Evaluation methods for unsupervised word embeddings

EMNLP 2015

Tobias Schnabel, Igor Labutov, David Mimno and Thorsten Joachims
Cornell University

September 19th, 2015

Motivation

- How similar (on a scale from 0-10) are the following two words?

(a) tiger	(b) fauna

- **Answer:** 5.62 (According to WordSim-353)
- **Problems:**
  - Large variance ($\sigma = 2.9$)
  - Aggregation of different pairs
- **Question:** How can we improve this?

Procedure design for intrinsic evaluation

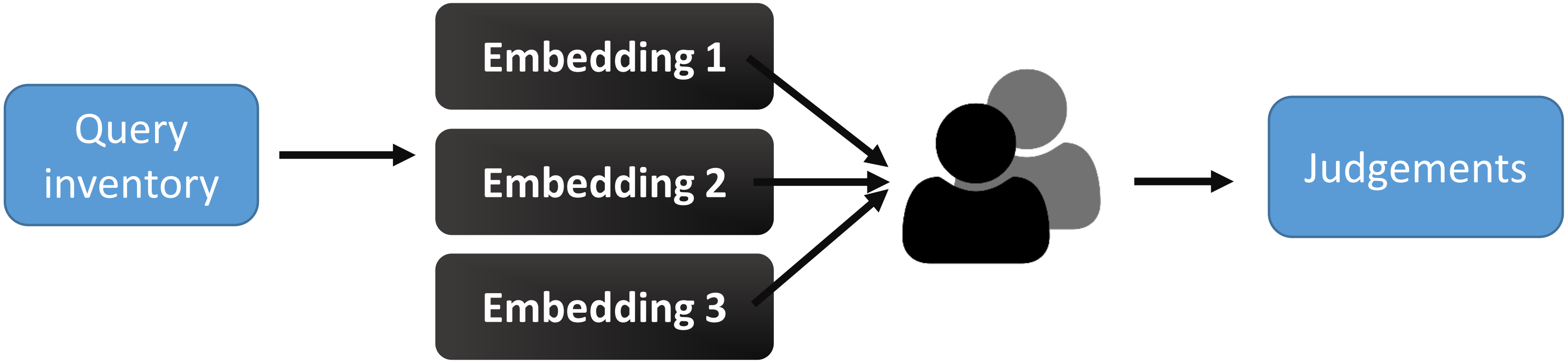
- Which option is most similar to the query word?

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

- **Answer:** 8/8 votes for (b)

Procedure design for intrinsic evaluation

Comparative evaluation (new):



Advantages:

- Directly reflects human preferences
- Relative instead of absolute judgements

Looking back

✓ How can we improve absolute evaluation?

- Comparative evaluation

... but

(a) tiger	(b) fauna

How should we pick these?

