TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Exploration of a Threshold for Similarity based on Uncertainty in Word Embedding

Navid Rekabsaz, Mihai Lupu, Allan Hanbury


@NRekabsaz


rekabsaz@ifs.tuwien.ac.at

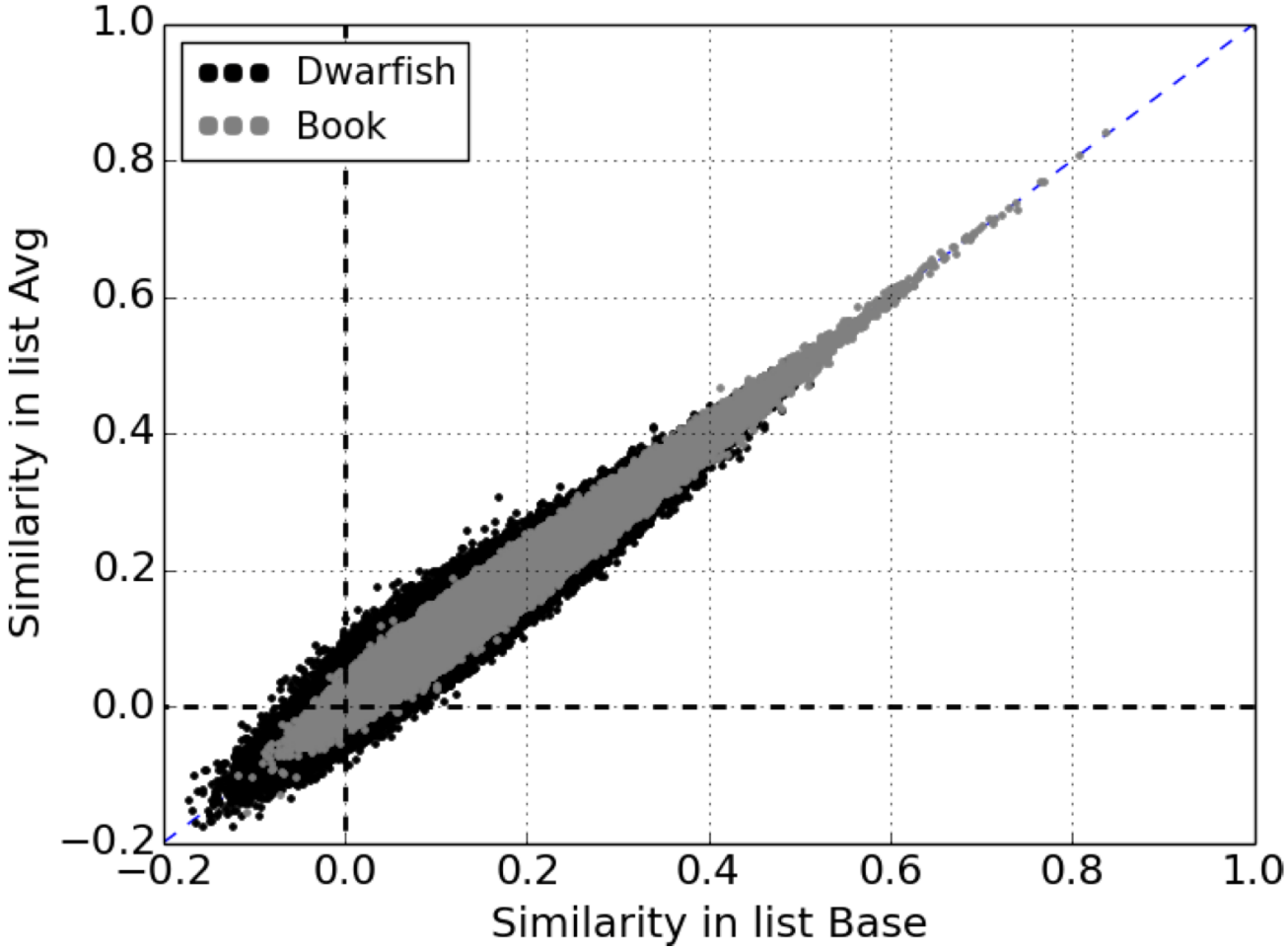
European Conference of Information Retrieval (ECIR)
Aberdeen, April 2017

