TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

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

NeuIR at SIGIR, 21st July 2016

Navid Rekabsaz, Mihai Lupu, Allan Hanbury


@NRekabsaz


rekabsaz@ifs.tuwien.ac.at

Similar Terms

Similar Terms

book		dwarfish	
books	0.82	corpulent	0.44
foreword	0.77	hideous	0.43
author	0.74	unintelligen	0.42
published	0.73	wizened	0.42
preface	0.69	catoblepas	0.42
republished	0.68	creature	0.42
reprinted	0.68	humanoid	0.41
afterword	0.67	grotesquely	0.41
memoir	0.67	tomtar	0.41

Similar Terms

TopN:

book		dwarfish	
books	0.82	corpulent	0.44
foreword	0.77	hideous	0.43
author	0.74	unintelligen	0.42
published	0.73	wizened	0.42
preface	0.69	catoblepas	0.42
republished	0.68	creature	0.42
reprinted	0.68	humanoid	0.41
afterword	0.67	grotesquely	0.41
memoir	0.67	tomtar	0.41

Similar Terms

TopN:

book		dwarfish	
books	0.82	corpulent	0.44
foreword	0.77	hideous	0.43
author	0.74	unintelligen	0.42
published	0.73	wizened	0.42
preface	0.69	catoblepas	0.42
republished	0.68	creature	0.42
reprinted	0.68	humanoid	0.41
afterword	0.67	grotesquely	0.41
memoir	0.67	tomtar	0.41

Using ***threshold*** in SIGIR 2016, tuned by brute-force search:

Scalable Semantic Matching of Queries to Ads in Sponsored Search Advertising.
Mihajlo Grbovic et al.

Robust and Collective Entity Disambiguation through Semantic Embeddings.
Stefan Zwicklbauer et al.

Contribution

Contribution

Our Contribution

- Analytical exploration of a general **threshold** on term similarity
  - Defined for all the terms in the lexicon
  - Tested on Ad Hoc retrieval
- Showing the advantage of threshold-based rather than TopN-based approach

Contribution

Our Contribution

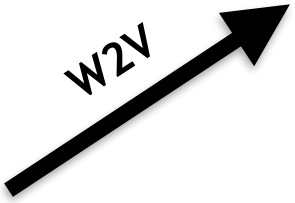
- Analytical exploration of a general **threshold** on term similarity
  - Defined for all the terms in the lexicon
  - Tested on Ad Hoc retrieval
- Showing the advantage of threshold-based rather than TopN-based approach

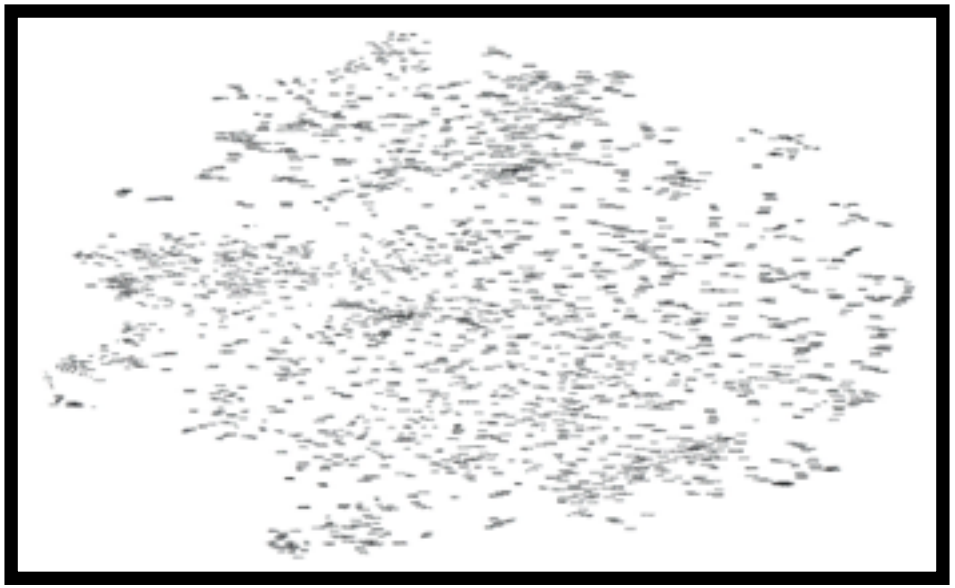
Roadmap

- *Uncertainty* in NN-based word embeddings
- Similarity distribution
- Proposing threshold
- Experiments

Observations

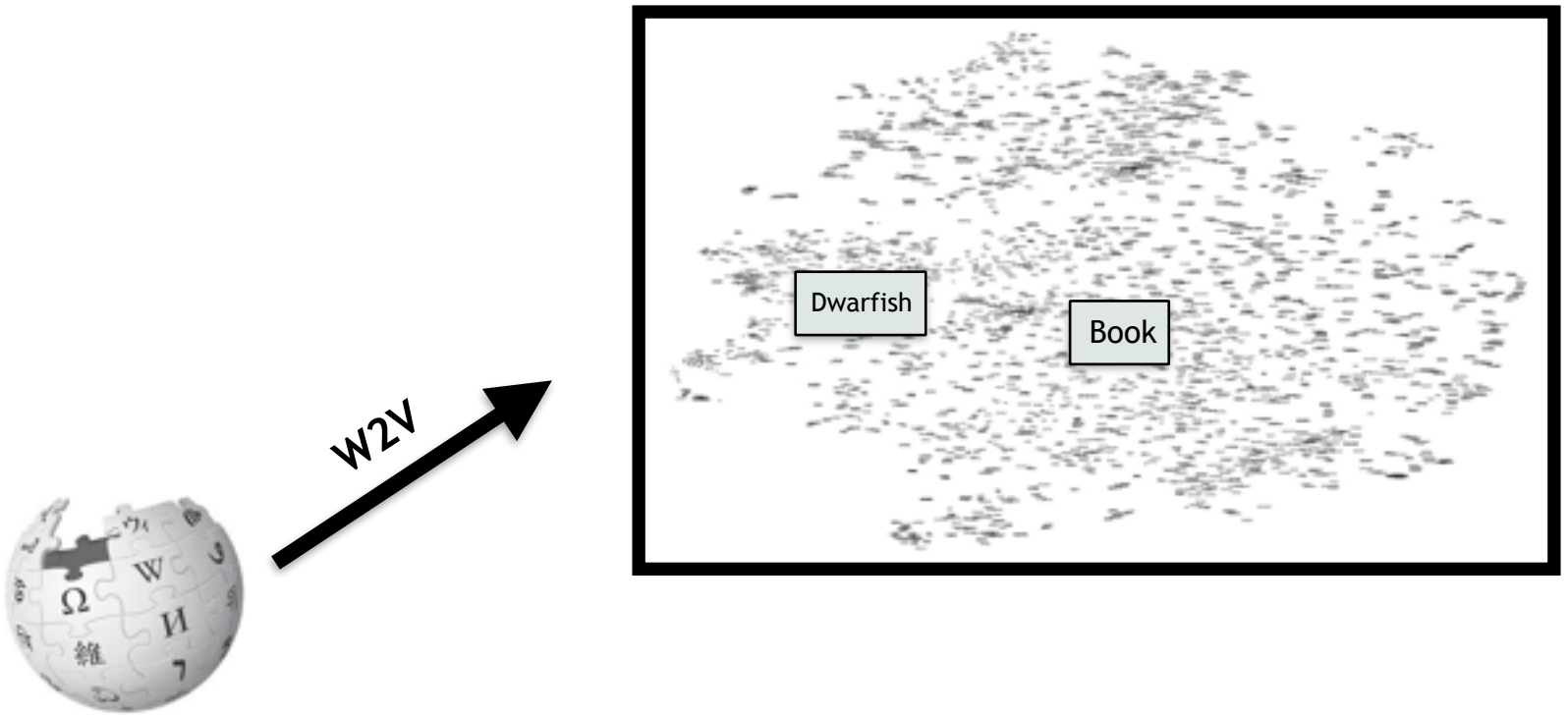
Observations





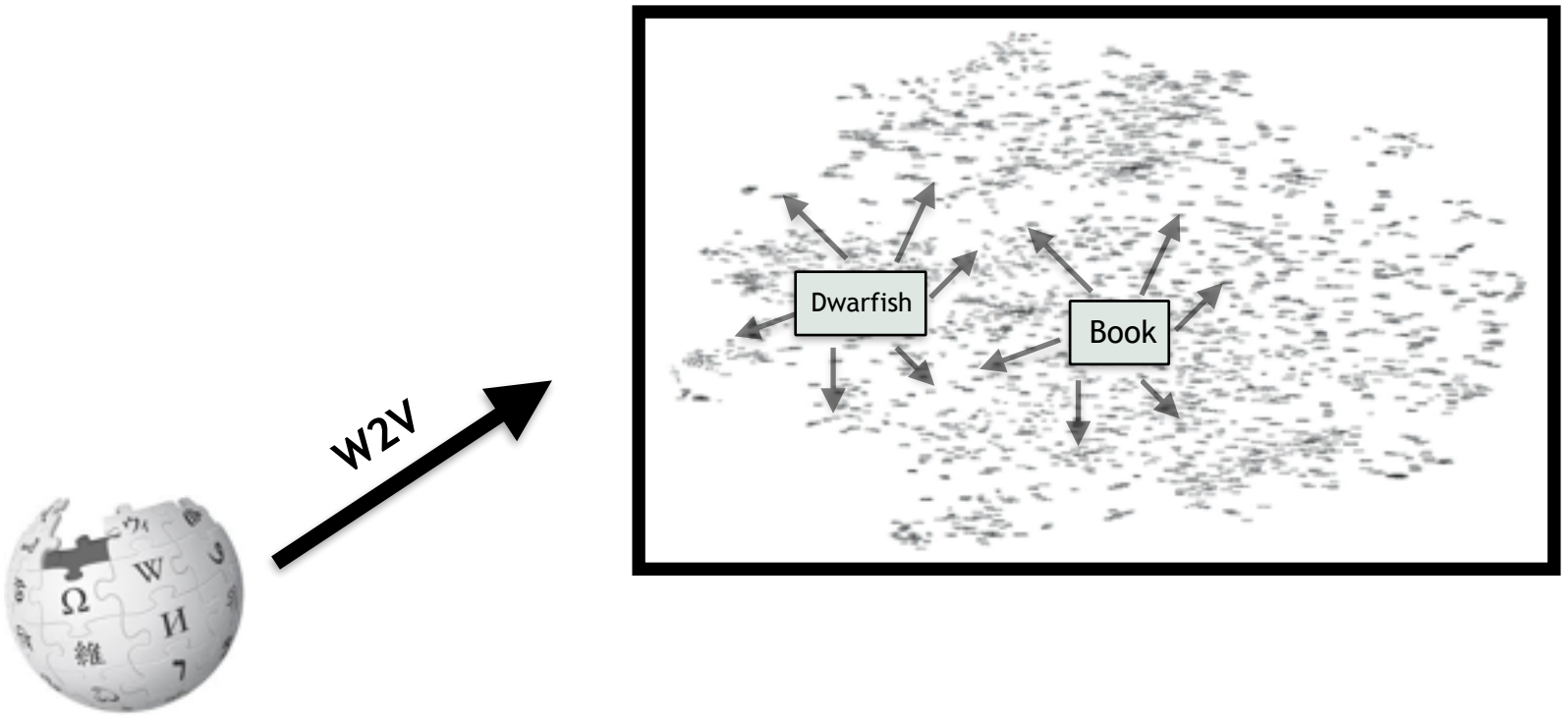
W2V: Word2Vec SkipGram with 200 Dimensions

