FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

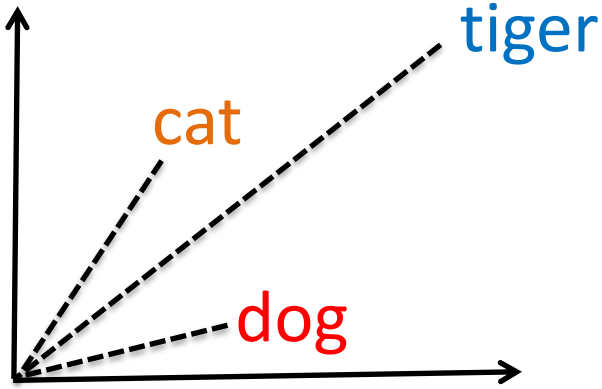
Jena University Language & Information Engineering Lab

The Influence of Down-Sampling Strategies on SVD Word Embedding Stability

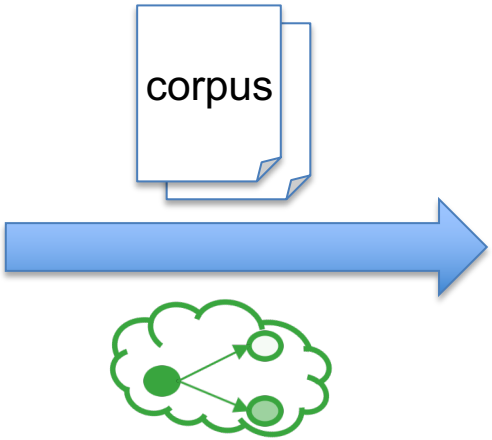
Johannes Hellrich, Bernd Kampe & Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich Schiller University Jena,
Jena, Germany
www.julielab.de

