Evento	Salão UFRGS 2014: SIC - XXVI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2014
Local	Porto Alegre
Título	Modelo de Referência para a Simplificação Lexical de Termos Compostos do Inglês
Autor	LUIZA DE AZAMBUJA HAGEMANN
Orientador	ALINE VILLAVICENCIO

O desenvolvimento de sistemas de simplificação de textos visa auxiliar o entendimento de textos para falantes de uma língua, assim possibilitando sua inclusão no ambiente literário. Por exemplo, na sociedade brasileira, baixas habilidades linguísticas e mesmo analfabetismo são uma realidade enfrentada por muitas pessoas: em 2012, estimava-se que 25% da população brasileira com até 15 anos era composta de analfabetos funcionais. Quando consideramos os falantes de inglês como segunda língua, estima-se que estes números sejam ainda maiores.

Para que os sistemas de simplificação tenham um bom desempenho, é preciso que exista um modelo de referência (um *gold standard*) para avaliar os diferentes modelos. Neste trabalho, apresentamos os passos desenvolvidos até o momento para o desenvolvimento de um *gold standard* de expressões compostas para simplificação lexical.

Baseamo-nos no trabalho desenvolvido por Belder e Moens, que criaram um *gold standard* com 201 termos utilizando dados da campanha de avaliação SemEval 2007 e as frequencias do *Internet Corpus of English*. Após processos de filtragem para exclusão dos termos julgados de fácil entendimento (por estarem em uma lista de termos simples), foram obtidas 43 palavras. Para cada palavra, foram selecionadas 10 sentenças que as contivessem.

Desta forma, obteve-se 430 sentenças. Estas foram analisadas por 5 linguistas, que escolheram uma simplificação para cada termo. Todos os termos originais e suas simplificações foram, então, passados para dois grupos de anotadores, que os ordenaram de acordo com sua complexidade. Expressões compostas – chamadas *multiword expressions* (MWE), que englobam os *phrasal verbs* do inglês (como "*walk away*" e "*wake up*") – foram, no entanto, descartadas devido à metodologia de escolha e o tamanho reduzido do grupo de palavras original.

Em complemento ao trabalho de Belder e Moens, este trabalho objetiva criar um *gold standard* para a simplificação lexical de palavras e expressões compostas do inglês. Como exemplo, temos casos como: "*simple-minded*", que seria simplificado para "*stupid*"; "*cinnamon-colored*", que seria simplifiado para "*brown*". Adicionalmente, são utilizados recursos lexicais, como *BabelNet* e *WordNet*, para garantir que a simplificação se aplique apenas ao contexto implicado (realizando a desambiguação dos termos).

A primeira etapa do processo consistiu em gerar uma lista com mais de 70 mil termos compostos extraídos da *WordNet* (um banco de dados lexical da língua inglesa), separados por classe gramatical (substantivos, verbos, adjetivos e advérbios). Em uma segunda etapa, a fim de evitar qualquer tipo de ambiguidade, cada termo foi filtrado pelo *BabelNet*, tornando possível a discriminação dos sentidos de cada palavra. Foram descartadas todas que, após a filtragem, mantiveram algum tipo de ambiguidade, e como resultado, foram obtidas 67,9 mil MWEs.

No estágio atual, as palavras estão sendo filtradas por frequência de ocorrência no *British English web corpus*, a fim de que variações linguísticas da mesma palavra (por exemplo, *color* e *colour*) sejam eliminadas. Nos próximos passos, cada um dos termos será pesquisado automaticamente em um motor de busca da Internet, de onde serão obtidas 100 sentenças. Estas serão desambiguadas da mesma forma feita na segunda etapa (as 10 primeiras serão selecionadas).

Por fim, serão gerados automaticamente sinônimos para cada uma das MWEs. A lista resultante será encaminhada a um grupo de linguistas, que farão a ordenação das sugestões de acordo com sua complexidade.

Desta forma, teremos um *gold standard* que será utilizado para melhorar a performance de programas de simplificação automática, garantindo um maior e melhor auxílio no entendimento de textos por um falante não nativo. Em trabalho futuro, esta metodologia será aplicada ao português.