Observations



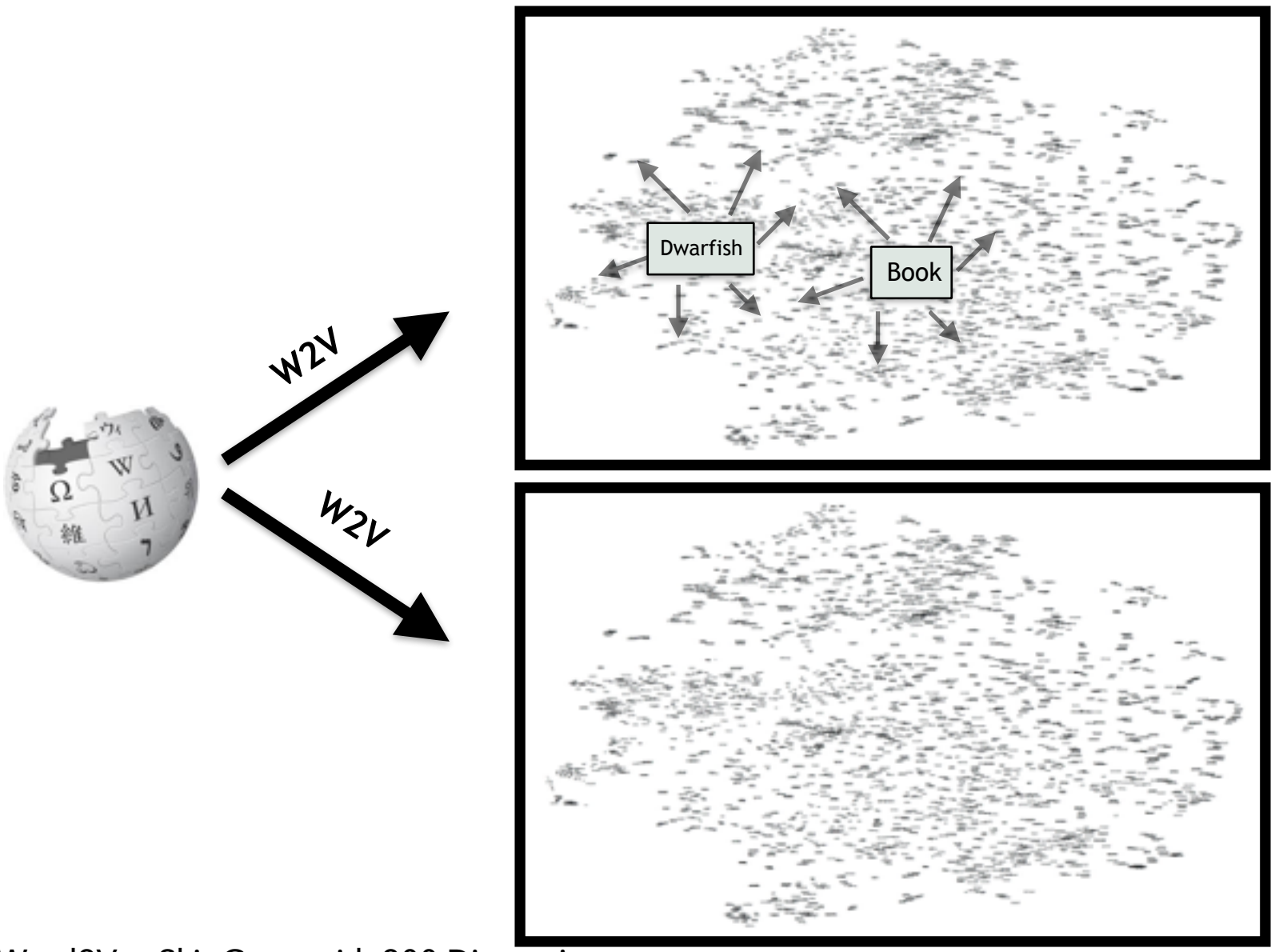
W2V: Word2Vec SkipGram with 200 Dimensions

Observations



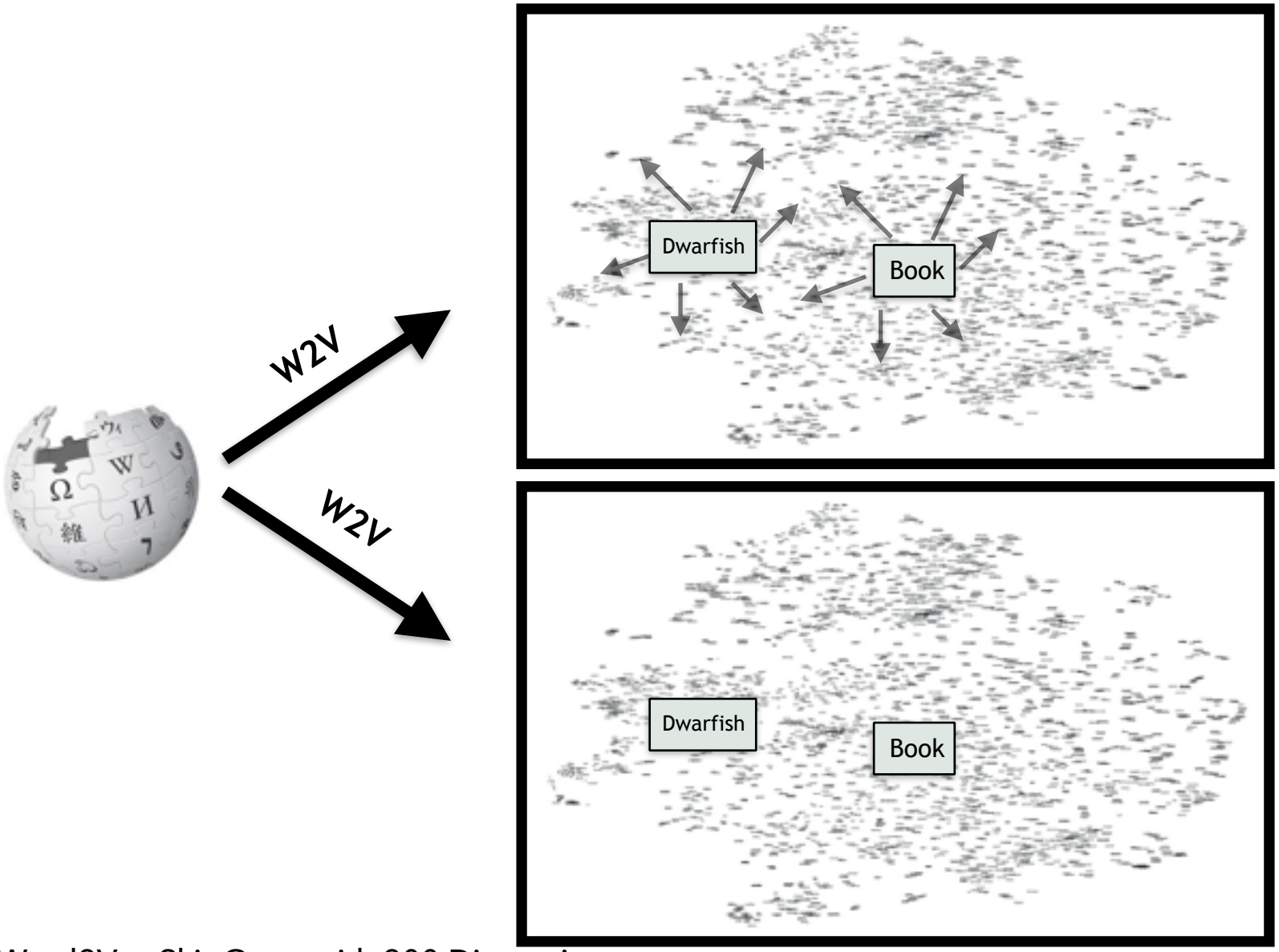
W2V: Word2Vec SkipGram with 200 Dimensions

Observations



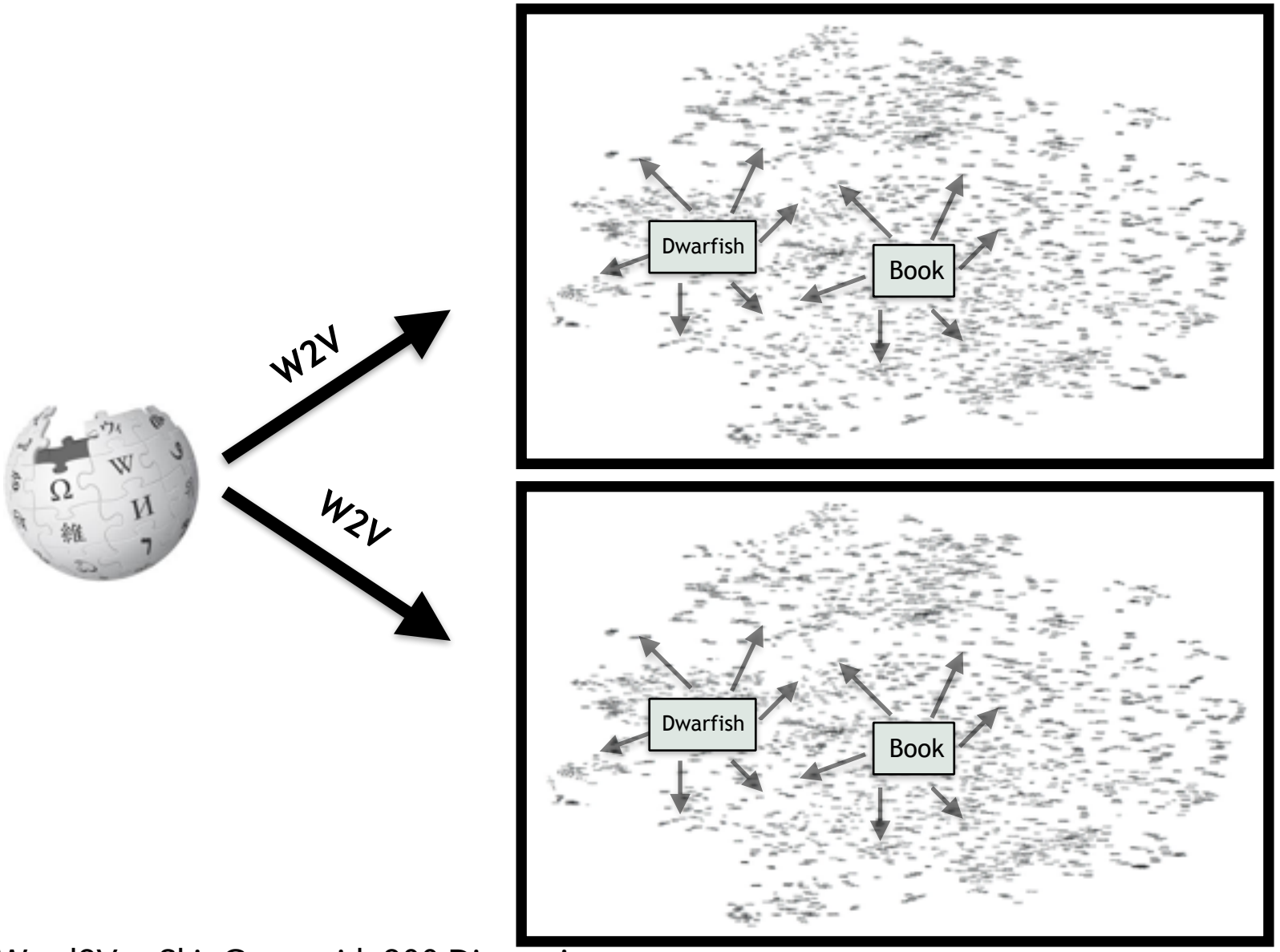
W2V: Word2Vec SkipGram with 200 Dimensions

