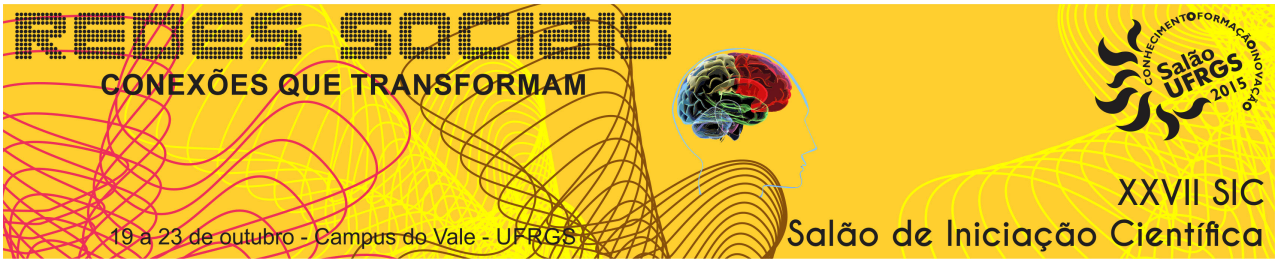

Evento	Salão UFRGS 2015: SIC - XXVII SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2015
Local	Porto Alegre - RS
Título	Preparação e análise de dados para estudos computacionais da aquisição de linguagem
Autor	JOÃO MARCOS FLACH
Orientador	ALINE VILLAVICENCIO

Projeto: Preparação e análise de dados para estudos computacionais da aquisição de linguagem
Autor: João Marcos Flach
Orientadora: Aline Villavicencio
Instituição: UFRGS
Resumo:
Substantivos compostos (doravante chamados SCs) como *orange juice* (suco de laranja) e *police car* (carro de polícia) são desafiadores para o processamento de textos, pois a relação entre os substantivos que os formam está implícita. SCs são comuns em textos e entender suas semânticas é importante para muitas aplicações de processamento de linguagem natural. Por exemplo, uma aplicação de tradução poderia traduzir *data base* literalmente como "base de dados", ou utilizar interpretações de SC e traduzir como "banco de dados" (tradução mais aceita).
Existem SCs de vários tipos, que dependem da classe das palavras que os formam. Alguns exemplos são: substantivo + substantivo = *bus stop* (parada de ônibus), adjetivo + substantivo = *full moon* (lua cheia), verbo + substantivo = *washing machine* (máquina de lavar). Além do tipo, o tamanho também pode variar. Como exemplo de uma SC com mais de duas palavras, temos *colon cancer tumor suppressor protein* (proteína supressora de tumor cancerígeno do colon). Este trabalho se foca nos SCs do tipo substantivo + substantivo com duas palavras.
Existem duas linhas gerais de pesquisa: a primeira deriva a semântica do SC da semântica dos nomes que o formam. A segunda modela a relação entre os nomes. Em qualquer caso, a semantica de um SC é tipicamente expressa por uma relação abstrata como CAUSA (*malaria mosquito* é um mosquito que causa malária), FONTE (*olive oil* é um óleo feito de oliva) ou PROPÓSITO (*migraine drug* é uma droga para enxaqueca), vinda de um pequeno inventário fixo. Na interpretação de SCs, verbos e preposições podem ser visto como padrões conectando os dois substantivos. Por exemplo, *iron knife* (faca de ferro) pode ser visto como *knife that is made of iron* (faca que é feita de ferro). Neste caso, *is made of* é um padrão que interpreta a relação de FONTE entre os substantivos.
Para ajudar nessa área de pesquisa, seria bom ter um grande conjunto de SCs para serem analizados. Então, o objetivo é construir um conjunto de centenas de milhares de SCs (primeiramente em inglês, reproduzindo o trabalho de *Kim e Nakov, 2011*, depois em português do Brasil), com cada um interpretado: (a) por uma relação semântica abstrata e (b) por um conjunto de verbos.
Para isso, é utilizada a linguagem *Python 2.7* com as bibliotecas *nltk* (para funções gerais de processamento de linguagem natural), *re* (para expressões regulares), *itertools* (para algumas funções de combinação de conjuntos), *glob* (para lidar com os arquivos de entrada/saída) e *pattern.en* (para funções de lematização). Além disso, utilizamos grandes quantidades de textos (doravante chamados *corpus*) sem anotações (formato .txt) da biblioteca *nltk*.
O algoritmo utiliza um processo de *bootstrapping*, o que significa que dado uma entrada inicial (neste caso, um pequeno conjunto de padrões) o programa se auto-sustenta a partir de então. Existem dois passos principais no algoritmo:
(a) Extração de SCs: para cada um dos padrões de entrada, todo o *corpus* é vasculhado para extrair os potenciais SCs. Depois de extraídos, eles são lematizados e então são aplicados filtros para eliminar possíveis erros.
(b) Extração de padrões: Para cada um dos SCs coletados, todo o *corpus* é vasculhado para extrair os potenciais padrões. Depois de extraídos, são aplicados filtros para eliminar possíveis erros.
Assim como esperado, os resultados obtidos até agora não diferem dos obtidos por *Kim e Nakov, 2011*. Como trabalho futuro, temos a replicação do trabalho para o português do Brasil, com a construção de um grande conjunto de SCs, que ajudará no avanço da área.