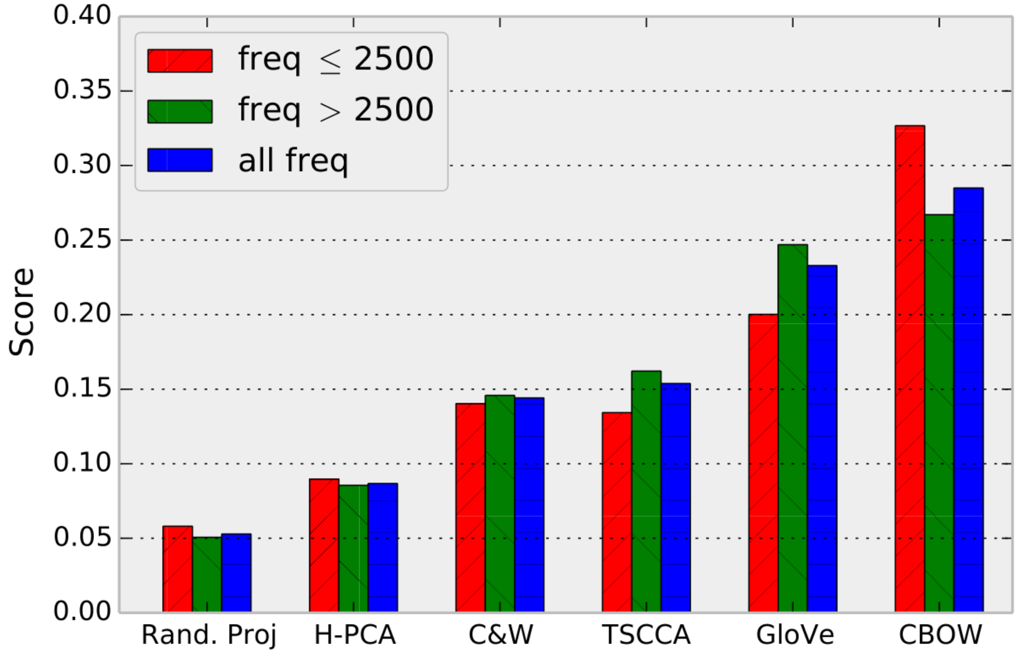
Inventory design

- **Often:** Heuristically chosen
- **Goal:** Linguistic insight
- Aim at **diversity** and **balancedness:**
    - 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)

Results

- **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

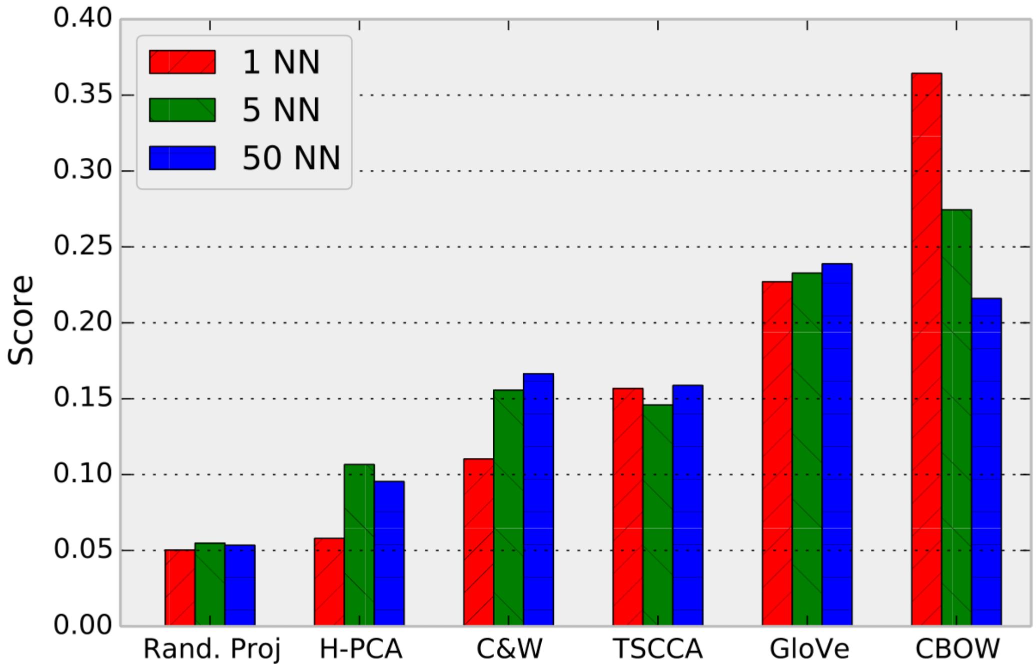
Results – by frequency



Normalized scores by global word frequency.

⇒ Performance varies with word frequency

Results – by rank



Normalized scores by nearest neighbor rank k.

⇒ Different falloff behavior

Results – absolute performance

	relatedness						categorization			sel. prefs		analogy			
	rg	ws	wss	wsr	men	toefl	ap	esslli	batt.	up	mcrae	an	ansyn	ansem	average
CBOW	**74.0**	**64.0**	**71.5**	**56.5**	**70.7**	66.7	**65.9**	**70.5**	**85.2**	24.1	13.9	**52.2**	**47.8**	**57.6**	**58.6**
GloVe	63.7	54.8	65.8	49.6	64.6	**69.4**	64.1	65.9	77.8	27.0	**18.4**	42.2	44.2	39.7	53.4
TSCCA	57.8	54.4	64.7	43.3	56.7	58.3	57.5	**70.5**	64.2	**31.0**	14.4	15.5	19.0	11.1	44.2
C&W	48.1	49.8	60.7	40.1	57.5	66.7	60.6	61.4	80.2	28.3	16.0	10.9	12.2	9.3	43.0
H-PCA	19.8	32.9	43.6	15.1	21.3	54.2	34.1	50.0	42.0	-2.5	3.2	3.0	2.4	3.7	23.1
Rand. Proj.	17.1	19.5	24.9	16.1	11.3	51.4	21.9	38.6	29.6	-8.5	1.2	1.0	0.3	1.9	16.2