Observations



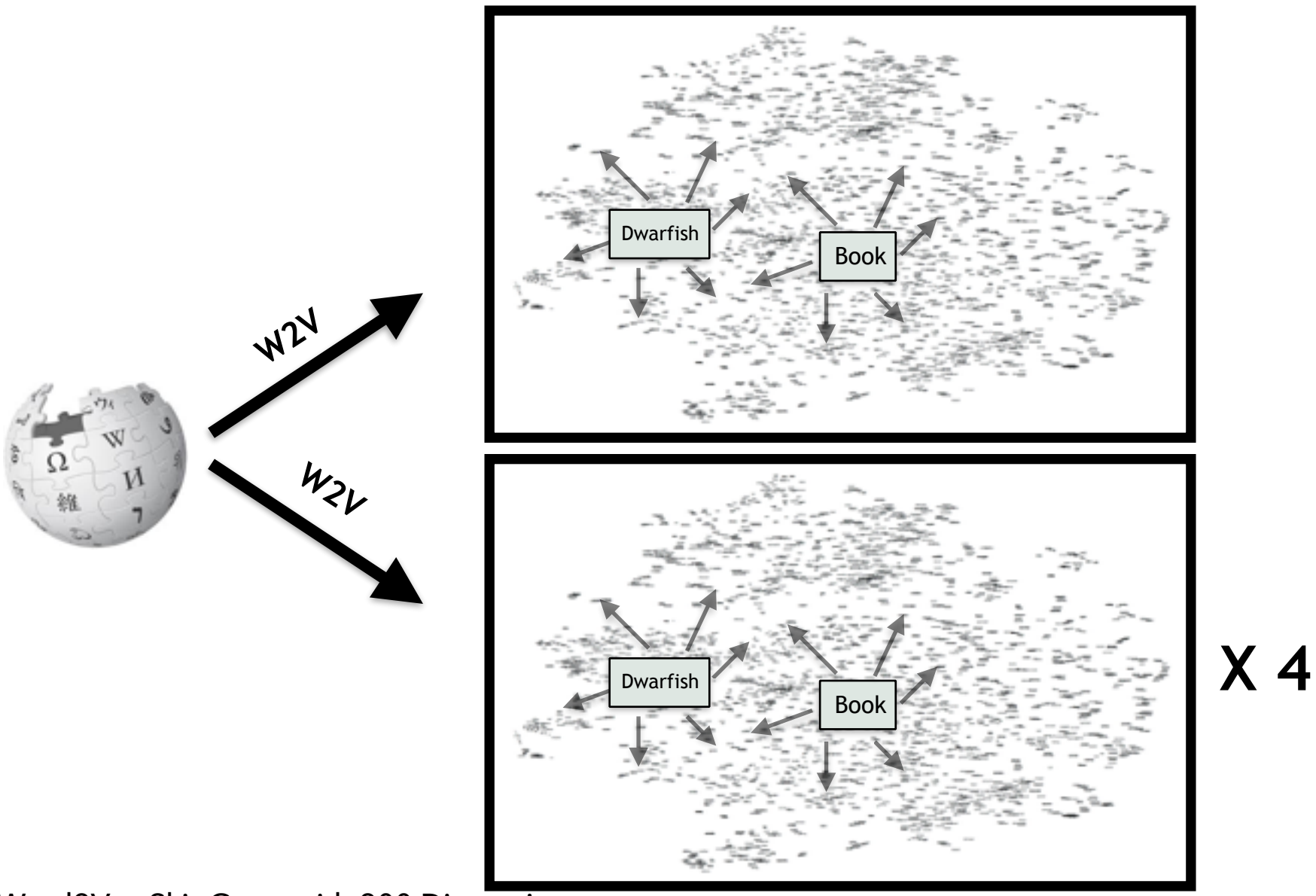
W2V: Word2Vec SkipGram with 200 Dimensions

Observations



W2V: Word2Vec SkipGram with 200 Dimensions

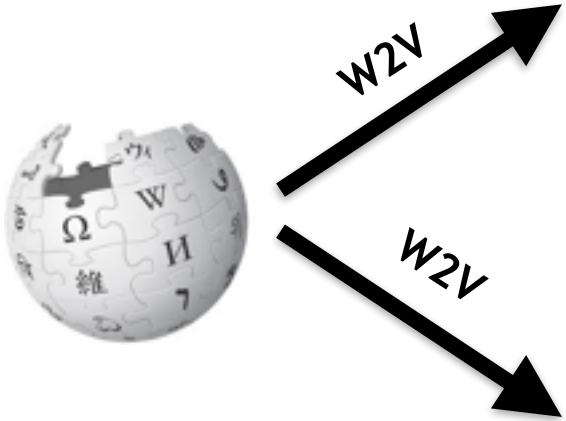
Observations

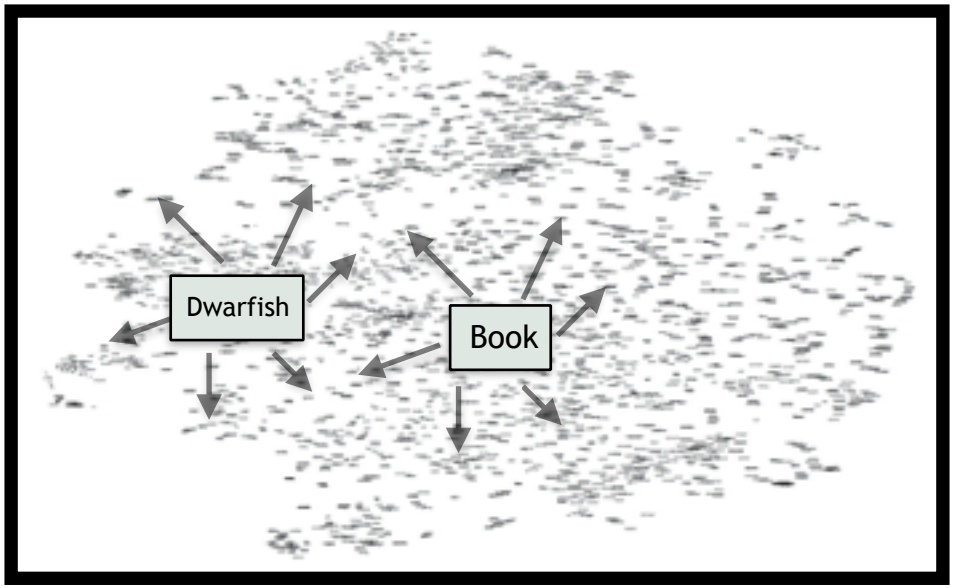


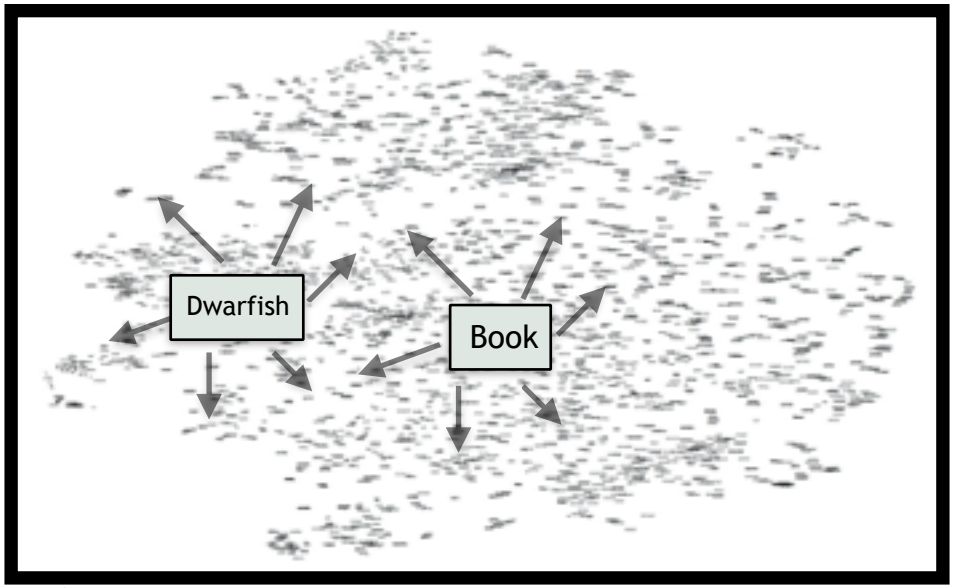
W2V: Word2Vec SkipGram with 200 Dimensions

