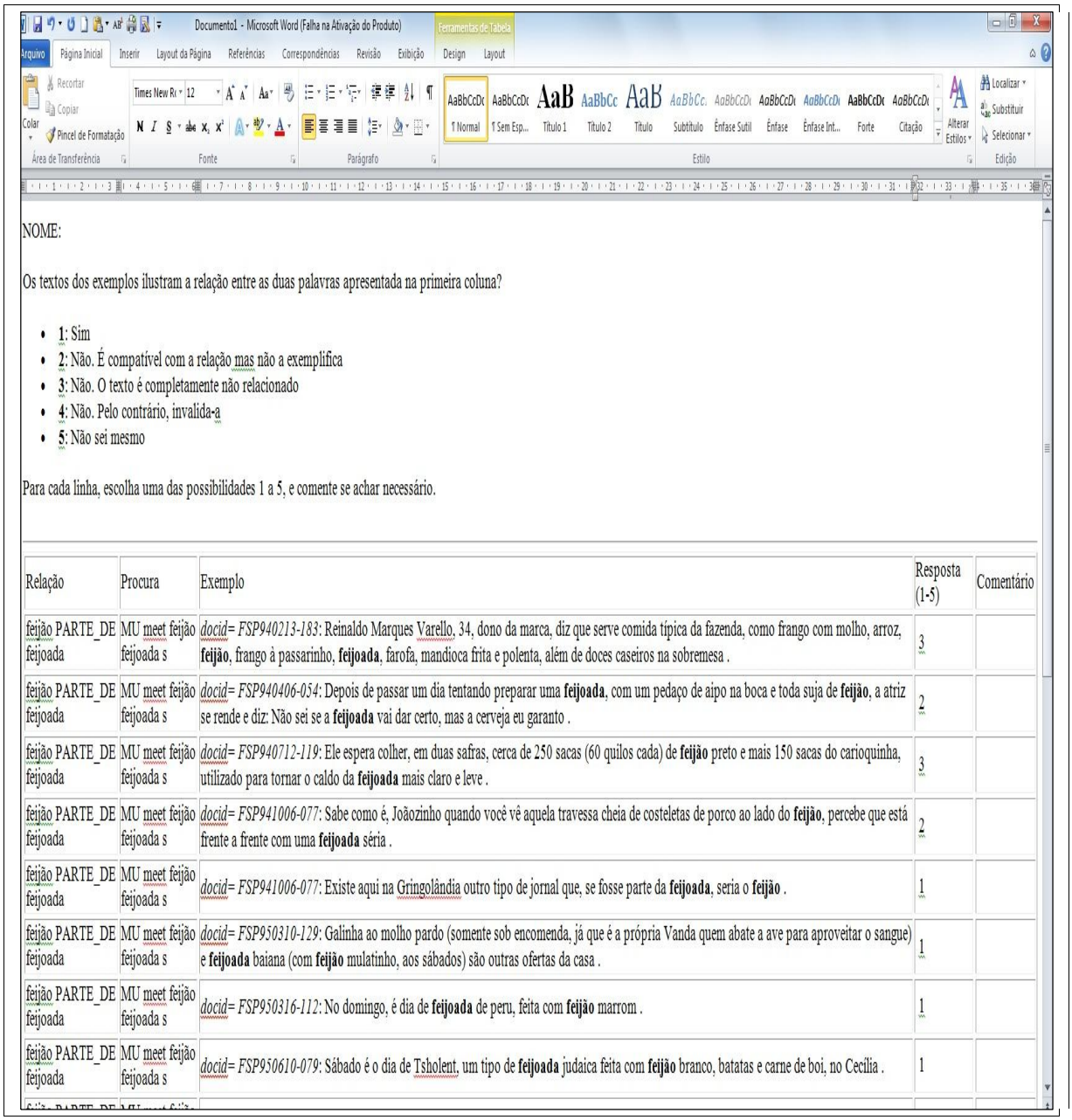
Esse projeto teve por objetivo avaliar as relações semânticas obtidas na ontologia do PAPEL, extraída semi-automaticamente de dicionário. Neste sentido, a bolsista foi estimulada a: i) apresentar sugestões de correção e adequação com relação à terminologia utilizada para caracterizar as relações em estudo, 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, conforme figura abaixo



Documento1 - Microsoft Word (Falha na Ativação do Produto)

NOME:

Os textos dos exemplos ilustram a relação entre as duas palavras apresentada na primeira coluna?