Results on absolute intrinsic evaluation

⇒ Similar results for absolute metrics

However: Absolute metrics less principled and insightful

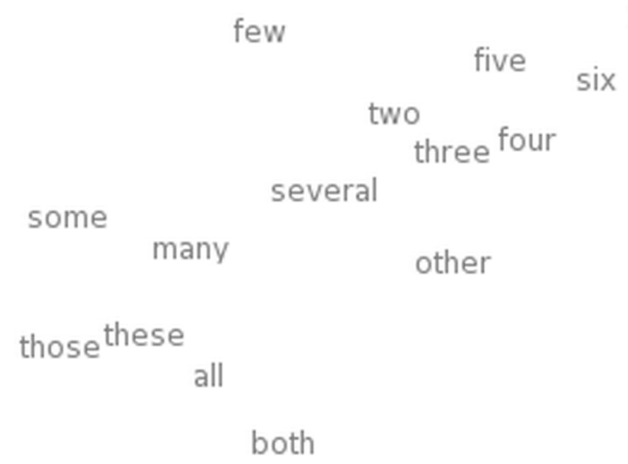
Looking back

- How can we improve absolute evaluation?
  - Comparative evaluation
- How should we pick the query inventory?
  - Strive for diversity and balancedness

… but

(a) tiger ⟷ (b) fauna

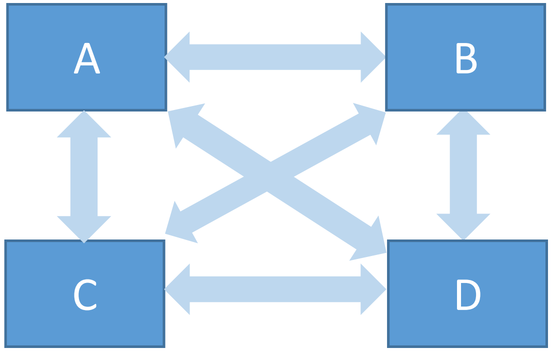
Are there more global properties?



Properties of word embeddings

- Common: Pair-based evaluation, e.g.,
  - Similarity/relatedness
  - Analogy

- Idea: Set-based evaluation
  - All interactions considered
  - Goal: measure coherence



Properties of word embeddings

- What word belongs the least to the following group?

(a) finally	(b) eventually
(c) put	(d) immediately

Answer: put (8/8 votes)

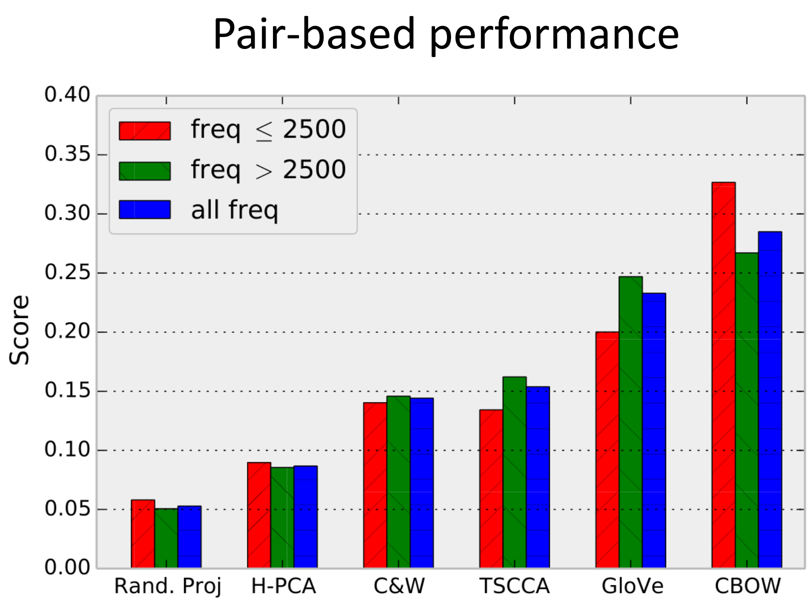
Properties of word embeddings

- Construction:

