Nome: Rafael Claro Ito (R.A.: 118430)
Resumo do artigo: Efficient Estimation of Word Representations in Vector Space

Neste artigo são propostas duas novas arquiteturas com o objetivo de se criar representações de palavras através de um vetor contínuo, partindo de datasets com grande volume de dados. A principal proposta dessas arquiteturas é apresentar uma melhora significativa de representação, com um pequeno custo computacional adicionado aos modelos anteriores. Além dessas duas arquiteturas propostas, outros três pontos fortes do artigo, segundo minha opinião, são: o desenvolvimento apresentado para medidas de similaridades sintáticas e semânticas entre palavras (feitas algebricamente), a criação do dataset "*Semantic-Syntactic Word Relationship test set*" e o forma de treinamento paralelo implementada em um *framework* chamado *"DistBelief"*.

Modelos mais simples como N-gram são capazes de lidar com volume grandes de dados. Entretanto, para tarefas mais complexas (reconhecimento automático de fala ou tradução), modelos mais complexos são necessários. A principal ideia aqui é tentar criar representações de vetores de palavras partindo de dataset com bilhões de palavras e milhões de palavras no vocabulário.

Antes de descrever as arquiteturas propostas, os autores citam dois tipos de redes: a NNLM (*Feedforward Neural Net Language Model*) e a RNNLM (*Recurrent Neural Net Language Model*). A fim de comparar a complexidade computacional dessas redes com as propostas no artigo, é dado que a complexidade computacional (*O*) é proporcional ao número de épocas do treinamento (*E*), número de palavras no dataset de treinamento (*T*) e um valor *Q: O = E x T x Q*

Para a rede NNLM, é mostrado que: $Q = (N \times D) + (N \times D \times H) + (H \times V)$, sendo $(H \times V)$ o termo dominante. Usando *softmax* hierárquica com o vocabulário sendo representado por uma árvore binária de Huffman, é possível reduzir esse termo para $(H \times log_2(V))$. Neste caso, o termo dominante passa a ser $(N \times D \times H)$.

Para a rede RNNLM, temos como principal característica a matriz recorrente que conecta a camada escondida à ela mesma, usando conexões *time-delayed*. Isso dá ao modelo uma memória de curto prazo (*short term memory*). A complexidade do modelo é dada por: $Q = (H \times H) + (H \times V)$, e, usando a mesma abordagem da rede NNLM, resulta no termo dominante ($H \times H$).

Finalmente, são propostos os modelos do artigo: CBOW (*Continuous Bag-of-Words Model)* e Skip-gram (*Continuous Skip-gram Model*), sendo ambos modelos log-linear. Esses modelos tentam diminuir a complexidade exposta nos dois parágrafos anteriores causada pela camada escondida não-linear. A ideia é trabalhar com modelos mais simples, mas passíveis de serem treinados com muito mais dados.

O modelo CBOW tenta prever a palavra atual, baseada no contexto *x* palavras anteriores e *x* palavras futuras. A camada escondida não-linear é removida e camada de projeção é compartilhada entre todas palavras, fazendo-as serem projetadas na mesma posição. Assim, a complexidade desse modelo é dada por: $Q = (N \times D) + (D \times log_2(V))$

O modelo Skip-gram tenta maximizar a classificação de uma palavra baseado em outra palavra na mesma frase. Nesse caso, a palavra atual é usada como entrada de um classificador não-linear com uma camada de projeção contínua, cuja saída são palavras antes e depois da atual dada uma distância *C*. Sua complexidade é dada por: $Q = C \times (D + D \times log_2(V))$, onde C é a distância máxima das palavras.

Os resultados mostrados comparam diferentes modelos, treinados com diferentes número de palavras, e até mesmo por épocas diferentes. Como conclusão, tem-se com esses modelos propostos (CBOW e Skip-gram) performance estado da arte para medidas de similaridades sintáticas e semânticas entre palavras, medidas através do dataset criado pelos autores.