- 1: Sim
- 2: Não. É compatível com a relação mas não a exemplifica
- 3: Não. O texto é completamente não relacionado
- 4: Não. Pelo contrário, invalida-a
- 5: Não sei mesmo

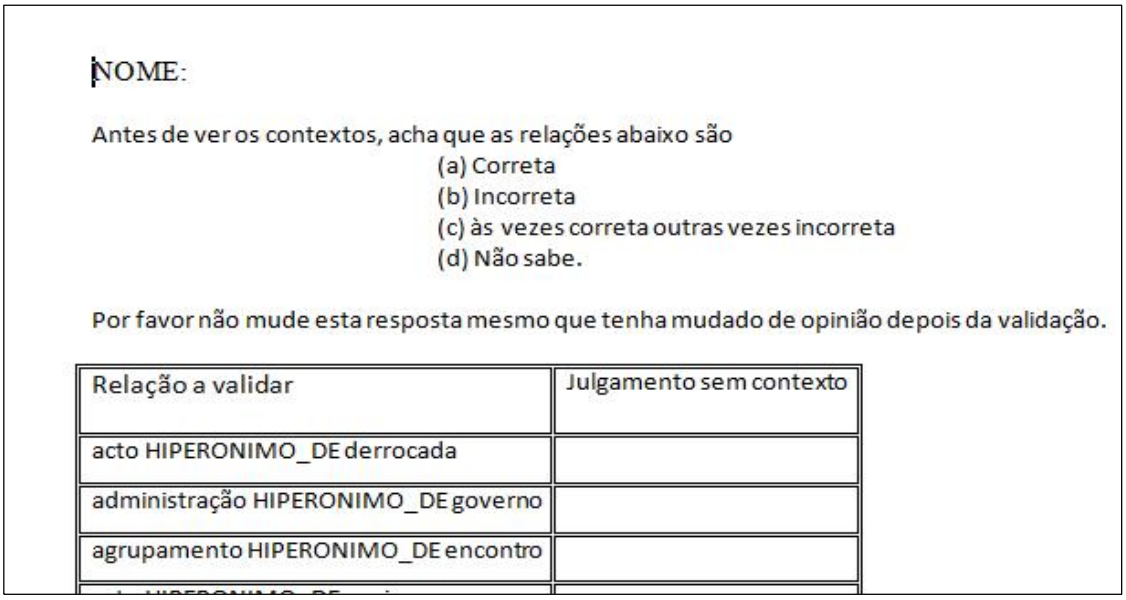
Para cada linha, escolha uma das possibilidades 1 a 5, e comente se achar necessário.

Relação	Procura	Exemplo	Resposta (1-5)	Comentário
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP940213-183*: Reinaldo Marques Varello, 34, dono da marca, diz que serve comida típica da fazenda, como frango com molho, arroz, **feijão**, frango à passarinho, **feijoada**, farofa, mandioca frita e polenta, além de doces caseiros na sobremesa .	3	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP940406-054*: Depois de passar um dia tentando preparar uma **feijoada**, com um pedaço de aipo na boca e toda suja de **feijão**, a atriz se rende e diz: Não sei se a **feijoada** vai dar certo, mas a cerveja eu garanto .	2	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP940712-119*: Ele espera colher, em duas safras, cerca de 250 sacas (60 quilos cada) de **feijão** preto e mais 150 sacas do carioquinha, utilizado para tornar o caldo da **feijoada** mais claro e leve .	3	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP941006-077*: Sabe como é, Joãozinho quando você vê aquela travessa cheia de costeletas de porco ao lado do **feijão**, percebe que está frente a frente com uma **feijoada** séria .	2	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP941006-077*: Existe aqui na Gringolândia outro tipo de jornal que, se fosse parte da **feijoada**, seria o **feijão** .	1	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP950310-129*: Galinha ao molho pardo (somente sob encomenda, já que é a própria Vanda quem abate a ave para aproveitar o sangue) e **feijoada** baiana (com **feijão** mulatinho, aos sábados) são outras ofertas da casa .	1	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP950316-112*: No domingo, é dia de **feijoada** de peru, feita com **feijão** marrom .	1	
feijão PARTE_DE feijoada	MU meet feijão feijoada s	*docid= FSP950610-079*: Sábado é o dia de Tsholent, um tipo de **feijoada** judaica feita com **feijão** branco, batatas e carne de boi, no Cecília .	1	

Figura 1: Dossiê criado pela equipe VARRA

Vale ressaltar que antes de proceder com a validação das relações, os varredores foram orientados a seguir as 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. A opção pela utilização das alternativas foi motivada levando-se em conta as informações pertinentes para a avaliação e melhoria do PAPEL, e por apresentarem a característica de uma objetividade capaz de preservar as nuances que o julgamento humano é

capaz de elaborar. Porém, 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, conforme figura abaixo:



NOME:

Antes de ver os contextos, acha que as relações abaixo são
(a) Correta
(b) Incorreta
(c) às vezes correta outras vezes incorreta
(d) Não sabe.

Por favor não mude esta resposta mesmo que tenha mudado de opinião depois da validação.

Relação a validar	Julgamento sem contexto
acto HIPERONIMO_DE derrocada	
administração HIPERONIMO_DE governo	
agrupamento HIPERONIMO_DE encontro	

Figura 2: Formulário com as relações sem contexto

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).

Acreditamos que o VARRA possa servir, também, como repositório que contenha não só as relações validadas como também as diferenças de opiniões entre determinadas relações e que podem, futuramente, após seus estudos, servir como instrumento para uma melhor compreensão da semântica da língua portuguesa.

Outra contribuição que o sistema pode conferir à Lingüística é no aprofundamento do estudo das contradições de julgamento que aparecem nas relações dentro e fora de contexto.

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 – Validação, Avaliação e Revisão de Relações 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.