(a) finally	(b) eventually
(c) put	(d) immediately

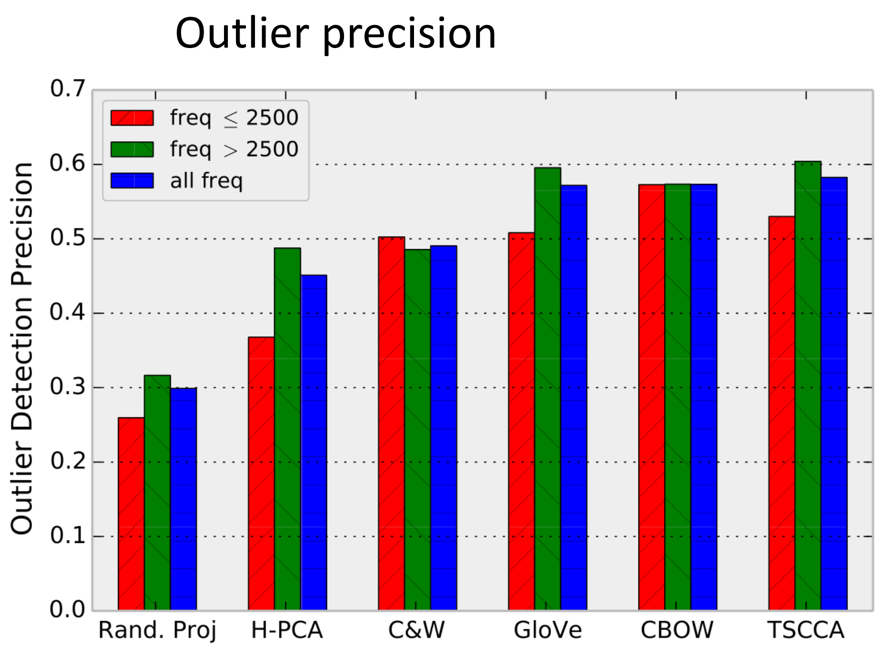
- For each embedding, create sets of 4 with one intruder

Results



(a) Normalized scores by global word frequency.

$\neq$



⇒ Set-based evaluation ≠ item-based evaluation

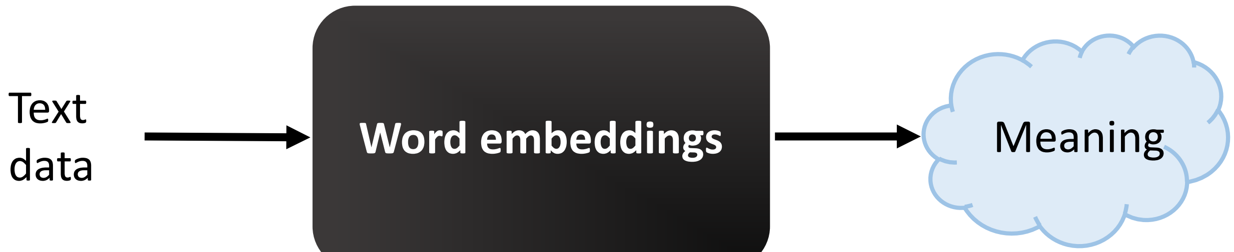
Looking back

- How can we improve absolute evaluation?
  - Comparative evaluation
- How should we pick the query inventory?
  - Strive for diversity and balancedness
- Are there other interesting properties?
  - Coherence

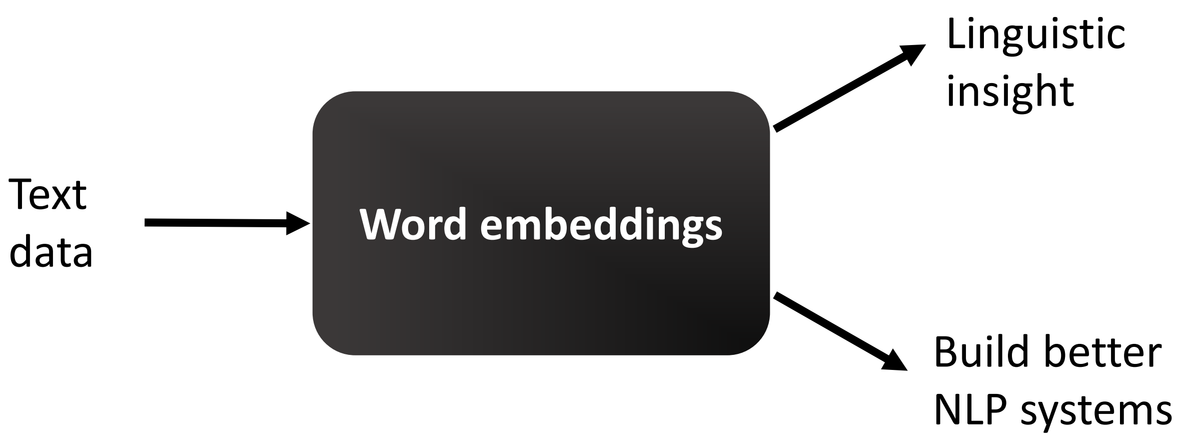
… but

What about downstream performance?