Word Embedding

journalist	
reporter	0.78
freelance_journalist	0.74
investigative_journalist	0.74
photojournalist	0.73
correspondent	0.71
investigative_reporter	0.68
writer	0.64
freelance_reporter	0.63
newsman	0.61

dwarfish	
corpulent	0.44
hideous	0.43
unintelligent	0.42
wizened	0.42
catoblepas	0.42
creature	0.42
humanoid	0.41
grotesquely	0.41
tomtar	0.41

Uncertainty

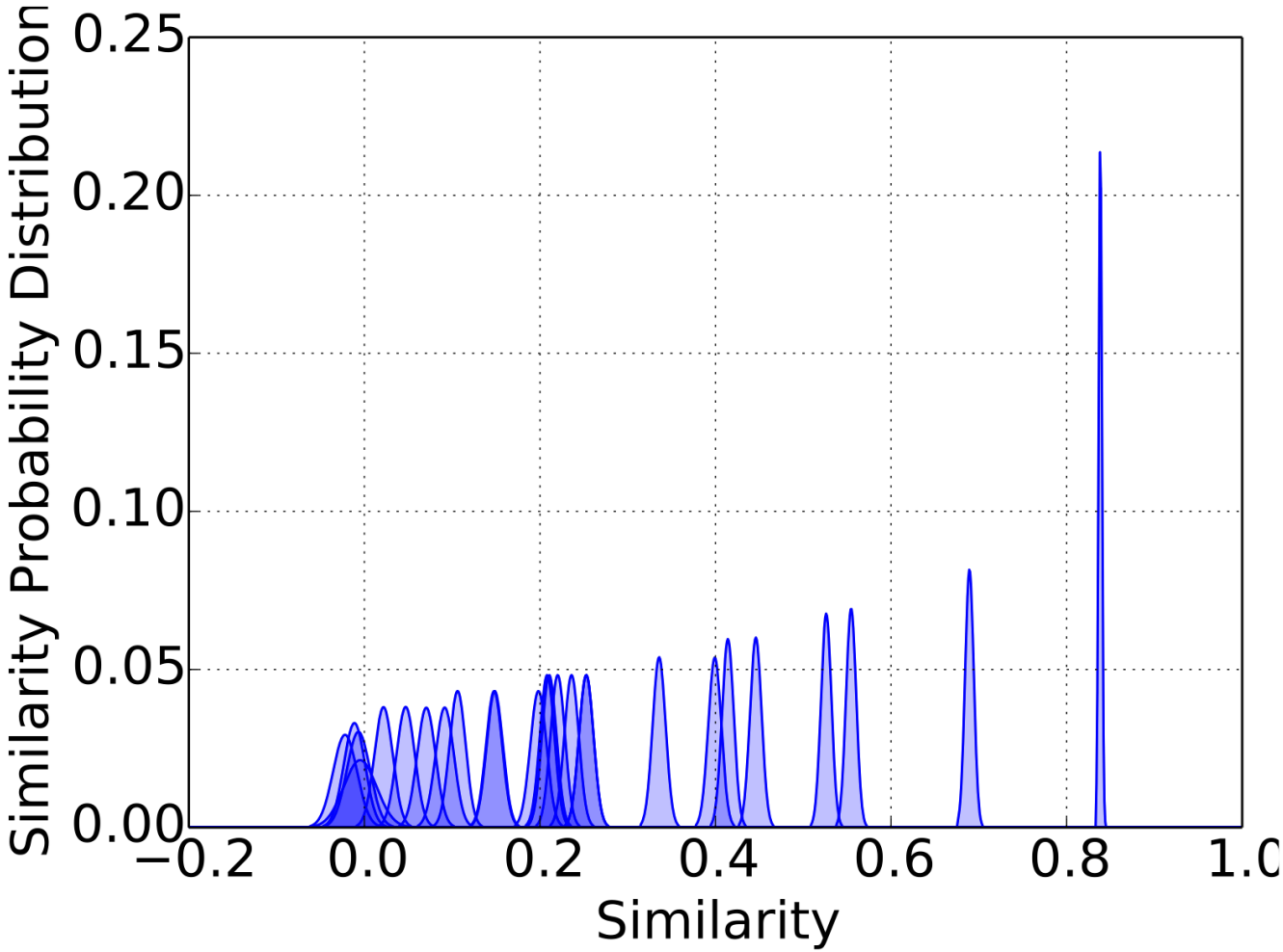


Uncertainty: $$\varrho(s) = \frac{1}{|\mathcal{S}_s|} \sum_{(x,y)\in\mathcal{S}_s} |sim(\vec{x}_M, \vec{y}_M) - sim(\vec{x}_P, \vec{y}_P)|$$

