SALÃO DE INICIAÇÃO CIENTÍFICA XXIX SIC

UFRGS PROPESQ

Salão UFRGS 2017

múltipla UNIVERSIDADE inovadora inspiradora

Evento	Salão UFRGS 2017: SIC - XXIX SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2017
Local	Campus do Vale
Título	Atributos Para Predição Da Composicionalidade De Compostos Nominais Em Inglês
Autor	MATHEUS WESTHELLE
Orientador	ALINE VILLAVICENCIO

Atributos Para Predição Da Composicionalidade De Compostos Nominais Em Inglês

Orientador: Aline Villavicencio
Autor: Matheus Westhelle
Universidade Federal Do Rio Grande Do Sul

Determinar se um composto nominal é composicional (por exemplo “police car”) ou idiomático (“loan shark”) se mostra muito útil em tarefas como simplificação lexical, e mais útil ainda no caso específico de compostos idiomáticos, em que o significado é opaco tomando apenas o significado de suas partes; por exemplo, não é intuitivo inferir que *loan shark* significa *agiota* apenas pelas palavras *loan* e *shark*. Nesse contexto, o presente trabalho visa investigar atributos que possivelmente indicariam a composicionalidade de compostos nominais. O trabalho procura complementar outro trabalho de Cordeiro et al, que procurava predizer a composicionalidade de compostos nominais através do uso de modelos de semântica distribucional. A primeira ideia de atributo que tivemos foi baseada no funcionamento de redes neurais recorrentes (RNN); RNNs são ferramentas eficientes para modelagem de sequências, sendo usadas para a modelagem de linguagem. Uma RNN tem um *hidden state* (representado por um vetor multidimensional), um estado atual que guarda informações da sequência que podem ser usadas para prever a próxima palavra de uma frase, por exemplo. A hipótese é que, na passagem da primeira palavra do composto para a segunda, esse estado sofreria uma mudança marcada para o caso de compostos idiomáticos, pois eles seriam mais imprevisíveis que compostos composicionais. Outra hipótese para um atributo foi calcular a razão da probabilidade de um composto nominal aparecer dado um contexto (uma frase) e o produto das probabilidades de cada composto aparecer no mesmo contexto: P(w1w2|c) / (P(w1|c) * P(w2|c). A fim de testar essas hipóteses, nos valemos do *dataset* utilizado no trabalho de Cordeiro et al, que contém frases de exemplo para cada composto nominal (3 frases para cada um de 90 compostos), bem como escores de composicionalidade de cada um obtido por avaliação humana, indo de 1 (totalmente idiomático) a 5 (totalmente composicional). Treinamos um modelo de linguagem usando RNN e o corpus text8 para testar a hipótese sobre RNNs, utilizando-o para obter os *hidden states* das frases de exemplo até a primeira e até a segunda palavra do composto (portanto, dois vetores) e obtivemos o cosseno desses vetores para verificar sua similaridade. Verificamos então o coeficiente de correlação Spearman entre esse resultado e os escores do *dataset*. Para testar a outra hipótese, usamos um modelo de linguagem já treinado com o corpus UKWaC para obter os valores das probabilidades e fazer o cálculo já mencionado, avaliando o resultado da mesma forma que na primeira hipótese. Para a primeira hipótese, obtivemos $\rho=0{,}144$ e para a segunda, $\rho=0{,}172$, valores muito baixos que refutam essas hipóteses. Prosseguimos investigando outros atributos, como entropia (uma medida de imprevisibilidade) de matriz de contagem de bigramas (conjuntos de duas palavras), por exemplo: talvez uma entropia maior signifique que o composto é idiomático, e vice-versa para o caso de ser composicional. Como trabalho futuro, pretendemos explorar outras ideias e também verificar se existe alguma possível interação de atributos que produza resultados interessantes.