Avaliação de corpus para geração de tesauros distribucionais

Eduardo Ferreira
edferreira@inf.ufrgs.br

Drª Aline Villavicencio
avillavicencio@inf.ufrgs.br

Instituto de Informática - Universidade Federal do Rio Grande do Sul

Objetivo

Avaliar o impacto do tamanho dos corpus utilizados de entrada na geração de tesauros distribucionais, tendo em vista que o texto representa a única fonte de informação linguistica disponível para a criação do tesauro.

Motivação

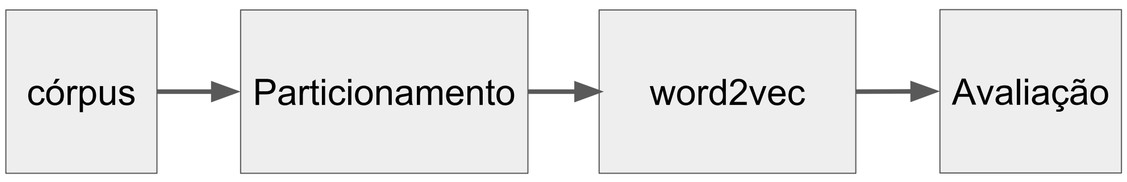
Tesauros distribucionais são importantes para aplicações em NLP, como simplificação e tradução automática de textos.

Conceitos

Córpus: conjuntos de textos escritos em uma determinada.
Tesauros Distribucionais: lista de relações de palavras organizada por relação de coocorrência. Como visto na tabela abaixo.

palavra 1	palavra 2	relação
narrar	poema	0,617
narrar	choramingar	0,096
narrar	importalizar	0,479
narrar	calar	0,067

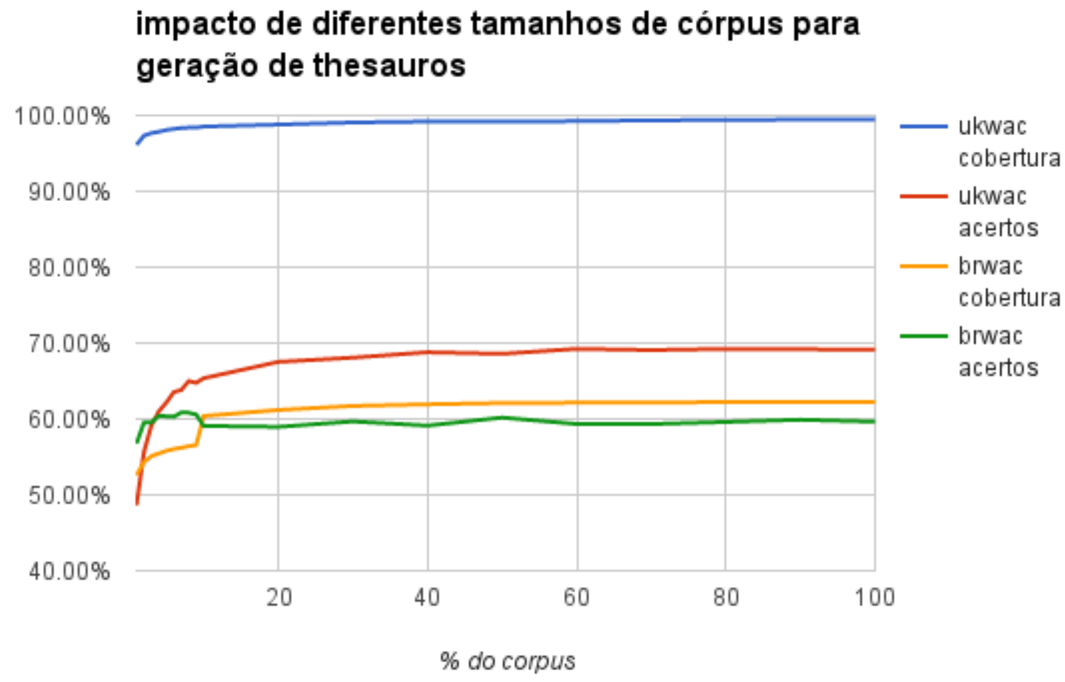
Metodologia



Córpus: ukWaC, $brwac_{original}$ e $brwac_{filtrado}$
ukWaC: rempção stop-words e noramlização de palavras
$brwac_{original}$: forma canonica das palavras
$brwac_{filtrad}$: brWaCoriginal com remoção de stop-words
Particionamento: dividido os córpus de 1% a 100%.
Word2vec: utilizado para criar os tesauros distribucionais.
Avaliação: gold standard de perguntas contendo 4 alternativas semânticamente relacionadas.

amora
(a) amoreira (c) mirante
(b) frontal (d) prognóstico

Resultados



Ambos os córpus mostram uma estabilização dos acertos e da cobertura.

Criado um novo thesauro utilizando 10% do brWaC, mas filtrando stop-words. O resultado segue na tabela abaixo.

	cobertura	acertos
$brwac_{original}$	56.41%	60.82%
$brwac_{filtrado}$	56.25%	62.36%

Conclusões

Estabilização devida a redundância nos córpora.
Redundância não acrescenta qualidade.
Remoção de stop-words melhora a qualidade do modelo.

Agradecimentos

Agradecemos ao Instituto de Informática da UFRGS pelo apoio à pesquisa. Parte dos resultados apresentados neste trabalho foram obtidos no projeto *Simplificação Textual de Expressões Complexas*, patrocinado pela Samsung Eletrônica da Amazônia Ltda. através da lei 8.248/91, e também contou com apoio do CNPq (113700/2015-6).