$$\mathcal{S}_s = \{(x,y) : sim(\vec{x}_M, \vec{y}_M) \in (s, s+\epsilon)\}$$

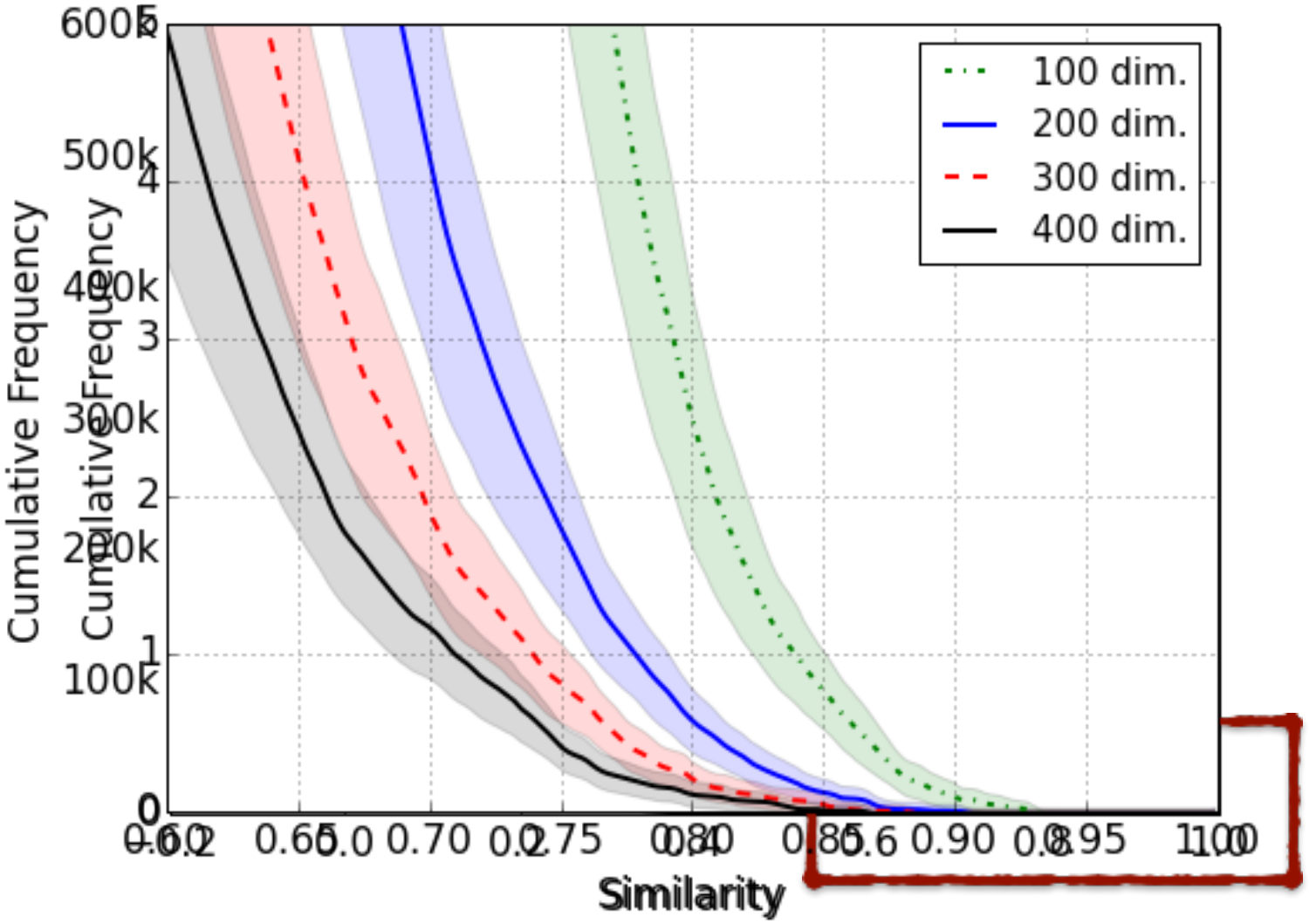
Similarity Probability Distribution

- Similarity between terms as probability distribution
- Normal distribution on observed similarities of 5 ‘identical’ models



