Atributos para Predição da Composicionalidade de Compostos Nominais em Inglês

Matheus Westhelle
matheus.westhelle@inf.ufrgs.br

Aline Villavicencio
avillavicencio@inf.ufrgs.br

Motivação

Determinar a idiomaticidade de compostos nominais é útil pro desafio de compreensão de linguagem natural e tarefas como simplificação lexical.

Loan shark

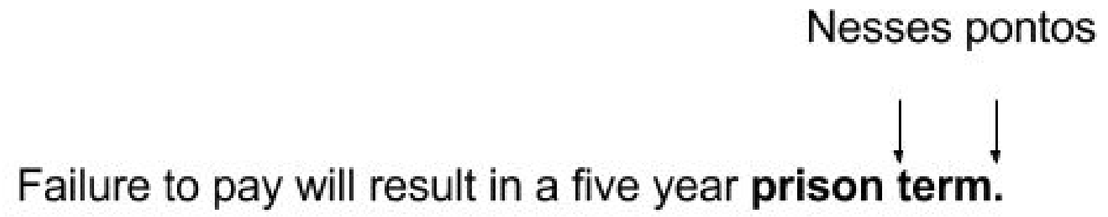
Police car

Dead end

Eager beaver

Hipóteses

RNN: RNNs capturam idiomaticidade e isto é refletido nos seus estados ocultos.



PMI: Medidas de associação estatística capturam a idiomaticidade e isto é refletido numa maior forca de associacão.

$$\frac{P(w1w2|c)}{P(w1|c)*P(w2|c)}$$

Metodologia

- Usamos o dataset *Compositionality of Nominal Compounds*, que contém frases de exemplo e escores de composicionalidade (variando de 1 a 5, obtido pela avaliação de humanos)
- Primeira hipótese:
  - Modelo treinado no corpus text8
  - Frases de exemplo usadas para obter hidden states
  - Obtivemos o cosseno dos vetores de hidden state para verificar similaridade
- Segunda hipótese:
  - Probabilidades obtidas de modelo treinado em Wikipedia + NewsCrawl
- Correlação Spearman dos resultados de cada hipótese com os escores de composicionalidade do dataset

Resultados e Trabalho Futuro

- Primeira hipótese: ρ=0,144
- Segunda hipótese: ρ=0,172

- Treinar modelo com corpus maior
- Investigar outras features