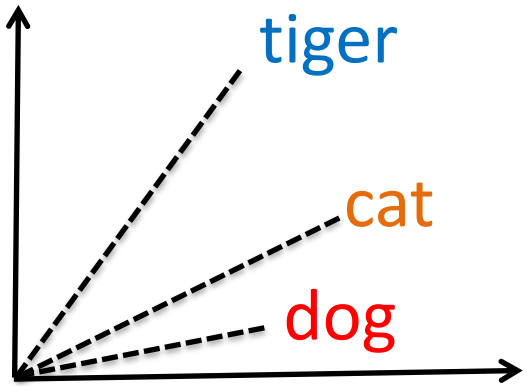
Typical Word Embeddings are Unstable



random embeddings

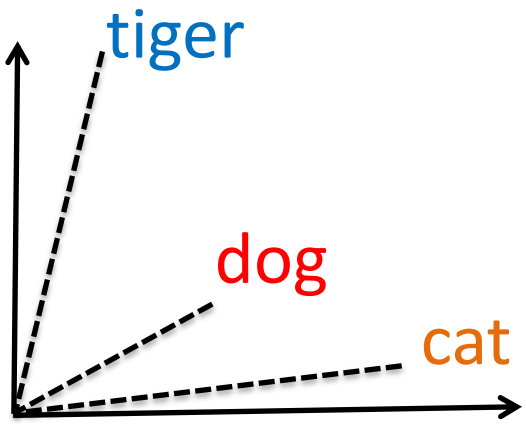


random processing

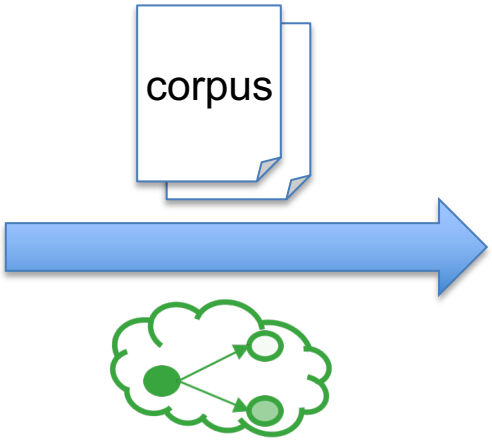


final embeddings

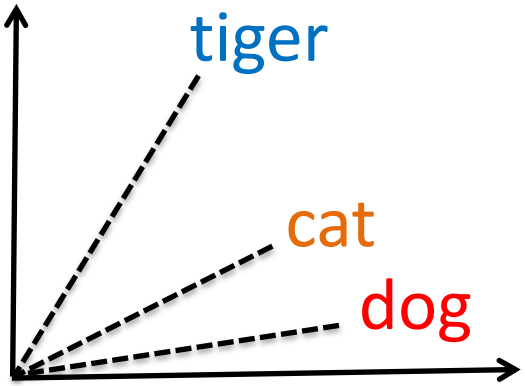
Typical Word Embeddings are Unstable



random embeddings

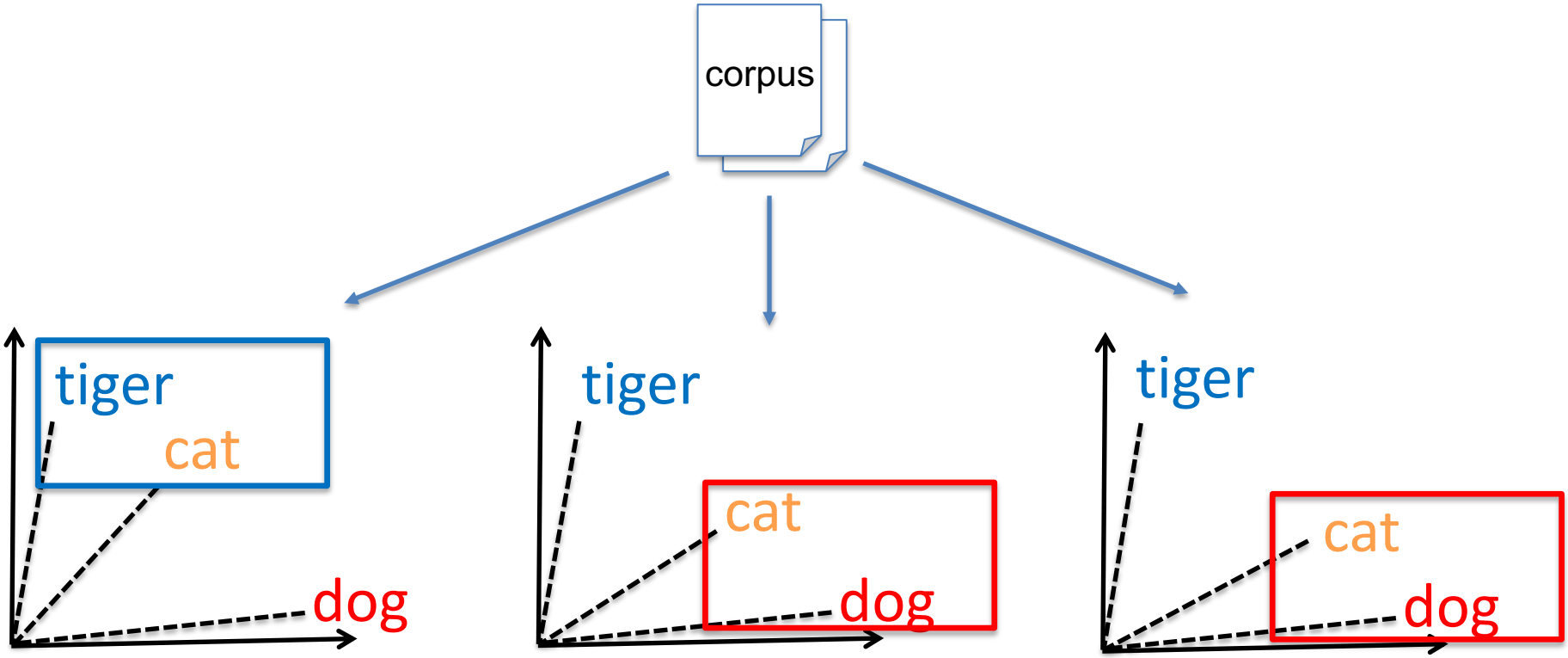


