Evento	Salão UFRGS 2014: SIC - XXVI SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2014
Local	Porto Alegre
Título	Simplificação de textos usando paráfrases
Autor	FERNANDO LUÍS SPANIOL
Orientador	ALINE VILLAVICENCIO

Paráfrase é um refraseamento de palavras compostas que podem não ser compreendidas por pessoas, porém quando explicadas por meio de uma paráfrase são melhores entendidas. Esta sendo uma importante estrutura de linguagem que ajuda a descrever eventos do cotidiano e que é muito usada porém difícil de ser tratada. Palavras compostas são palavras que sozinhas possuem certos significados mas quando postas juntas significam outra coisa. Métodos de descrição são frases que fazem a explicação usada na paráfrase. Por exemplo, termos compostos, como *Orange Juice, Cable Network, Chocolate Bar,* e métodos de descrição, como *be composed of, be comprised of, be made of* formam as paráfrases: *juice that is made of orange, network that is made of cable.*

Este trabalho consiste em criar uma lista de paráfrases (inicialmente focamos na língua inglesa e posteriormente na língua portuguesa). A motivação deste projeto é poder realizar melhores simplificações de textos para que possam ser lidos por pessoas que não tenham a habilidade necessária para os ler em seu vocabulário original.

Su Nam Kim e Preslav Nakov propuseram um método para descobrir paráfrases a partir de uma lista inicial de termos compostos e métodos de descrição. Para isso, o método verifica as combinações da palavra composta com os métodos de descrição e verifica qual retorna o maior número de resultados em uma busca no Bing. Para a criação deste banco de dados de palavras compostas e métodos de descrição, se deve começar com uma lista de palavras compostas e fazer a pesquisa das mesmas na web, após isso encontrar sequências de palavras que podem ser consideradas como método de descrição. Em seguida usa-se os métodos de descrição que temos disponíveis até o momento e fazemos a pesquisa somente deles, verificando possíveis palavras compostas retornadas junto com eles e voltamos para o passo anterior, repetimos o ciclo até que o número de palavras compostas novas achadas seja igual a 0 ou até que o tempo limite que definimos tenha esgotado.

Posteriormente basta realizar o cruzamento das palavras compostas com os métodos de descrição e fazer a sua busca para ver quais tem melhor resultado. Até o momento temos cerca de 62000 palavras compostas e 22 métodos de descrição para a língua inglesa. Com o cruzamento das palavras com os métodos verificamos que só é retornado cerca de 55000 pesquisas com pelo menos 1 resultado, ou seja, muitos termos ainda não possuem nenhum meio de serem parafraseados, dificultando a criação de uma lista mais completa.

Outro trabalho feito foi o de análise de dados de pacientes com deficit cognitivo. O objetivo era analisar os testes dos pacientes e ver se existe uma linha de raciocínio comum entre eles baseado na similaridade de palavras adjacentes. O objetivo deste trabalho é o melhor entendimento de como funciona o cérebro de tais pacientes e ter um avanço no tratamento dos mesmos.

Os pacientes foram submetidos a testes nos quais eles deviam falar quantas palavras distintas eles conseguissem em 1 minuto. Por exemplo: *palavras começadas com A*, *animais* e *verbos*. Estes dados foram submetidos a linguistas que definiram quando existia quebra de contexto entre as palavras. Após receber o resultado, analisamos os pares de palavras adjacentes e calculamos uma série de dados usando a WordNet, uma biblioteca para o auxílio de trabalhos em Python, tais como: similaridades, frequência e distância de uma palavra a outra na árvore de palavras da WordNet. E após termos cerca de 10 dados para cada par de palavras, usamos o Weka, que é um programa de manipulação de dados, e geramos uma fórmula que analisando a sequência de dados, previa quando existia uma quebra de contexto ou não.

Ao aplicarmos a fórmula a todos nossos resultados, verificamos que o acerto da fórmula era de apenas 40%, o que não pode ser considerado como um resultado confiável e concluímos que não existe um meio de se prever uma quebra de contexto entre pacientes com deficit cognitivo por meio de testes com similaridade entre palavras.