Document Embedding for Scientific Articles: Validation of word embeddings

Arjan Meijer, 11425555

April 9, 2018

1 Abstract

[TODO: Abstract]

2 Document Embedding

The basic theory of word embedding can be explained by the following quote of J.R. Firth: "You shall know a word by the company it keeps".
Word embedding as described by Lai et al, can capture both the semantic and syntactic information of words from a large unlabeled corpus.[1] Word embeddings are vector based representations of words, that can, depending on the model either predict the target word given context words, or predict context words, given the target word. Techniques based on word vectors have improved various NLP areas such as machine translation, [todo: add fields and references]

2.1 domain specific

Earlier work on this topic, concerning academic articles, by Truong[TODO: CITE] states the following: "Our findings clearly evinced that in-domain training of the word embeddings can drastically improve the process of document clustering. In fact, the effect is even stronger than the number of training examples and the model architecture. However, too isolated training can lead to a failure of several clustering algorithms, such as DBSCAN, due to a too dense vocabulary". This was also found by Lai et al, who state: "The corpus domain is more important than the corpus size. Using an in-domain corpus significantly improves the performance for a given task, whereas using a corpus in an unsuitable domain may decrease performance". These statements both indicate that an in-domain corpus improves the performance of word-vectors for those specific domains. However, domain specific validation techniques do not exist currently. Multiple generic validation sets have been published such as:

the Rare-word dataset introduced by Luong et al in their paper "Better Word Representations with Recursive Neural Networks for Morphology"[**?**], the MEN test collection[**?**] and the WordSimilarity-353 test collection[**?**].

These sets have been used in multiple studies of word-vectors[TODO REFERENCE] which corpus existed of similar generic information [TODO: INSERT CORPUS DETAILS]. These validation methods are limited to the non domain specific fields.

2.2 Embedding validation

Not published results of the study by Truong show high error rates on the validation scores, while the word-vectors seem to work well on document clustering. This seems to indicate that the word-vectors are well suited, but no objective validation method can confirm or deny this. Therefore, we propose the validation of word-embeddings through [SOLUTION].

	WordSim	Men	RareWords
Lowest	0.38	0.54	0.29
Highest	0.49	0.61	0.32
Average	0.45	0.59	0.32
Baseline	0.64	0.68	0.34

References

[1] Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14, 2016.