random processing



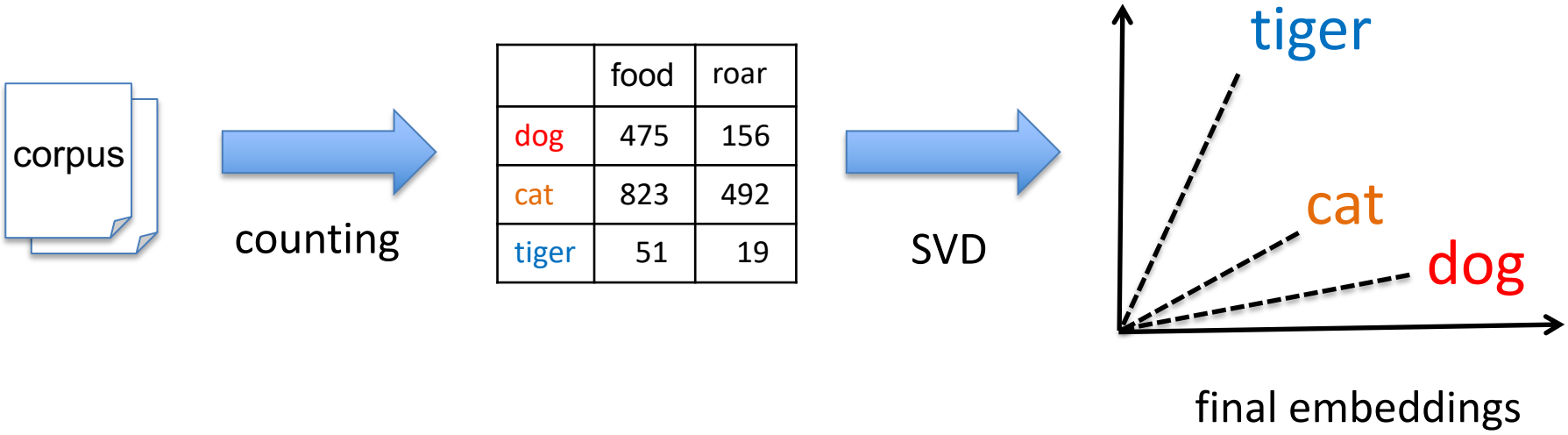
final embeddings

Measuring Stability

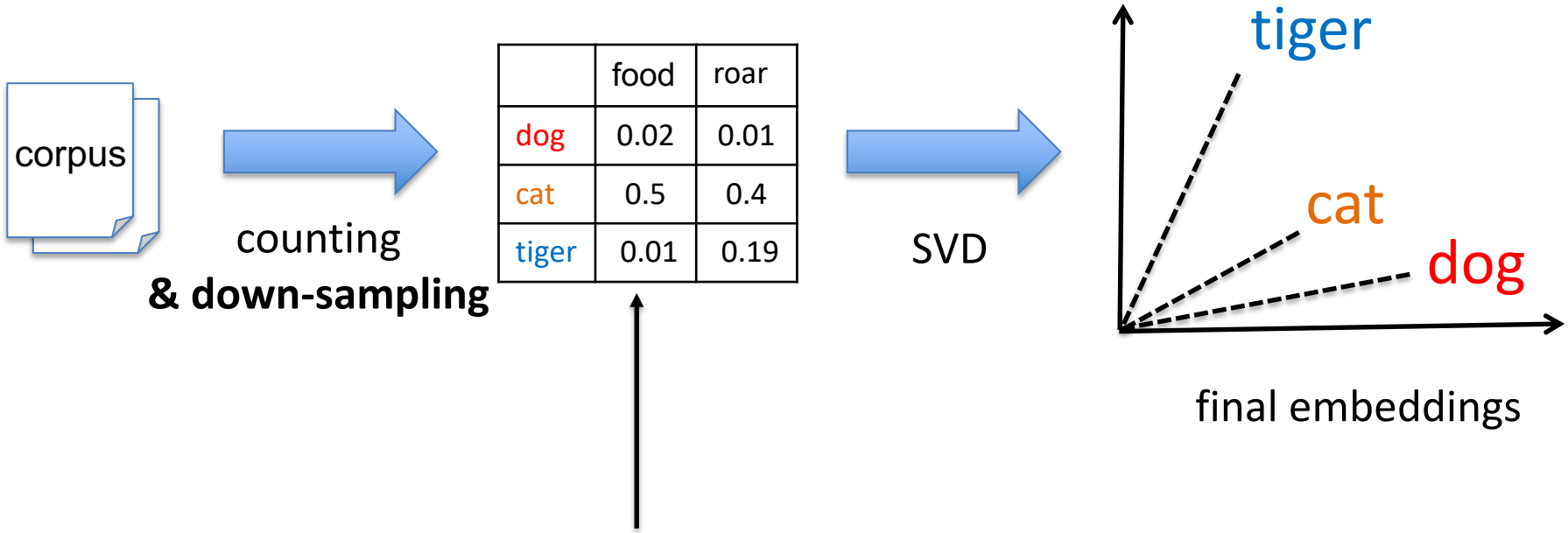


$$j@n := \frac{1}{|A|} \sum_{a \in A} \frac{|\bigcap_{m \in M} \text{msw}(a, n, m)|}{|\bigcup_{m \in M} \text{msw}(a, n, m)|}$$

Why SVD Embeddings?



Why SVD Embeddings?



Why Down-Sampling?

