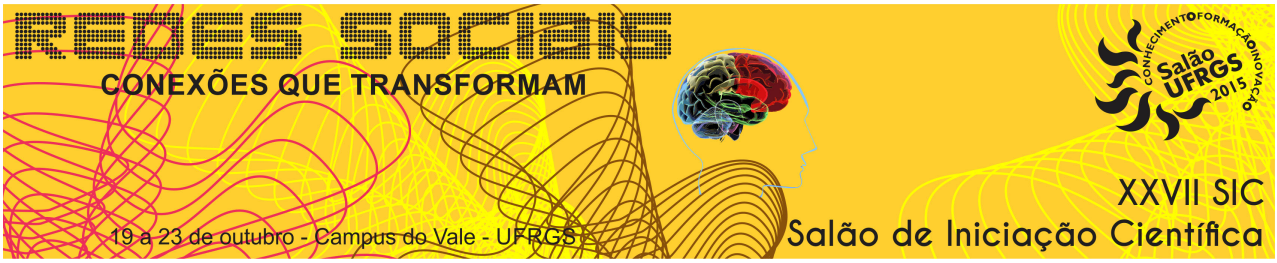

Evento	Salão UFRGS 2015: SIC - XXVII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2015
Local	Porto Alegre - RS
Título	Avaliação de corpus para geração de tesauros distribucionais
Autor	EDUARDO DELAZERI FERREIRA
Orientador	ALINE VILLAVICENCIO

Título: Avaliação de corpus para geração de tesauros distribucionais

Autor: Eduardo Delazeri Ferreira
Orientadora: Aline Villavicencio
Instituição de Origem: Universidade Federal do Rio Grande do Sul (UFRGS)

Um tesauro é uma lista de palavras agrupadas pela sua proximidade semantica, tais como sinonimos (por exemplo: comprar, adquirir, etc.). Através de tesauros é possível analisar significados de palavras através da similaridade entre elas. A criaçao destes tesauros manualmente é uma tarefa muito demorada, pois exige conhecimento especializado. Uma alternativa é automatizar a sua construçao, na forma de tesauros chamados distribucionais, que utilizam técnicas de distribuição de ocorrência de palavras em grandes quantidades de textos para identificar suas relações semânticas. Estes tesauros são importantes para uma série de aplicações, tais como simplificação de textos e tradução automática. Como estas técnicas são baseadas em co-ocorrências, a qualidade deles pode ser influenciada por fatores como a qualidade e a quantidade dos textos utilizados e também as características gerais da língua.

Este trabalho objetiva avaliar o impacto de alguns fatores como o tamanho dos textos de entrada na geração de tesauros distribucionais. Isso tendo em vista que o texto representa a única fonte de informação linguistica disponível para a criação do tesauro e se assume que quanto maior o texto maior a qualidade do recurso resultante.

Para geraçao do tesauro utilizamos uma ferramenta (word2vec) que cria representações vetoriais de palavras em um espaço n-dimensional. Uma das propriedades deste espaço é que algumas operações sobre os vetores correspondem a propriedades semânticas e morfossintáticas das palavras. Para avaliar a qualidade do tesauro resultante sera usado um conjunto de testes similar ao TOEFL, que é um teste de múltipla escolha para identificaçao de palavras semanticamente relacionadas: dada uma palavra alvo e um conjunto de 4 palavras candidatas como alternativas, o tesauro deve identificar qual é a mais relacionada (pergunta: abacaxi - alternativas: maça, régua, cadeira, montanha). O resultado é reportado em termo de acurácia (a porcentagem de vezes que a resposta correta foi atingida).

Para avaliar o impacto do tamanho do texto na geração desses tesauros, foram utilizados dois textos como base. Um deles em inglês, limpado anteriormente, e outro em português, apenas contendo os lemmas. Ambos os textos foram particionados (variando de 1% até 100%). Cada um desses pedaços foi avaliado utilizando o mesmo conjunto de parâmetros, resultando em tesauros que podem indicar a convergência do tamanho necessário do corpus de treino.

Ao avaliarmos os tesauros distribucionais gerados em inglês, identificamos que com 1% obtemos uma cobertura de 96.12% e uma precisão de 48.62%. Com 10% a cobertura fica em 98.54% e com precisão de 65.37%. Com 100% a cobertura é de 99.47% e a precisão de 69.12%. Para o português identificamos que com 1% obtemos uma cobertura de 52.59% e um acerto de 56.77%. Com 10% a cobertura é de 60.39% e a taxa de acertos de 59.09%. Com 100% a cobertura é de 62.27% e a taxa de acertos de 59.65%.

O importante de se analisar nesses testes é a acurácia em relação a cobertura, visto que a cobertura em si tem uma forte dependência da facilidade do teste aplicado. Analisando a

acurácia dos testes aplicados é perceptivel que os resultados começam a estabilizar, o ganho de 1% para 10% é muito maior que o ganho de 10% para 100%. Portanto a variação no tamanho do corpus é importante, mas o maior limitador é a qualidade dos textos utilizados. Finalmente um ultimo teste foi realizado com o objetivo de aumentar a qualidade do tesauros para o português. Uma parte do texto de 10% foi limpa de forma semelhante ao do inglês, mantendo apenas palavras com valor semântico, os resultados obtidos foram: cobertura 56.25% e a acuracia em 62.36%. Portanto e possível perceber o aumento de qualidade desses 10% foi suficiente para o resultado superar inclusive os 100% originais.