Evento	Salão UFRGS 2014: SIC - XXVI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2014
Local	Porto Alegre
Título	Simplificação lexical de substantivos e multiword expressions
Autor	MARCELY ZANON BOITO
Orientador	ALINE VILLAVICENCIO

Este trabalho consiste na construção de um pipeline independente de linguagem para simplificação textual lexical com foco em expressões multipalavras e substantivos.

Simplificação lexical é uma forma de simplificação que funciona através da identificação e substituição de palavras consideradas complexas – segundo algum determinado parâmetro – por equivalentes (hipônimos, hiperônimos e sinônimos) mais simples. Assim, essa modalidade de simplificação auxilia pessoas com baixo letramento, perca cognitiva e falantes não-nativos da língua, facilitando a compreensão de textos.

O foco em expressões multipalavras (abreviadamente MWE, de *multiword expressions*) é porque tais expressões são constituídas por vários "pedaços" separados que representam uma única ideia, muitas vezes completamente independente do significado individual dos termos que as compõem. Dessa forma, na simplificação de texto, elas são um desafio, pois o tratamento separado de seus termos leva ao erro e à perda de significado. Por exemplo, a expressão *run out*, que significa 'acabar', poderia erroneamente ser simplificada para *walk out*, que significa 'ir embora', em vez de *to end* ou *to finish*.

Para a classificação das palavras entre complexas e simples, numa primeira abordagem, usamos listas de frequências de palavras, que são geradas utilizando grandes conjuntos de textos. Para essa medida, admitimos que palavras que ultrapassam um determinado valor de frequência, são simples.

A justificação da frequência como indicador de complexidade é baseada na comparação de vários parâmetros (como tamanho da palavra e polissemia) em que fica evidente que, tanto para inglês quanto para o português, a medida mais relevante para medir a complexidade de uma palavra é a sua frequência em um determinado corpus. Esse trabalho ("*Size does not matter. Frequency does. A study of features for measuring lexical complexity.*") foi submetido para a conferência Iberamia 2014 e aguarda avaliação.

A informação de frequência utilizada nessa ferramenta é obtida de corpora externa, o ukWac (corpus de 2 bilhões de palavras retirado do domínio .uk) e WFWSE (*Word Frequencies in Written and Spoken English: Based on the British National Corpus*, Livro de listas de frequências para inglês), e através dela as palavras do texto alvo são separadas entre complexas e simples.

Para as palavras consideradas complexas, a ferramenta gera sugestões de possíveis substituições utilizando um recurso lexical, o WordNet, que contem informações de relações entre as palavras tais como sinonímia, hiponimia e hiperonímia, e para simplificação lexical, escolheu-se usar a sinonimia, pois é a que representa menor mudança de significado. As possíveis alternativas de simplificação são classificadas verificando que combinação de palavras ocorre com maior frequência, e a escolhemos como melhor simplificação. Assim, uma frase como '*The acquiescence of the crew'* pode encontrar como correspondente '*The acceptance of the crew'*, que, embora perca um pouco o significado original, fornece uma forma mais fácil de entendimento.

Para a avaliação do sistema de simplificação é utilizada como *gold standard* a *simple wikipedia*, uma versão manualmente simplificada da wikipedia em inglês. No momento, este pipeline está sendo testado em inglês, para a futura exportação para o português.