- Avoids over-representing frequent words
- Closer context words are more salient than distant ones

→ Increased Performance (Mikolov, NIPS 2013)

Down-Sampling Mechanism

Probabilistic

- word2vec
- SVD_{PPMI}

Weighting

- GloVe
- New: SVD_{wPPMI}

Experimental Design I/II

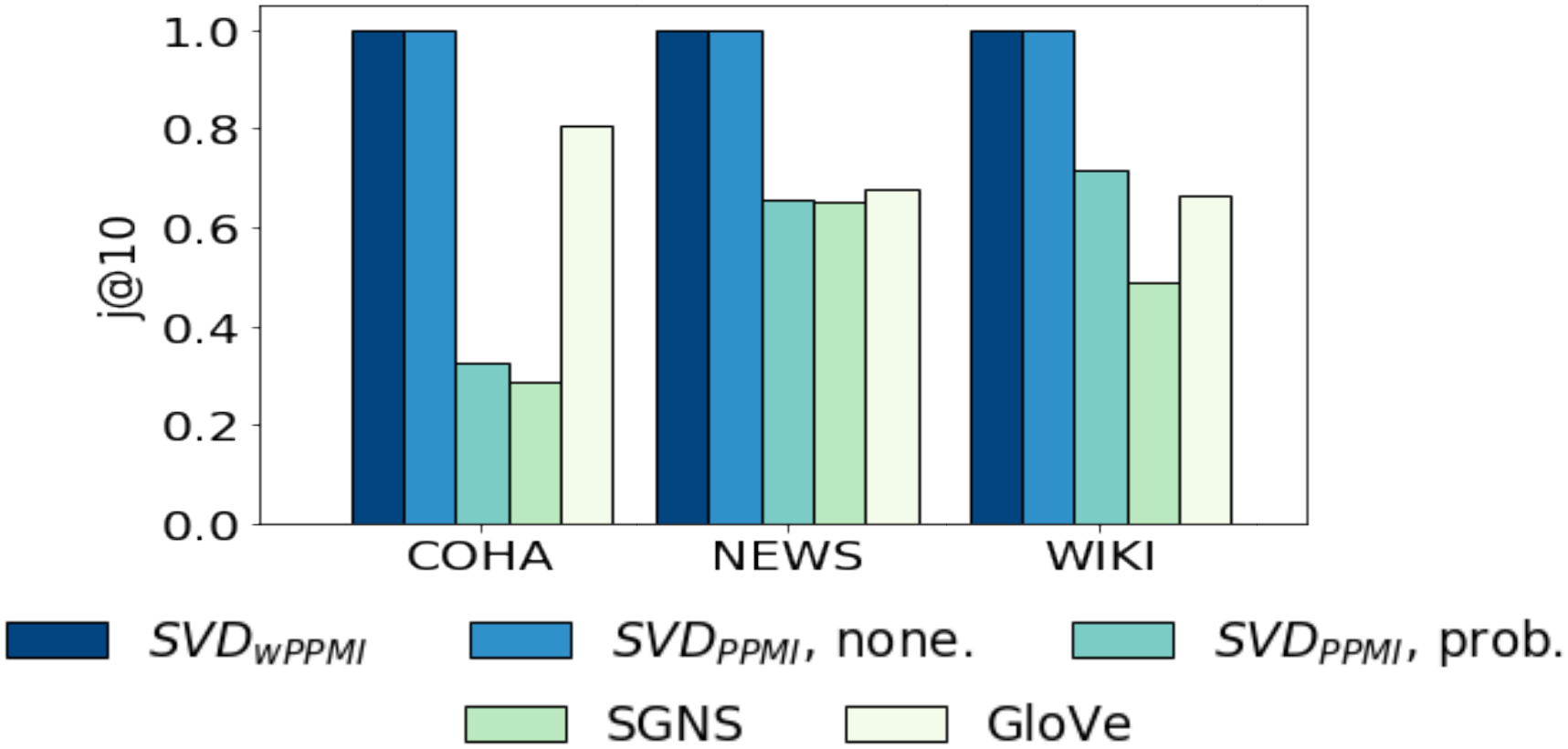
- Three Corpora:
  - Corpus of Historical American English 2000s decade (COHA; 28M tokens.)
  - English News Crawl Corpus (NEWS; 550M tokens)
  - Wikipedia (WIKI; 1.7G tokens)

→ Other studies used mostly COHA-sized corpora!

