Assignment - Research Paper Summary

Devashish Kamble

The following is a summary of the paper -

DON'T COUNT, PREDICT! A systematic comparison of context-counting vs. context-predicting semantic vectors

1 Introduction

By comparing neural networks (NN) with the traditional count-based models, the authors provide insights on the influence of NNs within the Distributional Semantics domain, which is grounded on the Distributional Hypothesis: *similarity of meaning correlates with similarity of distribution.* A typical way to get an approximation of the meaning of words in distributional semantics is to look at their linguistic context (i.e the other words that appear next to it) in a very large corpus.

1.1 Count Models

In traditional count models, the frequency of certain context word next to the main word is considered by choosing a window size. In order to achieve higher performance two weighting schemes, **Pointwise Mutual Information** and **Local Mutual Information** were considered. The vectors are also compressed using **SVD** and **Non-negative Matrix Factorization** that produces easy to use smaller dense vectors.

1.2 Predict Models

The authors considered the word2vec implementation of **Continuous Bag-of-Words (CBOW)** as their choice of predict model. CBOW takes vectors of context words as input and predicts the main word. In order to speed up the probability computations, **Hierarchical Softmax** and **Negative Sampling** are applied. The word2vec also downsizes the function words as they are not informative.

2 Evaluation

The models were trained on a corpus of 2.8 billion tokens, combining ukWac, the English Wikipedia and the British National Corpus. The models were tested on the following benchmarks:

Benchmark	Task	Models
Semantic relatedness	Compute correlation based on the relatedness score between two words	**rg, ws, wss, wsr, men**
Synonym detection	Find the correct synonym for a word based on multiple choices	**toefl**
Concept categorization	Group concepts into semantic categories	**ap, esslli, battig**
Selectional preferences	Decide how likely a noun is to be a subject or object for a specific verb	**up, mcrae**
Analogy	Find suitable word to solve analogy questions	**an, ansyn, ansem**

3 Results

The neural networks win by a large margin, showing extraordinary performance for the semantic relatedness, synonym detection and analogy detection benchmarks and equal or slightly better on categorization and selectional preferences.

Best Parameter Choice				
Model	Window	Weight	Dimension	
Count	2	PMI	300K	No compression
Predict	5	Subsampling of frequent words	400	Negative Sampling

4 Conclusion

The results are in favor of the claim that predict models (NNs) do a better job at distributional similarity tasks. The way forward is to focus on parameters and extensions of the predict models to make them even better.

rg: 65 noun pairs
wsr: Relatedness subset Wordsim353
ap: 402 concepts in 21 categories
up: 221 word pairs
ansyn: Syntax subset of analogy

ws: 353 Wordsim353 word pairs
men: 1000 word pairs
esslli: 44 concepts in 6 categories
mcrae: 100 noun-verb pairs
ansem: Semantic subset of analogy

wss: Similarity subset Wordsim353
toefl: 80 mcqs with 4 synonyms
battig: 83 concepts in 10 categories
an: 19,500 analogy questions