Observations

Similarity values:
Base





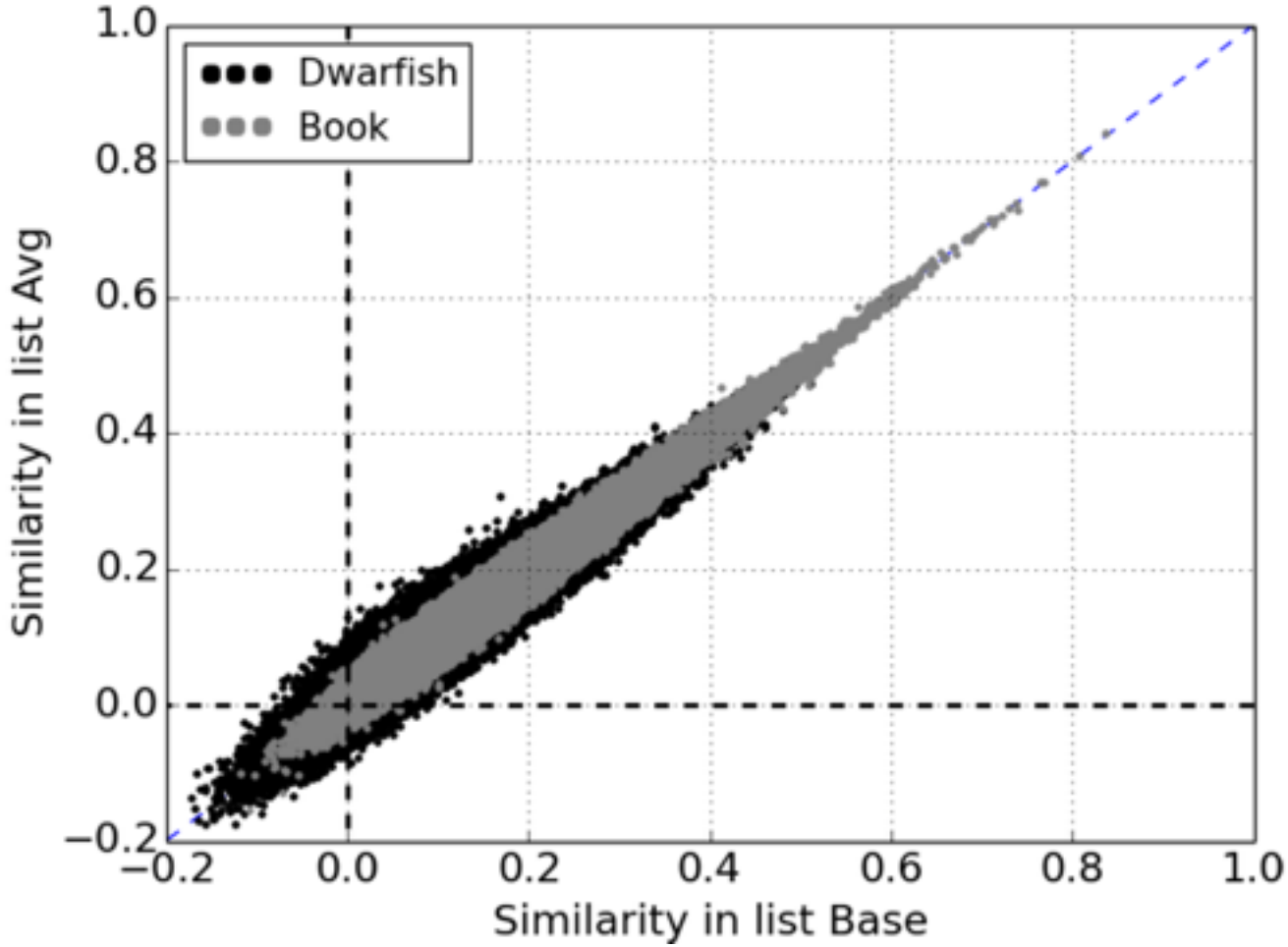


Average of similarity values of 5 models:
Avg

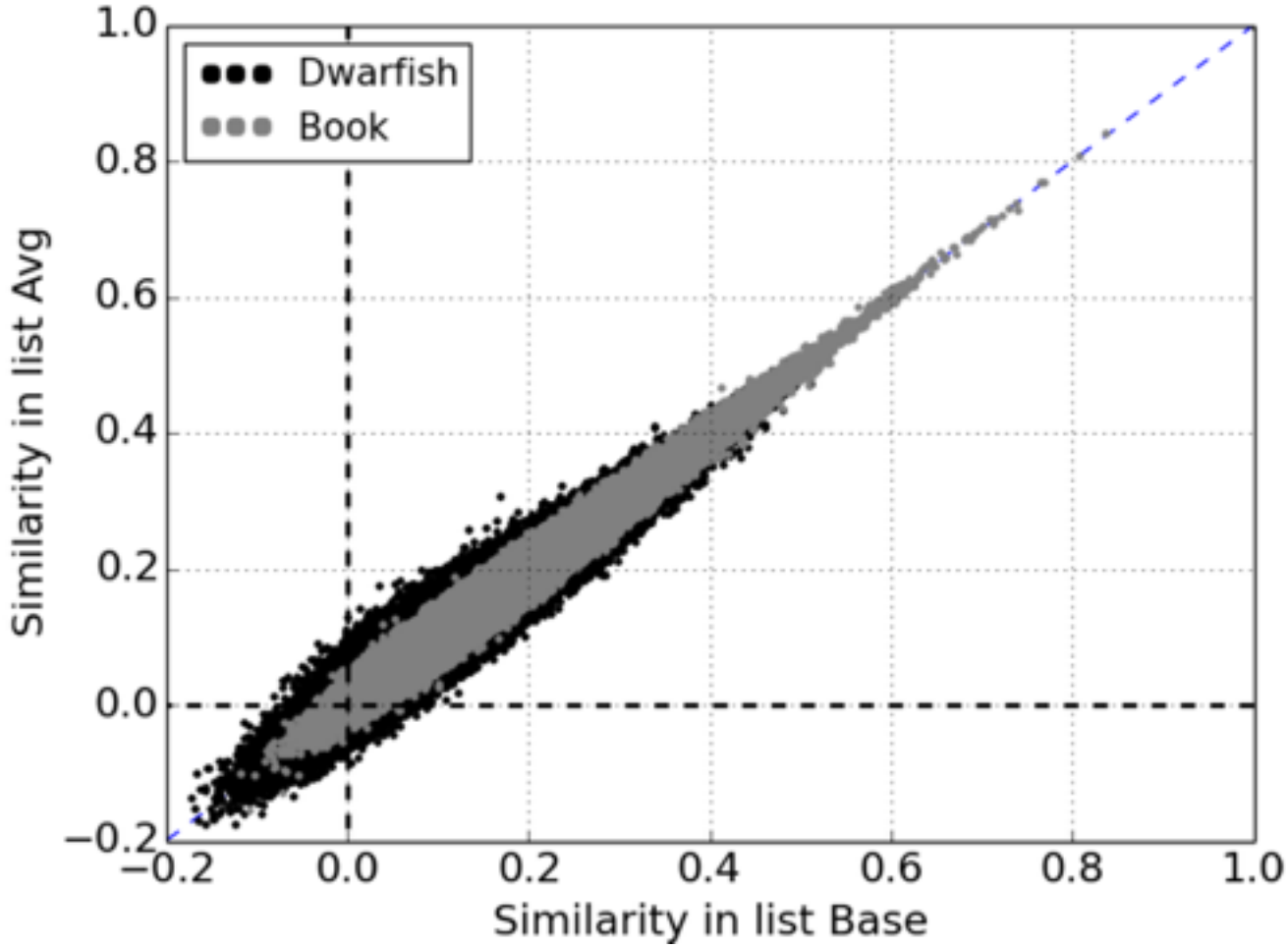
X 4

W2V: Word2Vec SkipGram with 200 Dimensions

Uncertainty



Uncertainty



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

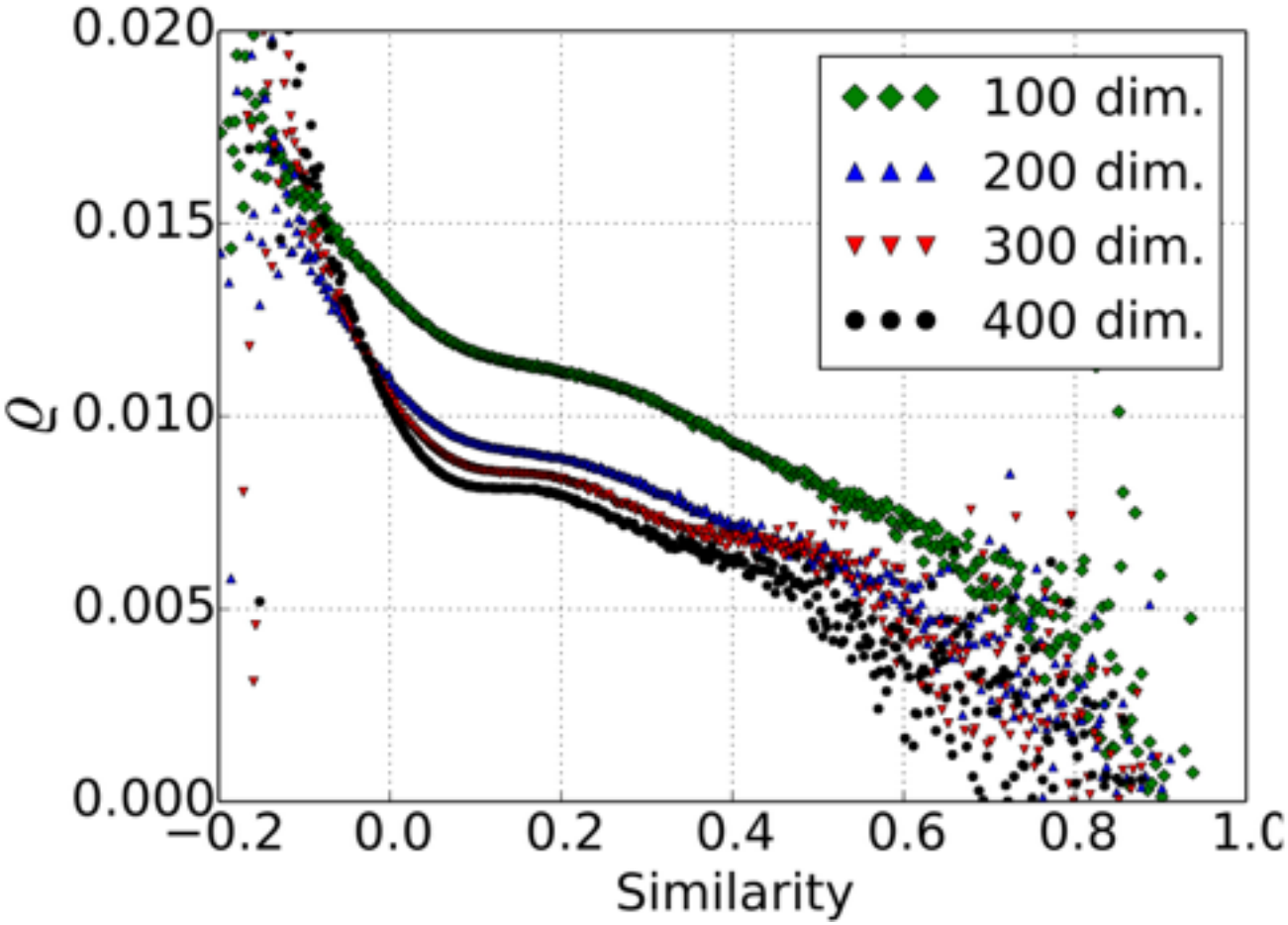
$$\mathcal{S}_s = \{(x,y) : sim(\vec{x}_M, \vec{y}_M) \in (s, s+\epsilon)\}$$

Uncertainty

- **Arbitrary term:** Average of 100 representative terms [Schnabel et al. 2015]

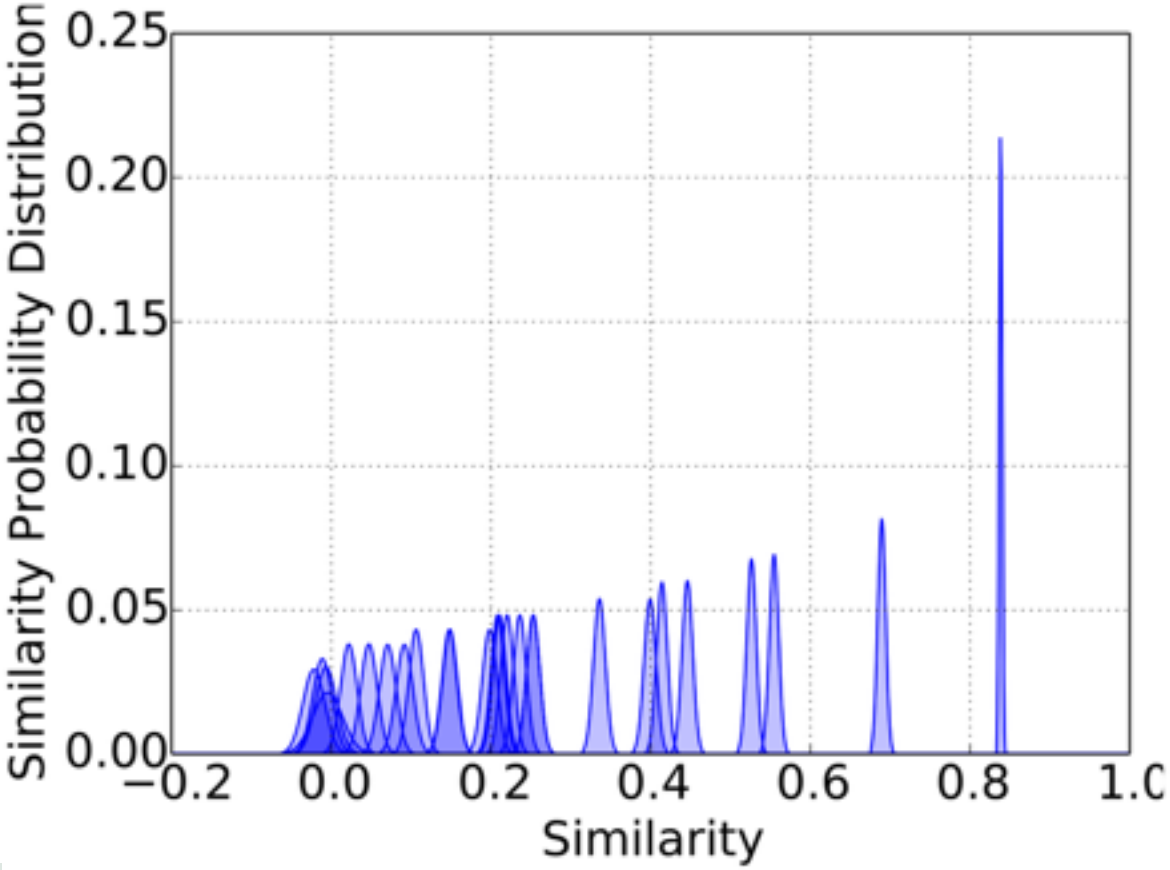
Uncertainty

- **Arbitrary term:** Average of 100 representative terms [Schnabel et al. 2015]
- Less uncertainty in higher similarity values
- Less uncertainty in higher dimensions



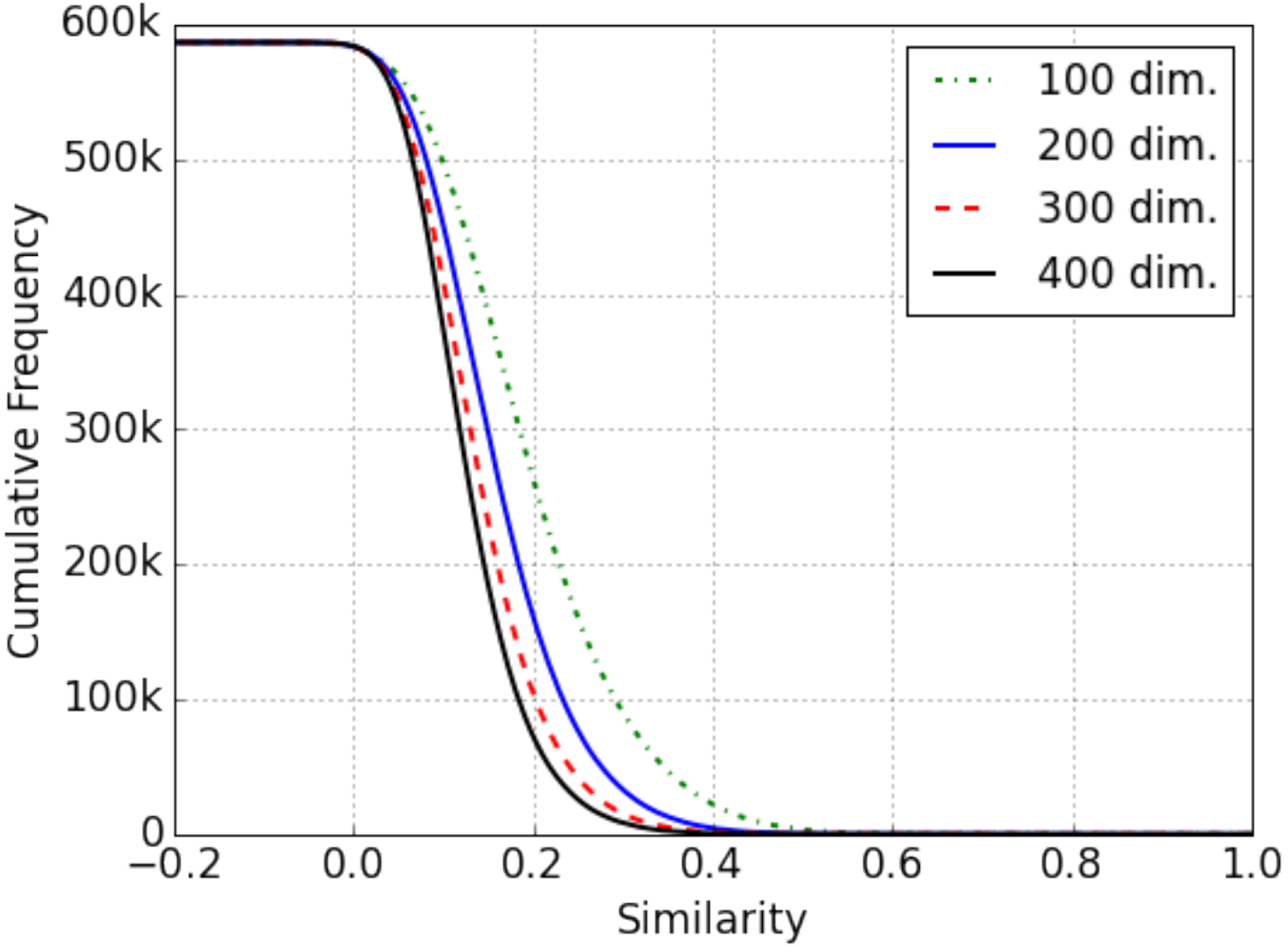
Similarity Probability Distribution

- Similarity between terms as probability distribution instead of concrete values
- Assumption: normal distribution based on 5 observed similarities



