



Interpretability

Advanced Machine Learning for NLP

Jordan Boyd-Graber

NEED FOR INTERPRETABILITY

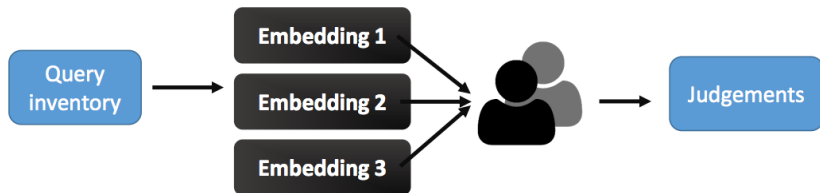
Slides adapted from Tobias Schnabel

Evaluating Word Embeddings

Query: skillfully		
(a) swiftly	(b) expertly	(c) cleverly
(d) pointedly	(e) I don't know the meaning of one (or several) of the words	

- Collected without reference to embeddings
 - Balance rare and frequent words (e.g., play vs. devour)
 - Balance POS classes (e.g., skillfully vs. piano)
 - Balance abstractness/concreteness (e.g., eagerness vs. table)
- See if embeddings can answer questions
- Perhaps not right questions to distinguish methods

Proposal

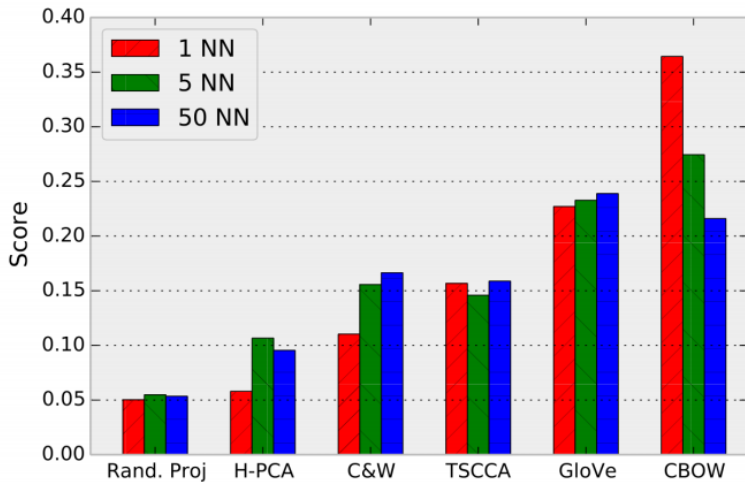


Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims.
Evaluation methods for unsupervised word embeddings, EMNLP 2015.

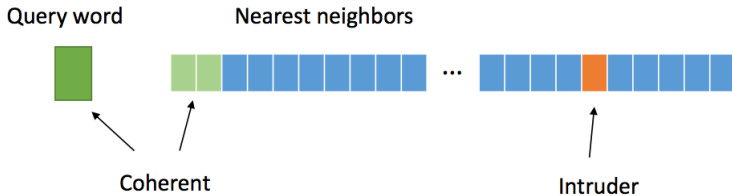
Setup

- Embeddings
 - Prediction-based: CBOW and Collobert&Weston (CW)
 - Reconstruction-based: CCA, Hellinger PCA, Random Projections, GloVe
 - Trained on Wikipedia (2008), made vocabularies the same
- Details
 - Options came from position $k = 1, 5, 50$ in NN from each embedding
 - 100 query words x 3 ranks = 300 subtasks
 - Users of Amazon Mechanical Turk answered 50 such questions
- Win score: Fraction of votes for each embedding, averaged

Winners



What about Intruders?



What about Intruders?

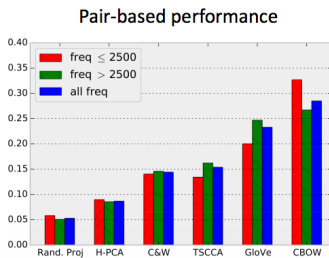
(a) finally

(b) eventually

(c) put

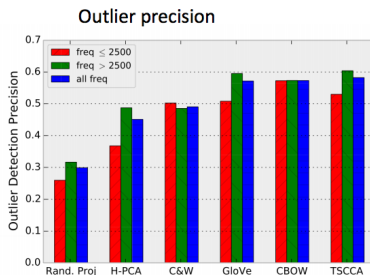
(d) immediately

What about Intruders?

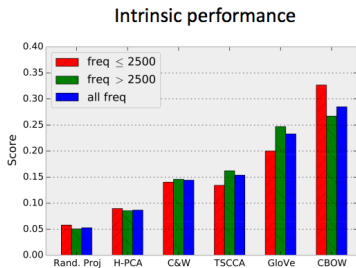


(a) Normalized scores by global word frequency.

≠

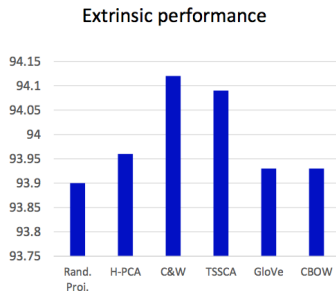


Downstream Tasks?



Normalized scores by global word frequency.

\neq



F1 chunking results