Cumulative Similarity Distributions

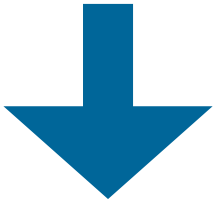
Y axes: Expected number of neighbors in a similarity value, averaged over 100 terms



Filtering Neighbors

What is the best threshold for filtering the related terms?

Hypothesis: it can be estimated based on the average number of synonyms over the terms



What is the expected number of synonyms for a word in English?

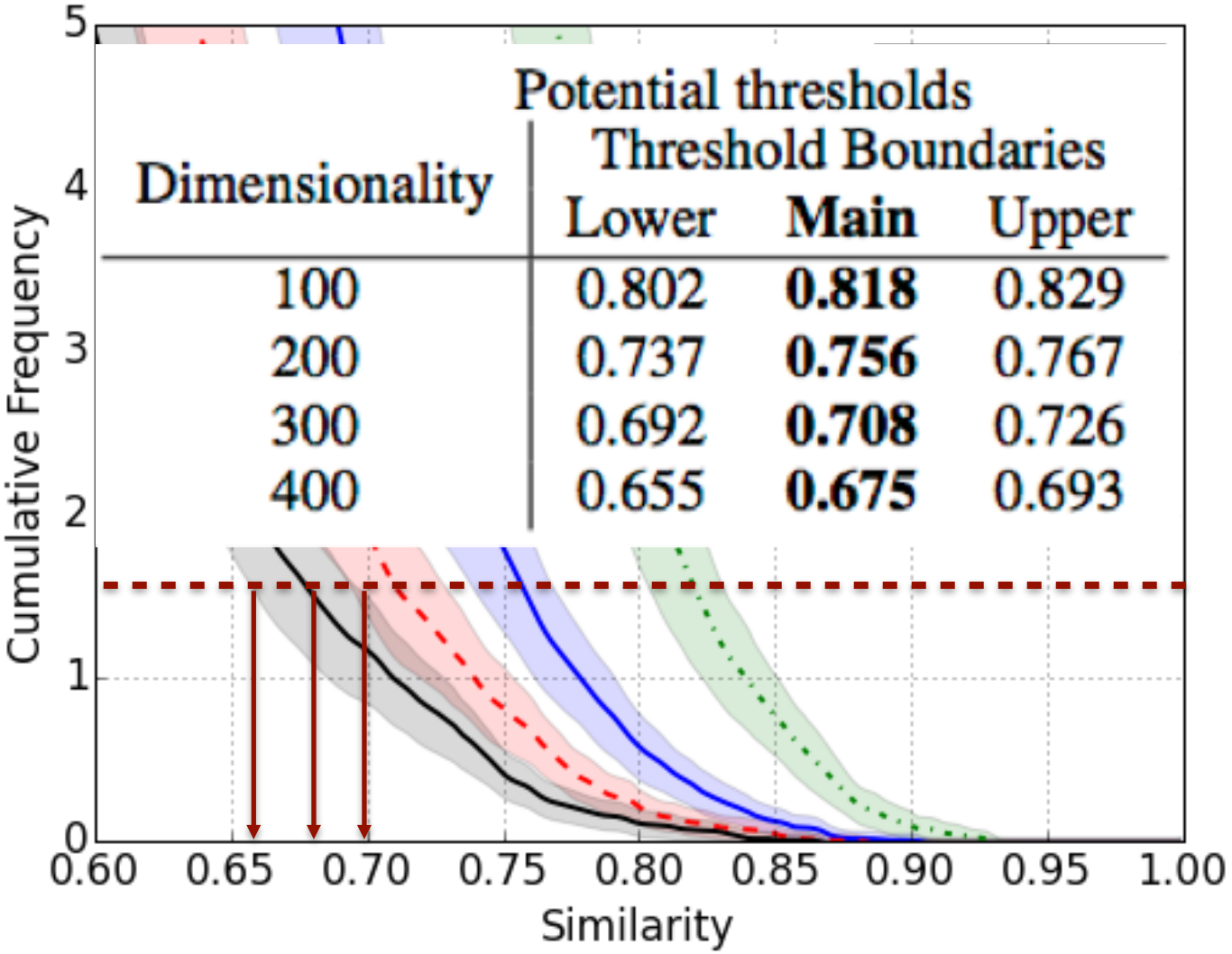
# of terms: 147306

Average # of synonyms per term: **1.6**

Standard deviation : 3.1

Threshold