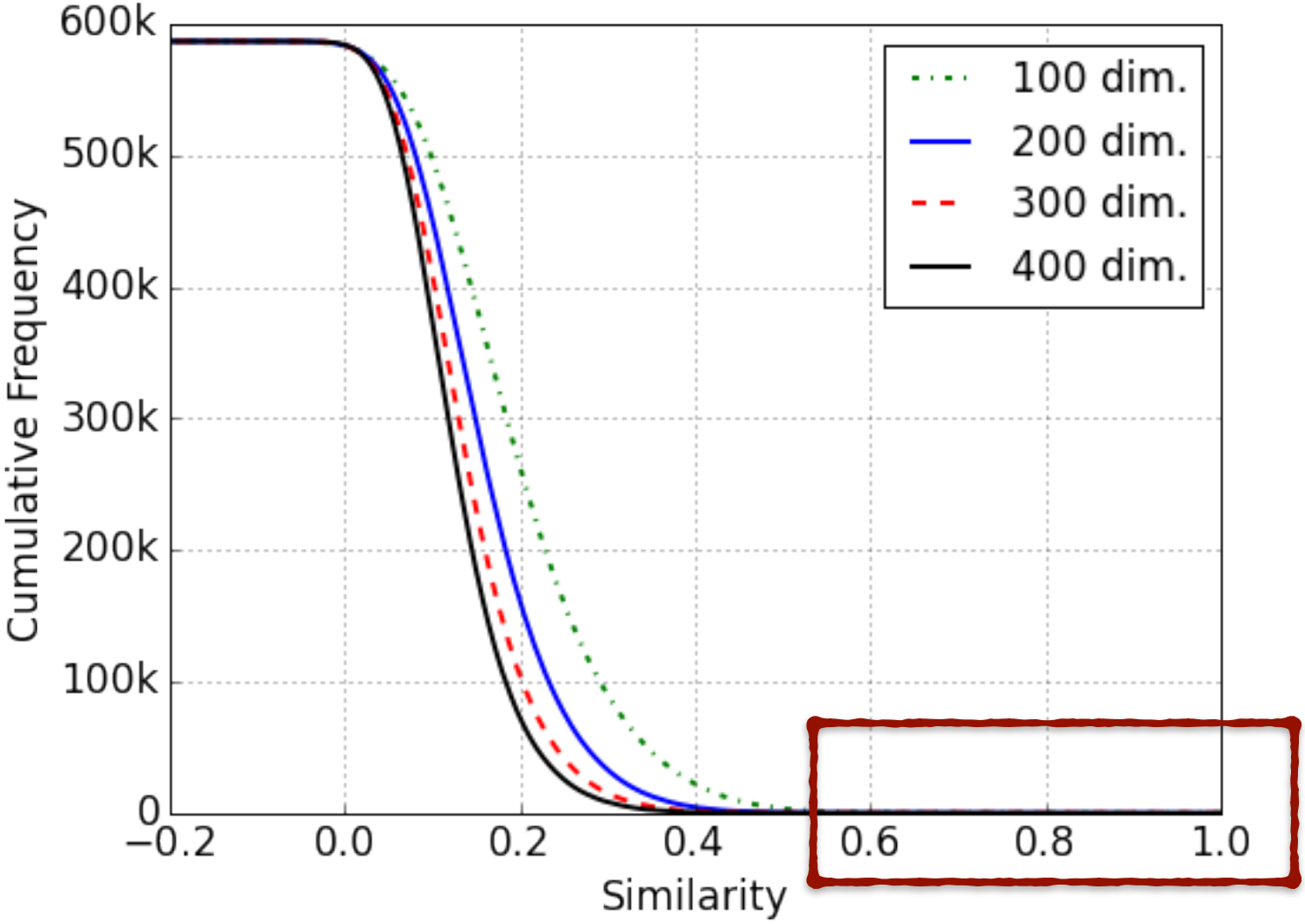
Mixture of Cumulative Similarity Distributions

- Y axes: number of neighbors, located in the space between the X value and the term



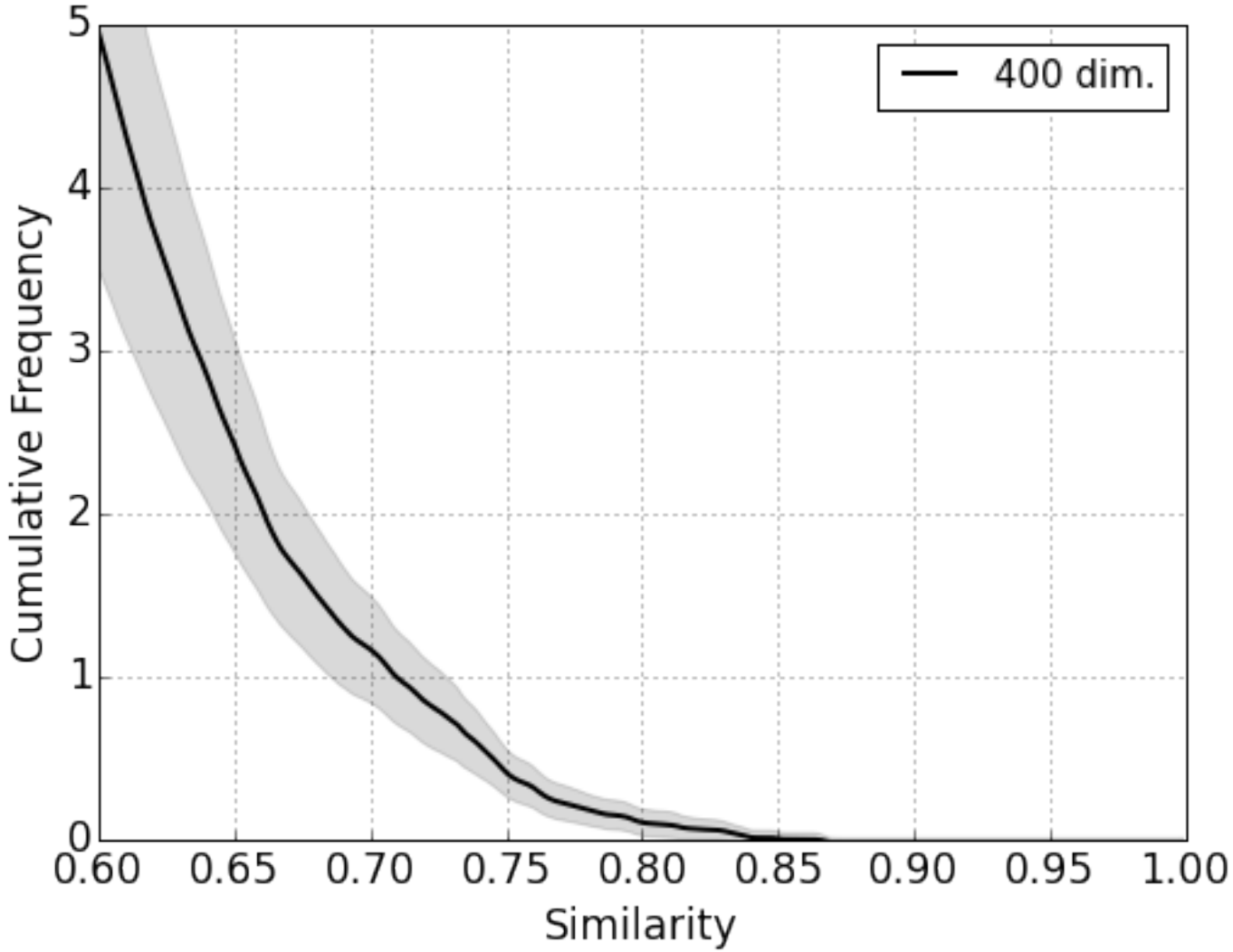
Mixture of Cumulative Similarity Distributions

- Y axes: number of neighbors, located in the space between the X value and the term



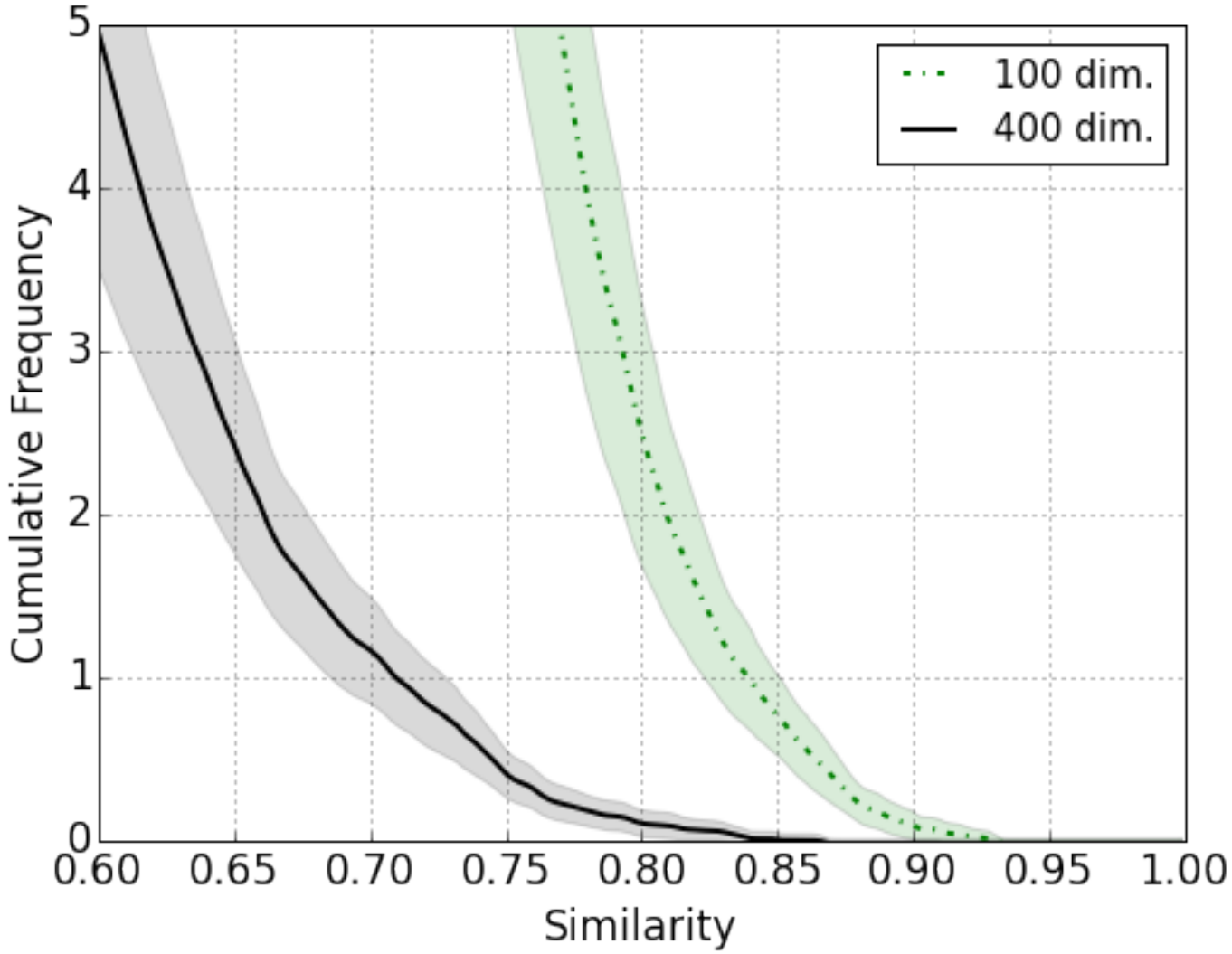
Mixture of Cumulative Similarity Distributions

- Y axes: number of neighbors, located in the space between the X value and the term



