Uncertainty in Neural Network Word Embedding
Exploration of a Threshold for Similarity

Navid Rekabsaz, Mihai Lupu, Allan Hanbury
TU Wien, Vienna, Austria
surname@ifs.tuwien.ac.at

Problem

- In word embedding models, one can measure the ***semantic similarity*** of *any* two words, e.g. by Cosine function between the vector representations of the terms
- However an essential challenge in using this similarity measure is:

How to distinguish between similar and dissimilar terms?

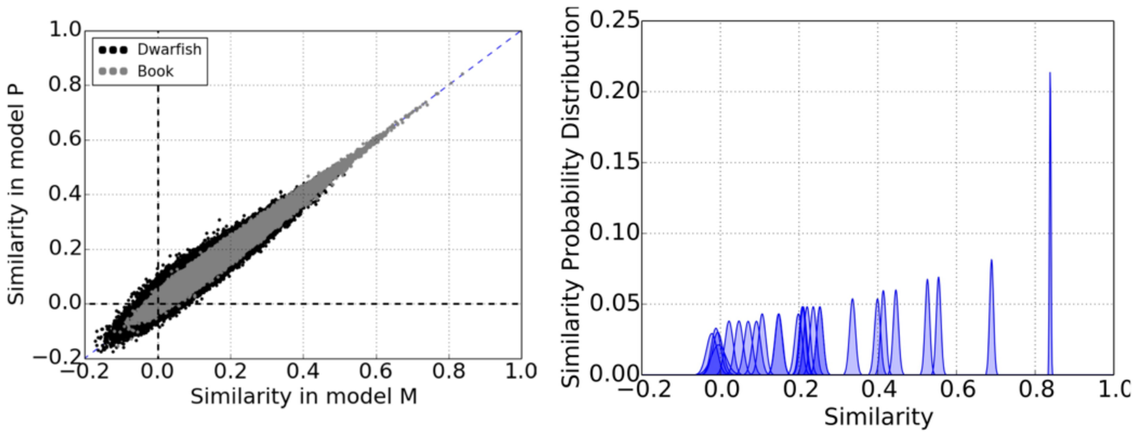
Proposed Solution

- We suggest a ***general threshold*** which effectively filters out dissimilar/less similar terms
- We explore this threshold through ***inherent uncertainty*** in the similarity of neural network-based word embedding models

Term Similarity Threshold

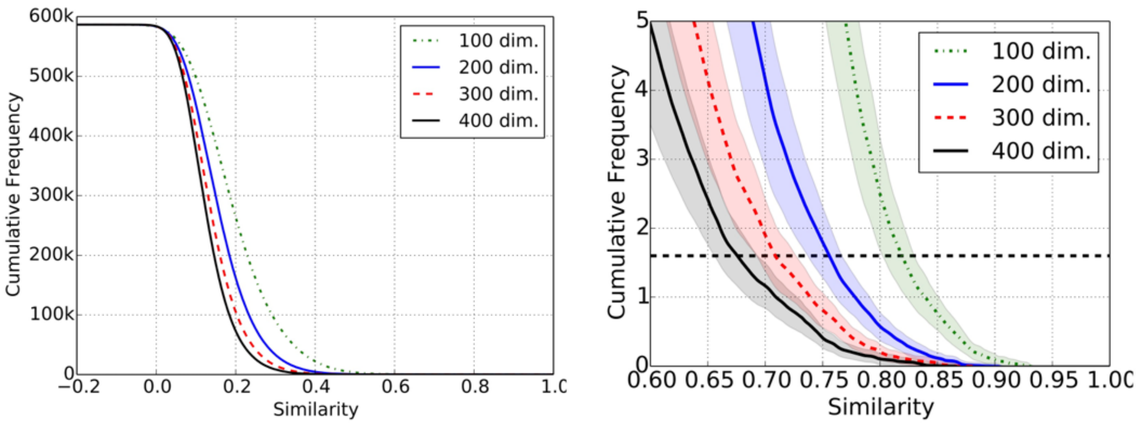
Uncertainty in Similarity:

Left: Similarity of two words to all terms in two models.
Right: Term-to-term similarity as probability distribution.



Continuous Distribution of Neighbours:

- Create the Cumulative Distribution Function (CDF) of similarities for 100 terms and average their values.
- Y-values: The expected number of neighbors.



Threshold:

- Average # of similar terms per term in WordNet is ~1.6
- Our thresholds is achieved by projecting 1.6 on CDF diagram.

Dimensionality	Threshold Boundaries Lower	**Main**	Upper
100	0.802	**0.818**	0.829
200	0.737	**0.756**	0.767
300	0.692	**0.708**	0.726
400	0.655	**0.675**	0.693

Experiments on Document Retrieval

To use word embeddings in retrieval models, we use:
Generalized Translation Models (GT) &
Extended Translation Models (ET)
for *BM25* and *Language Modeling (LM)*.

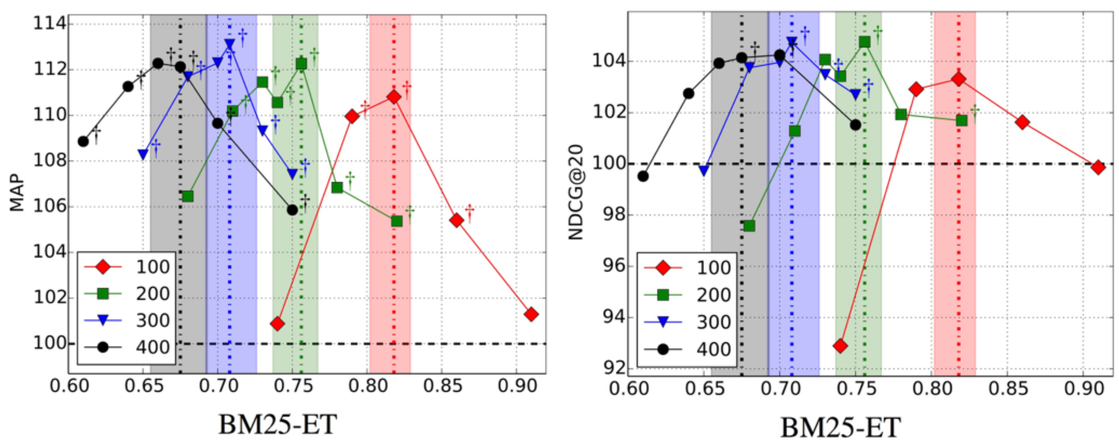
Rekabsaz et al. [2016], Zuccon et al. [2015]

Experiments:

Evaluation measures:
MAP, NDCG

Name	Collection	# Documents
TREC 6	Disc4&5	551873
TREC 7 and 8	Disc4&5 without CR	523951
HARD 2005	AQUAINT	1033461

- We exhaustively evaluate the collections on various thresholds in [0.6-1.0] and dimensions of [100-400].
- Significance test against the original models without word embedding (BM25 and LM).



Results:

- The optimal results in all the dimensions are the same or in the range of the proposed thresholds.
- Results with the optimal thresholds are significantly better than the original models.

Conclusion

- We propose a novel representation of the probability of the # of neighbors around a term
- We propose optimal thresholds to select effective terms in different dimensions

This work is funded by: Self-Optimizer (FFG 852624) in the EUROSTARS programme, funded by EUREKA, the BMWFW and the European Union, and ADMIRE (P 25905-N23) by FWF.

FACULTY OF !NFORMATICS