Proposed Threshold: cumulative frequency equal to 1.6



Integrating Similarity in IR Models

BM25

$$\sum_{t \in T_d \cap T_q} \frac{(k_1+1)\overline{tf_d(t)}}{k_1+\overline{tf_d(t)}} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \log \frac{|D|+0.5}{df_t+0.5}$$

BM25 Generalized Translation (GT)

$$\sum_{t \in \widehat{T}_d \cap T_q} \frac{(k_1+1)\overline{\widehat{tf}_d(t)}}{k_1+\overline{\widehat{tf}_d(t)}} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \log \frac{|D|+0.5}{df_t+0.5}$$

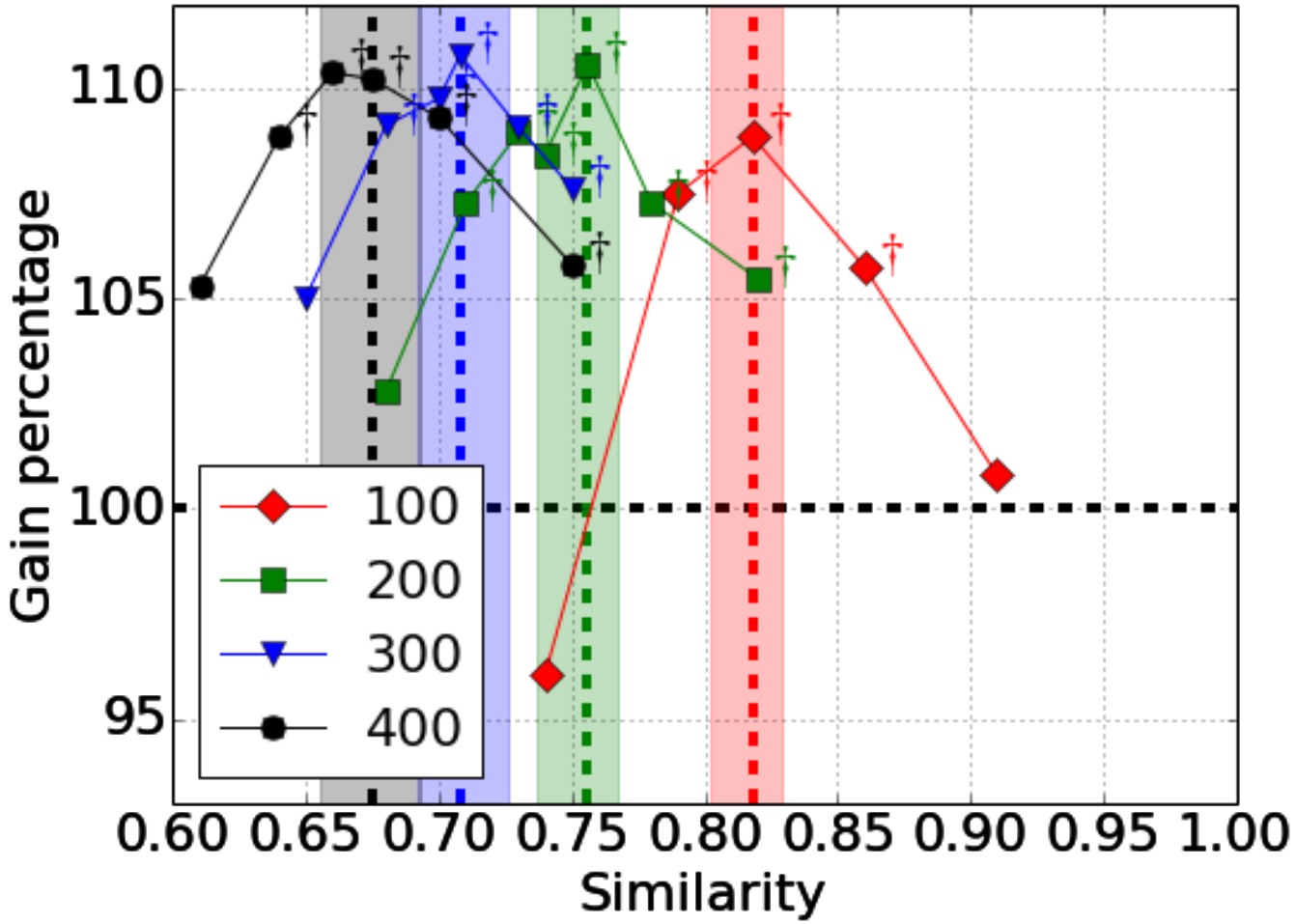
BM25 Extended Translation (ET)