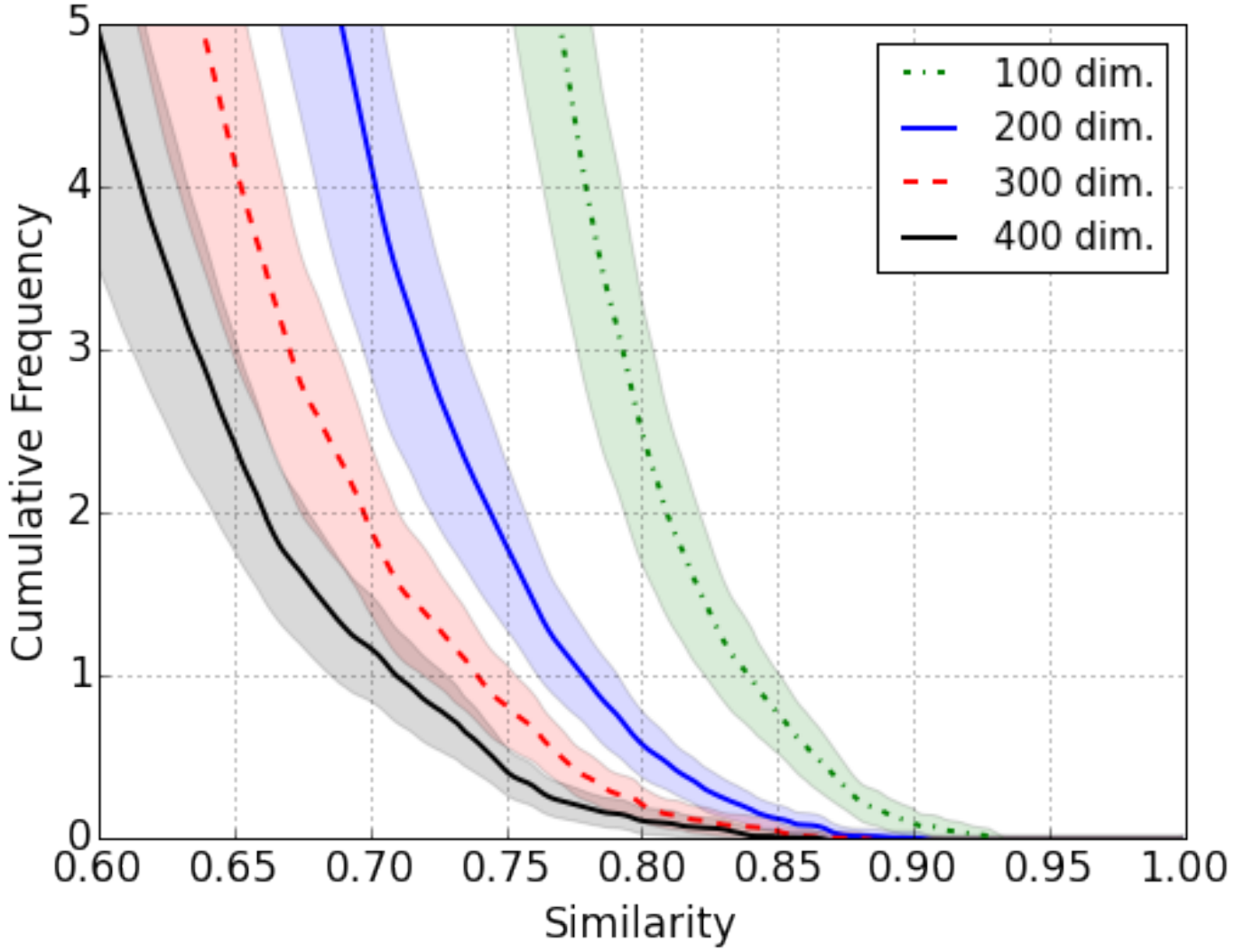
Mixture of Cumulative Similarity Distributions

- Y axes: number of neighbors, located in the space between the X value and the term



Mixture of Cumulative Similarity Distributions

- Y axes: number of neighbors, located in the space between the X value and the term



Filtering Neighbors

What is the best threshold for filtering the related terms?

Filtering Neighbors

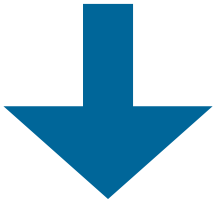
What is the best threshold for filtering the related terms?

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

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

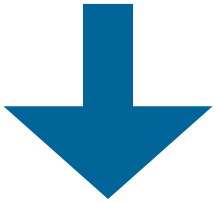


What is the expected number of synonyms for a word in English?

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?