Experimental Design II/II

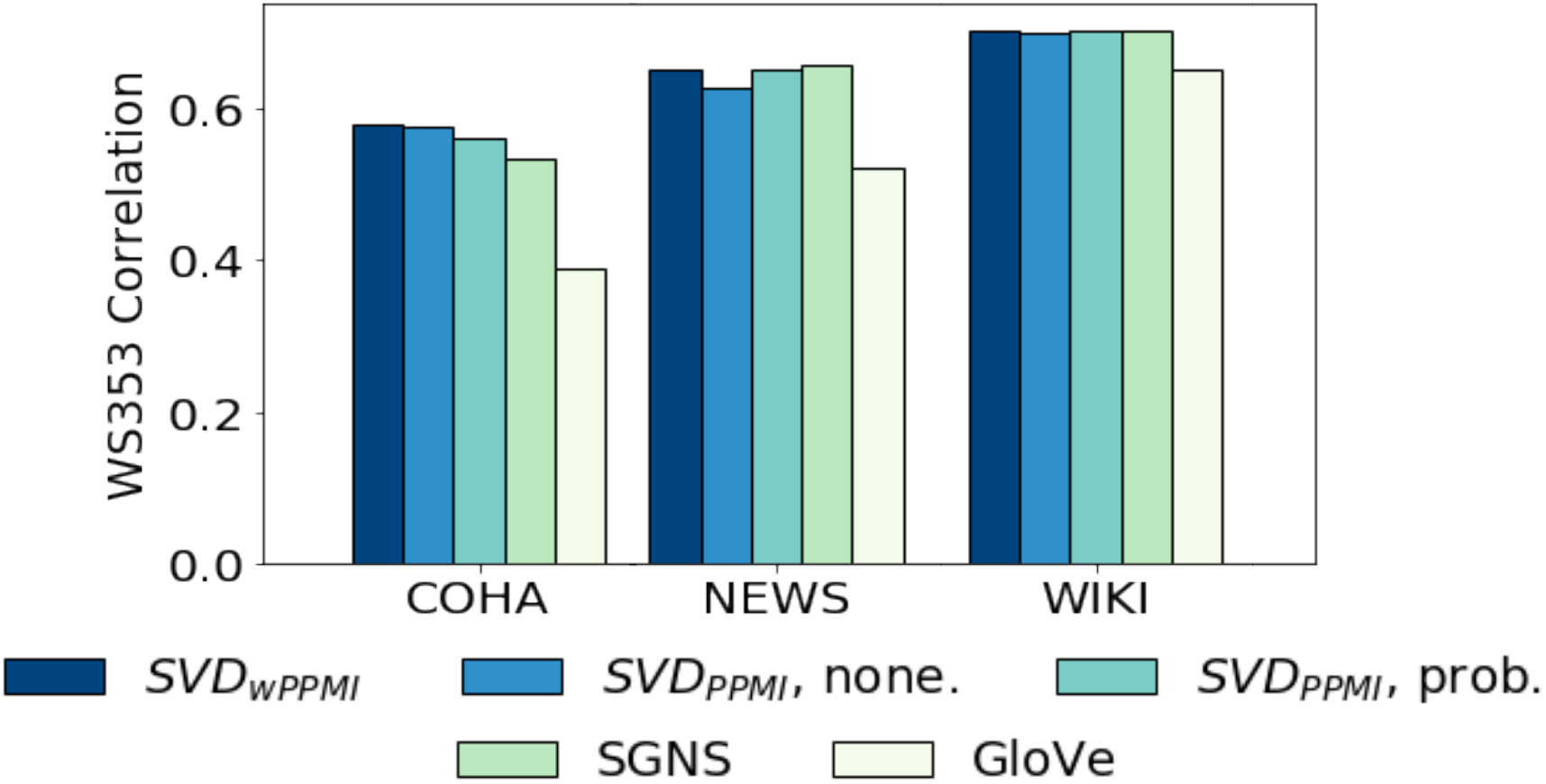
- Train 10 models each with SGNS, GloVe, SVD_{PPMI} (none / prob. down-sampling), SVD_{wPPMI}
- Evaluate intrinsically with four word similarity & two analogy test sets
- Measure stability with j@10 for 1k most frequent words

Stability Results



GloVe's high stability (Antoniak & Mimno, TACL 2018; Wendlandt et al., NAACL 2018) is true only for small corpora

Exemplary Accuracy Results



Wilcoxon rank-sum test shows SVD_{wPPMI} and SGNS to be indistinguishable in accuracy over all test sets and corpora

Conclusion

- Typical word embeddings are unstable
- Down-sampling details greatly affect stability
- GloVe's stability is worse than claimed in literature
- SVD_{wPPMI} embeddings provide SGNS-like performance and perfect stability
- See paper for additional results (and bootstrapping)

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Jena University Language & Information Engineering Lab

The Influence of Down-Sampling Strategies on SVD Word Embedding Stability

Johannes Hellrich, Bernd Kampe & Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich Schiller University Jena,
Jena, Germany
www.julielab.de