$$\sum_{t \in \widehat{T}_d \cap T_q} \frac{(k_1+1)\overline{\widehat{tf}_d(t)}}{k_1+\overline{\widehat{tf}_d(t)}} \frac{(k_3+1)tf_q(t)}{k_3+tf_q(t)} \log \frac{|D|+0.5}{\widehat{df}_t+0.5}$$

Generalizing Translation Models in the Probabilistic Relevance Framework
Rekabsaz et al., CIKM 2016

Experiments Results

- Gain of MAP over standard BM25, averaged on collections.
- Optimal threshold is either the same or in the confidence interval of the proposed threshold.



Take Home Message

WE OBSERVED

- **Uncertainty** in **similarity value** of neural network word embedding models:
  - depends on similarity range
  - depends on dimensionality

WE PROPUSE

- **Threshold** to filter **most similar terms** :
  - Proposed threshold as good as optimal threshold

Come for a chat!

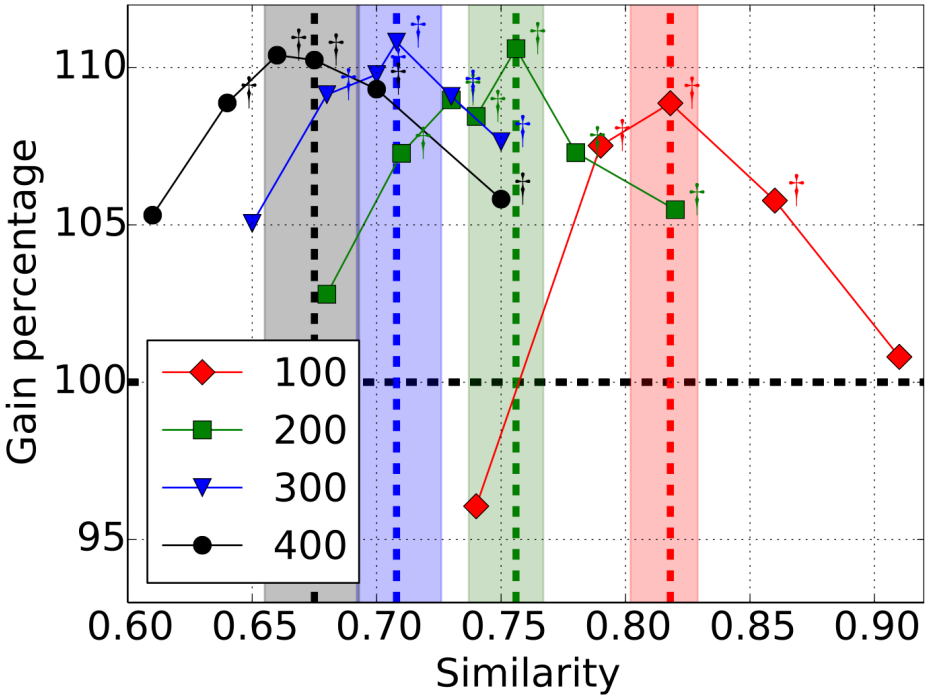
@NRekabsaz

rekabsaz@ifs.tuwien.ac.at

Threshold vs. TopN

- Conclusion2: Threshold outperforms TopN

Threshold-based



TopN