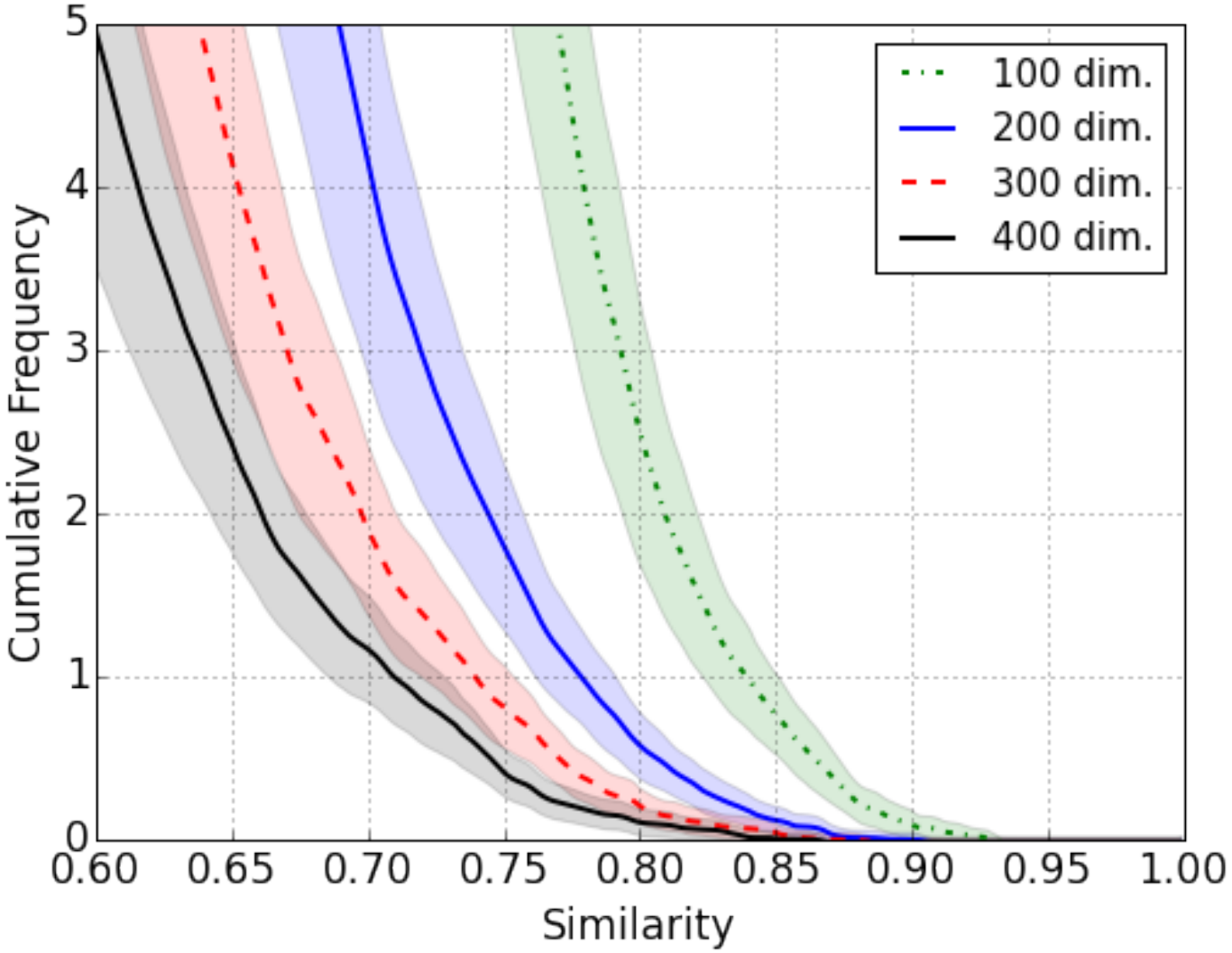
# of terms: 147306

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

Standard deviation : 3.1

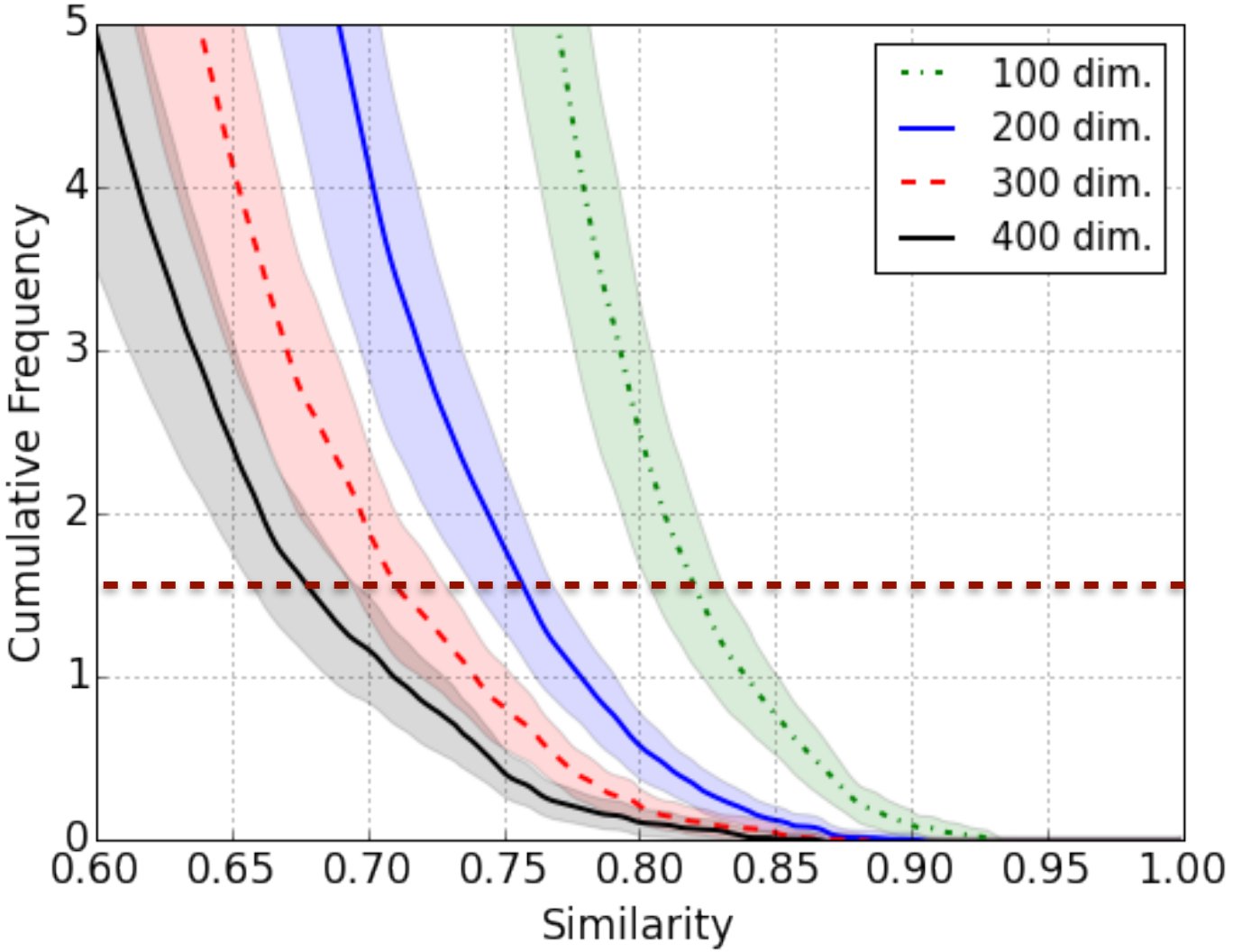
Threshold

- Proposed Threshold: cumulative frequency equal to 1.6



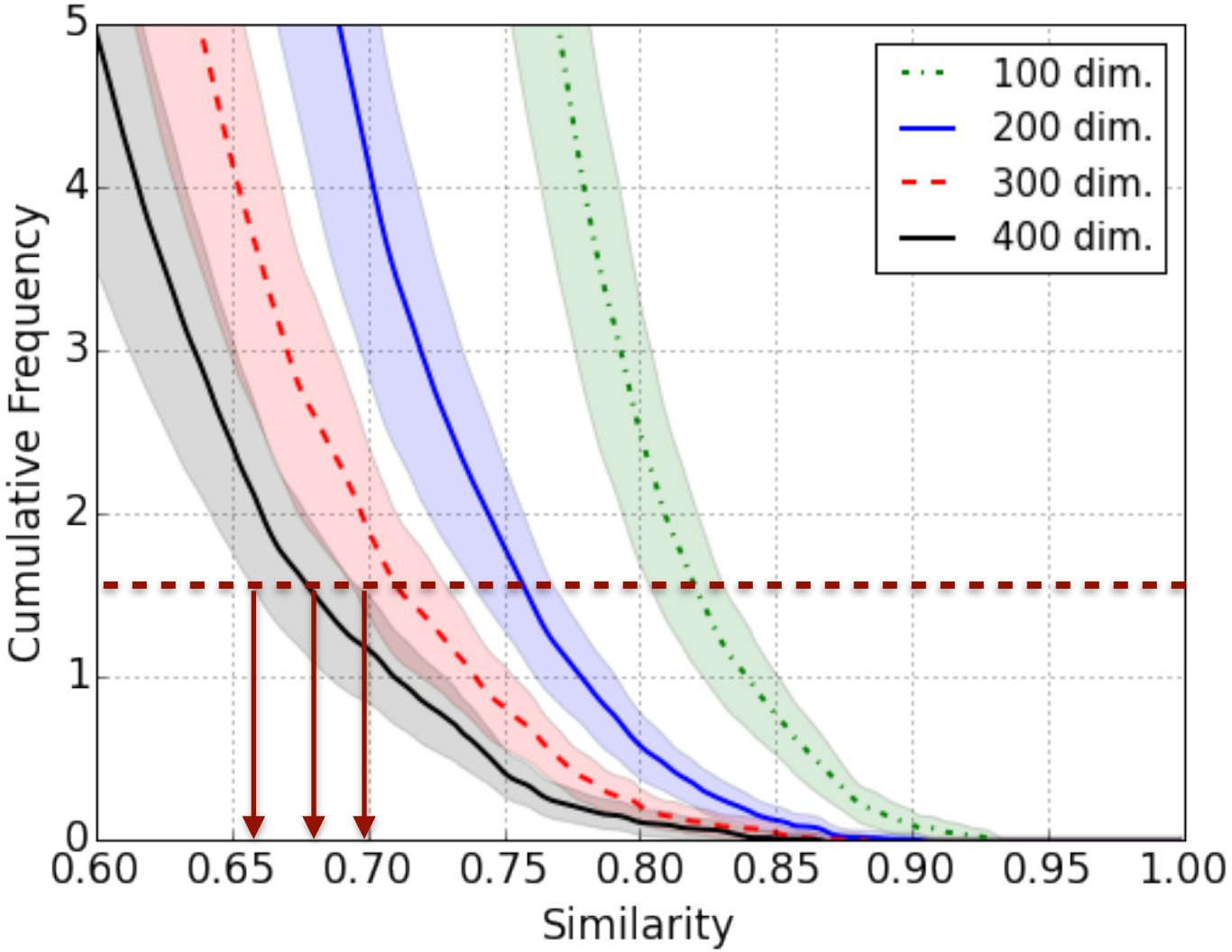
Threshold

- Proposed Threshold: cumulative frequency equal to 1.6



Threshold

- Proposed Threshold: cumulative frequency equal to 1.6

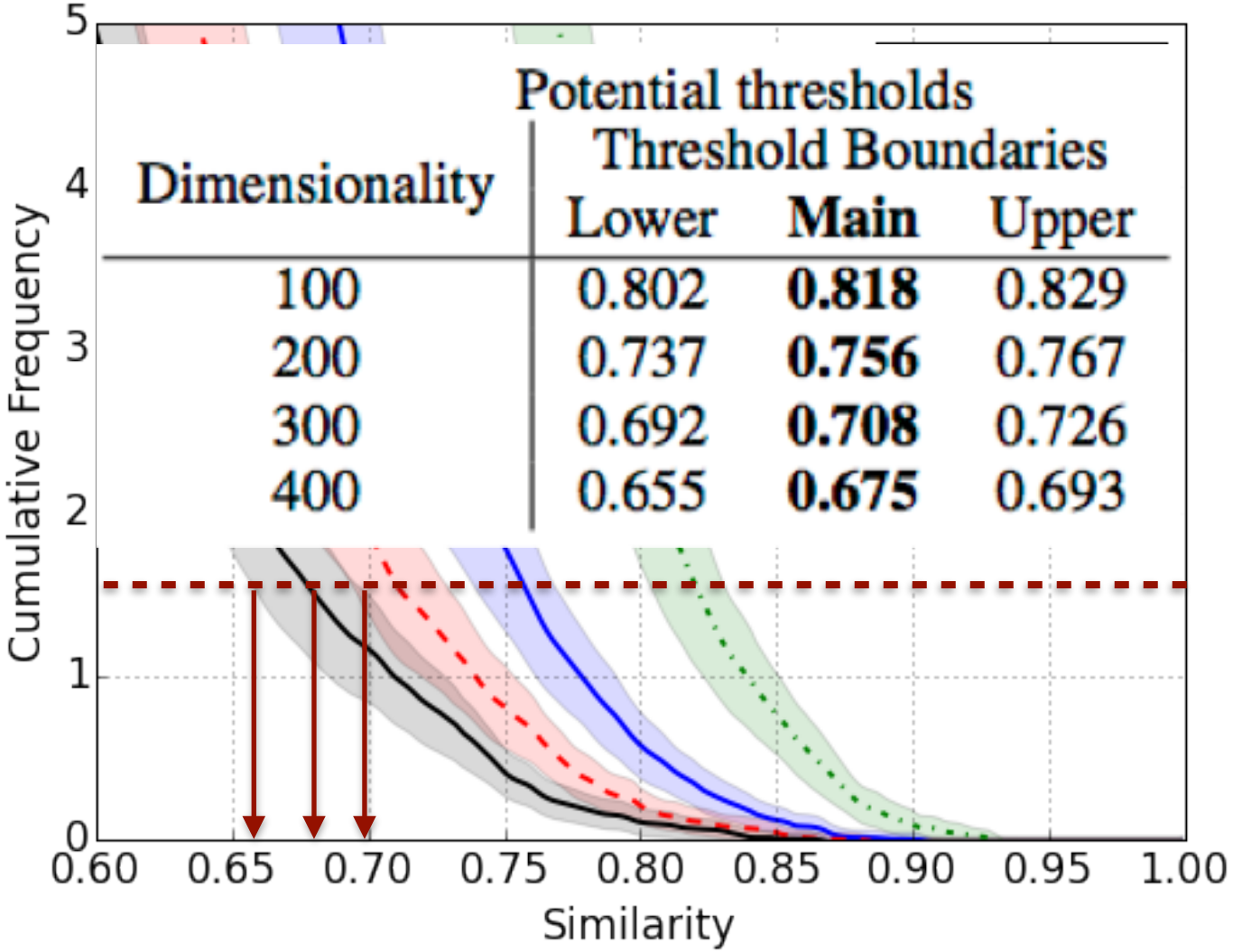


Threshold

- Proposed Threshold: cumulative frequency equal to 1.6

Potential thresholds

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 Setup

- Translation Language Model with Dirichlet smoothing
- Use word embedding similarity value for translation probability
  Zuccon et al. [2015]
- Apply the proposed threshold to select the similar words
- Brute-force search to find the optimal threshold

$$P(q|M_d) = \prod_{t_q \in q} \left(\sum_{t_d \in d} P_T(t_q|t_d) P(t_d|M_d) \right)$$

Experiments Setup

- Translation Language Model with Dirichlet smoothing
- Use word embedding similarity value for translation probability
  Zuccon et al. [2015]
- Apply the proposed threshold to select the similar words
- Brute-force search to find the optimal threshold

$$P(q|M_d) = \prod_{t_q \in q} \left(\sum_{t_d \in d} P_T(t_q|t_d) P(t_d|M_d) \right)$$

- Test collections:

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

Experiments Setup

- Translation Language Model with Dirichlet smoothing
- Use word embedding similarity value for translation probability
  Zuccon et al. [2015]
- Apply the proposed threshold to select the similar words
- Brute-force search to find the optimal threshold

$$P(q|M_d) = \prod_{t_q \in q} \left(\sum_{t_d \in d} P_T(t_q|t_d) P(t_d|M_d) \right)$$

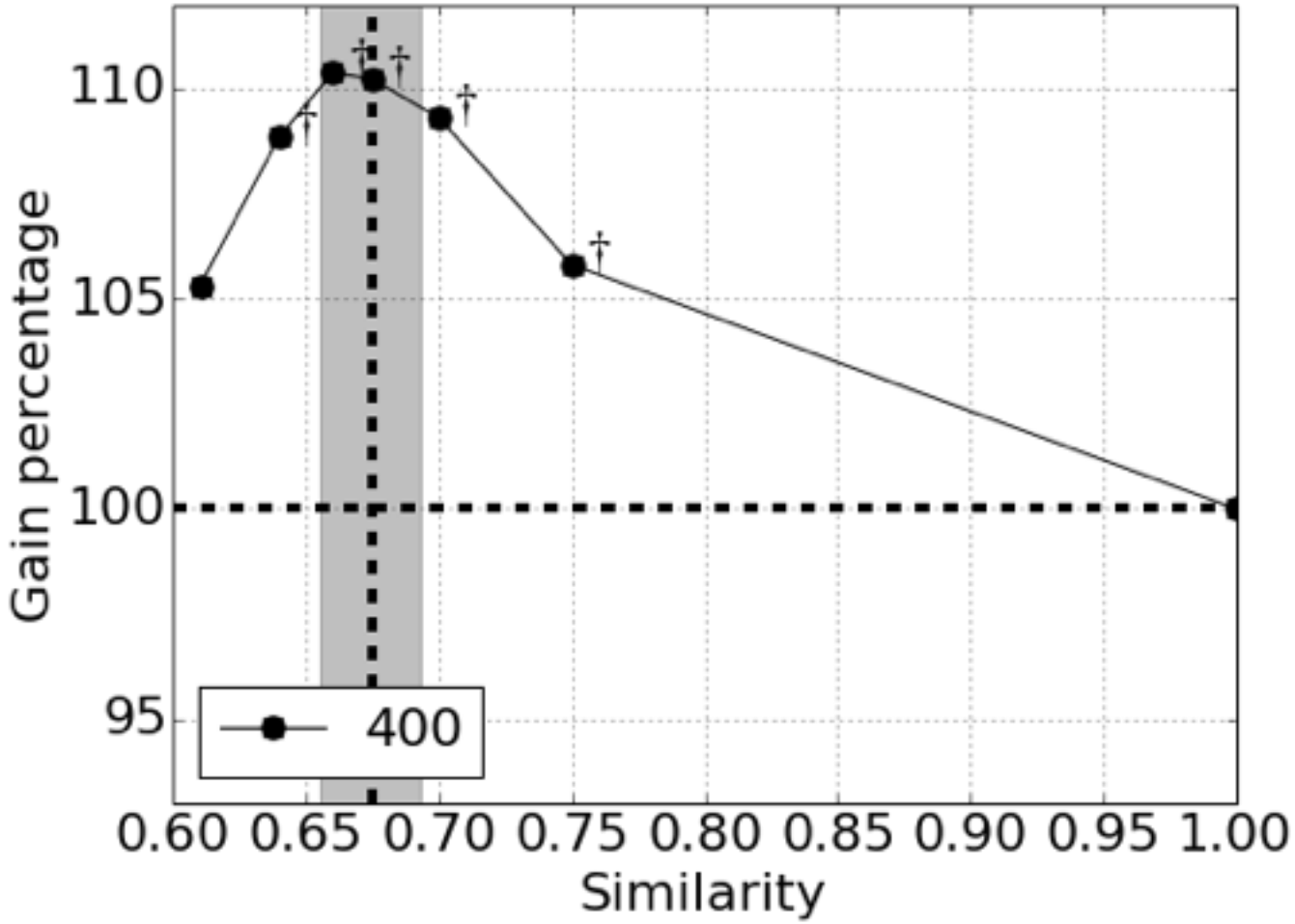
- Test collections:

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

- Baseline: language model (Dirichlet smoothing)
- Significance Test : T-Test $p<0.05$

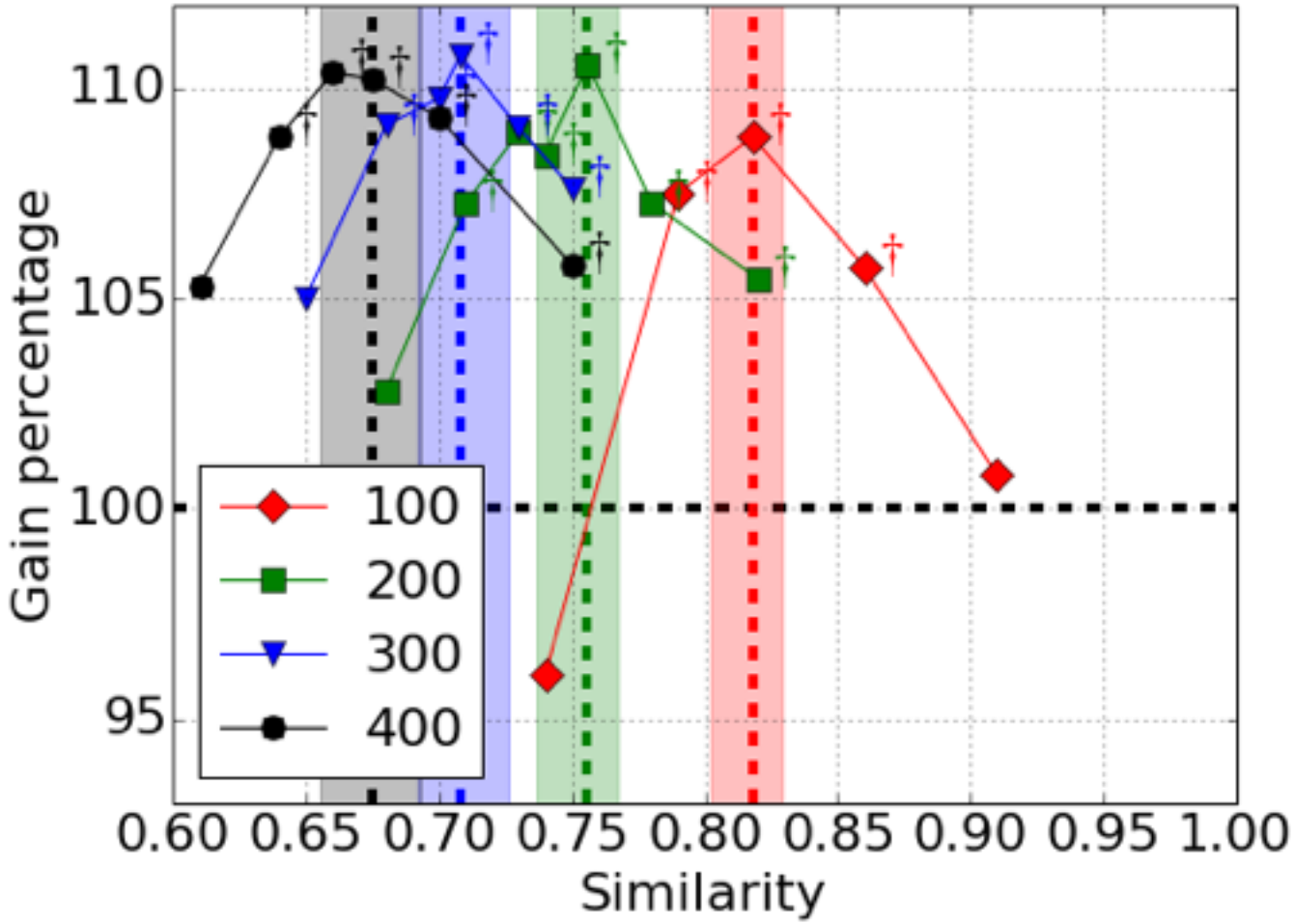
Experiments Results

- Gain of MAP over baseline, averaged on four collections.
- Conclusion1: Optimal threshold is either the same or in the confidence interval of the proposed threshold.



Experiments Results

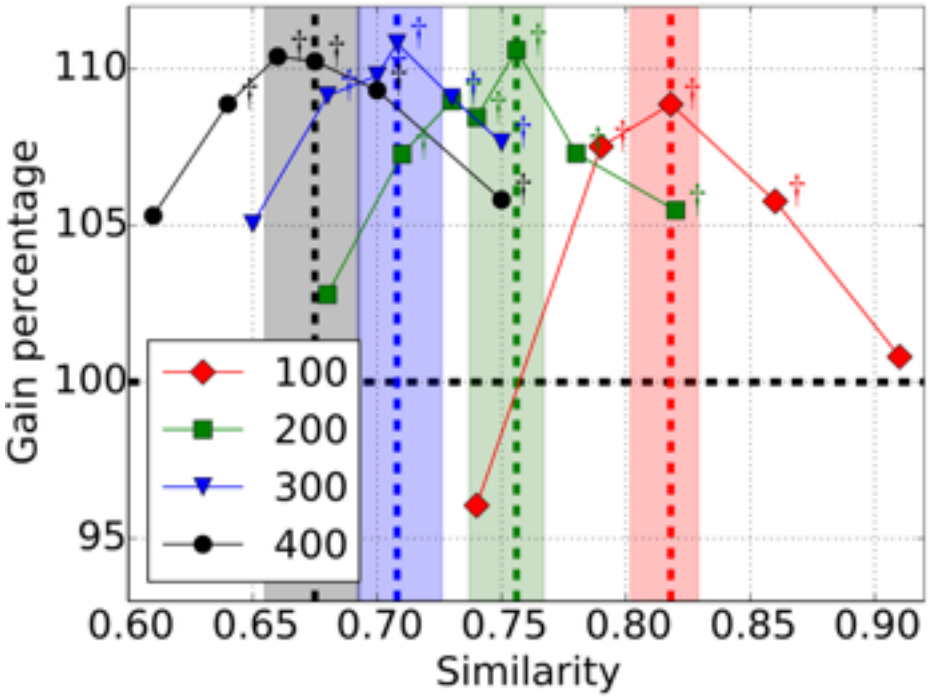
- Gain of MAP over baseline, averaged on four collections.
- Conclusion1: Optimal threshold is either the same or in the confidence interval of the proposed threshold.



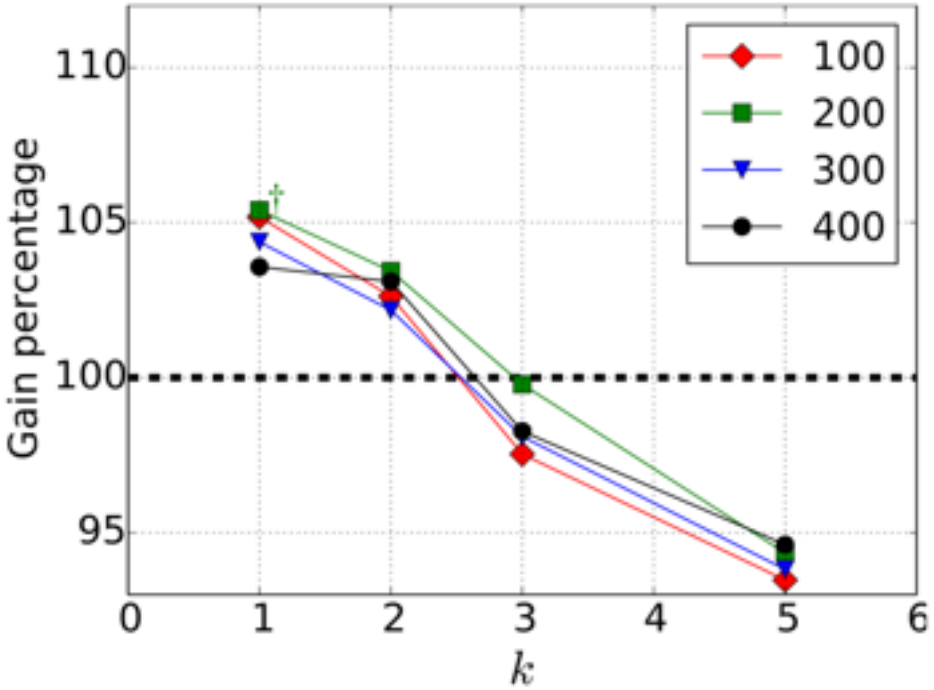
Threshold vs. TopN

- Conclusion2: Threshold outperforms TopN

Threshold-based



TopN



Take Home Message

- Uncertainty in neural network word embeddings:
  - depends on similarity value
  - depends on dimensionality

- Threshold to filter *unrelated* terms :
  - Proposed threshold as good as optimal threshold
  - Threshold approach much better than Top-N approach

Questions?

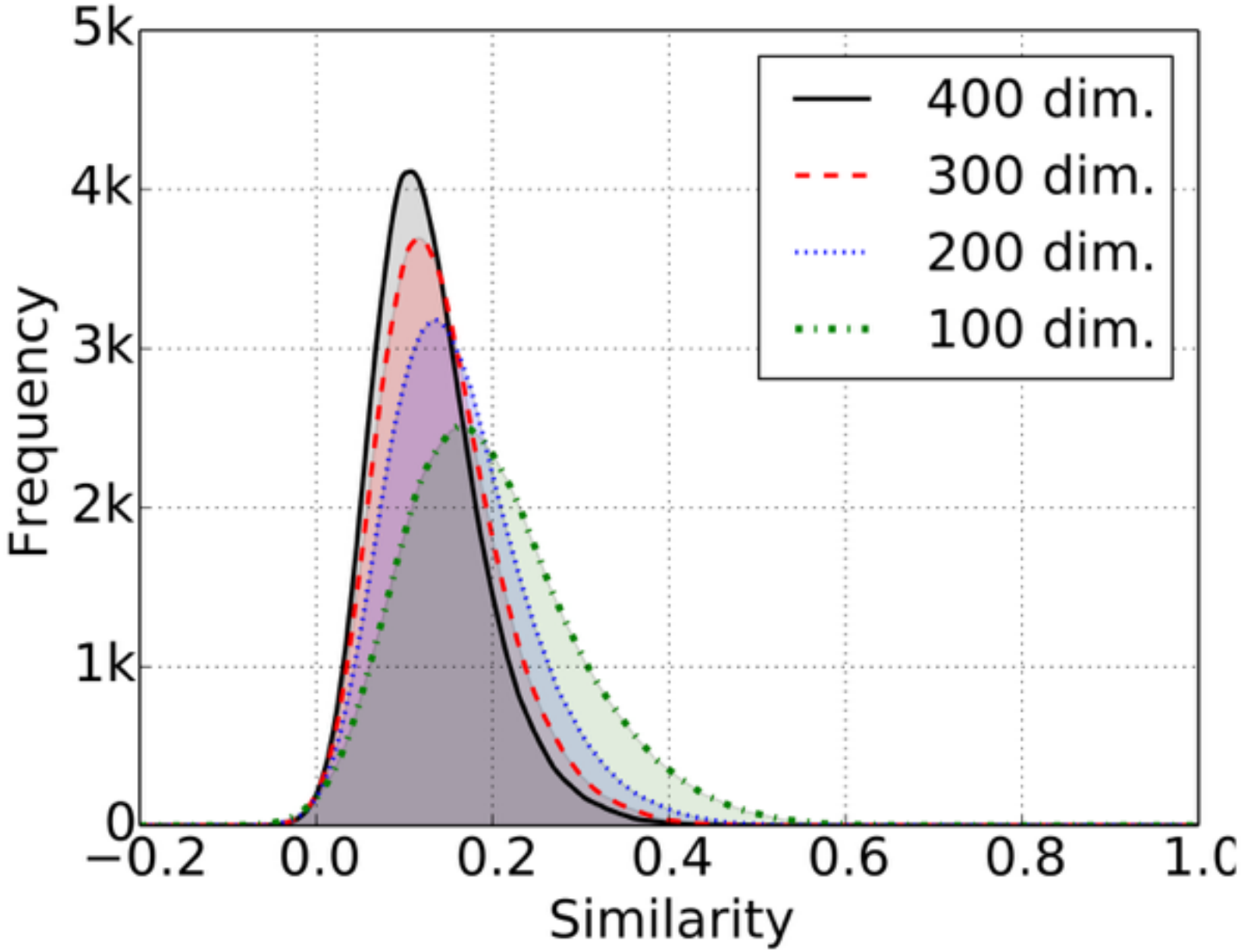
Ideas!

Follow-up paper in CIKM 2016:
Generalizing Translation Models in the Probabilistic Relevance Framework
Navid Rekabsaz, Mihai Lupu, Allan Hanbury


@NRekabsaz

rekabsaz@ifs.tuwien.ac.at

Dimensionality



Proposed vs. Optimal Threshold