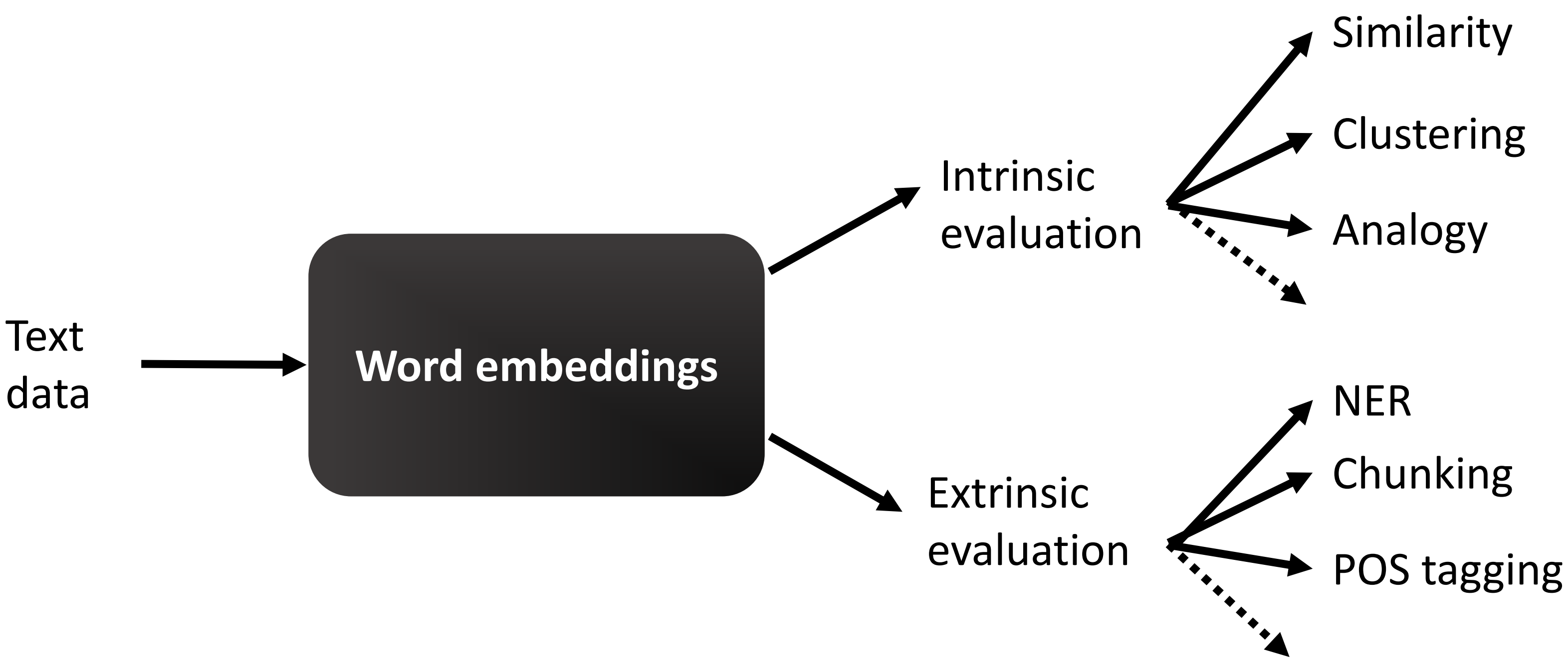
The big picture



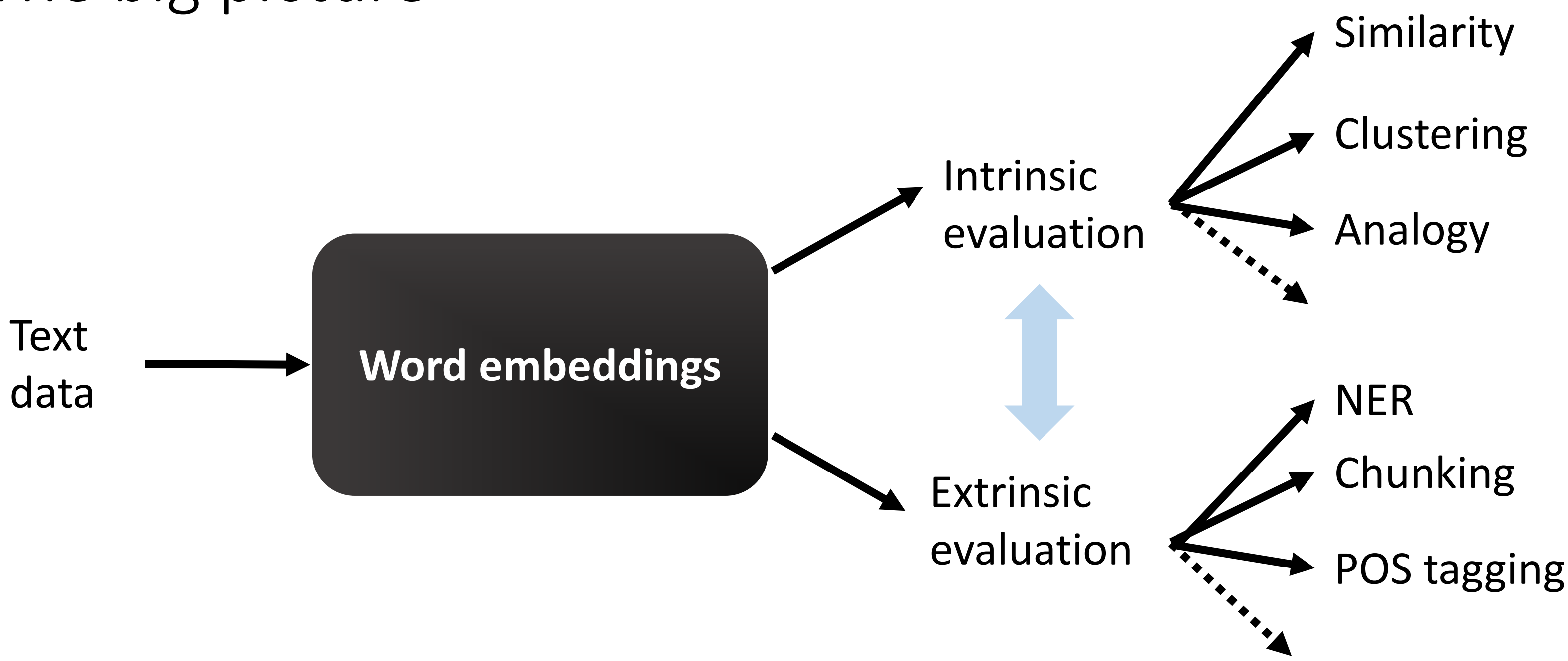
The big picture



The big picture



The big picture

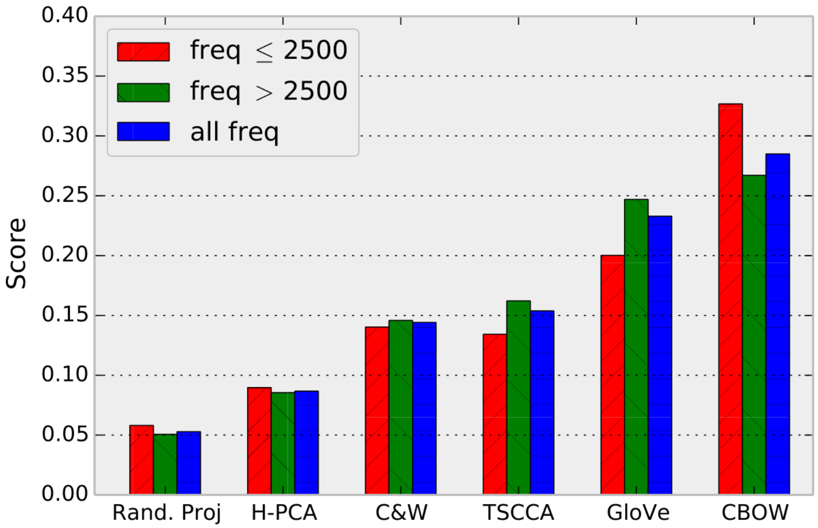


Extrinsic vs. intrinsic performance

- Hypothesis:
  - Better intrinsic quality also gives better downstream performance
- Experiment:
  - Use each word embedding as extra features in supervised task

Results – Chunking

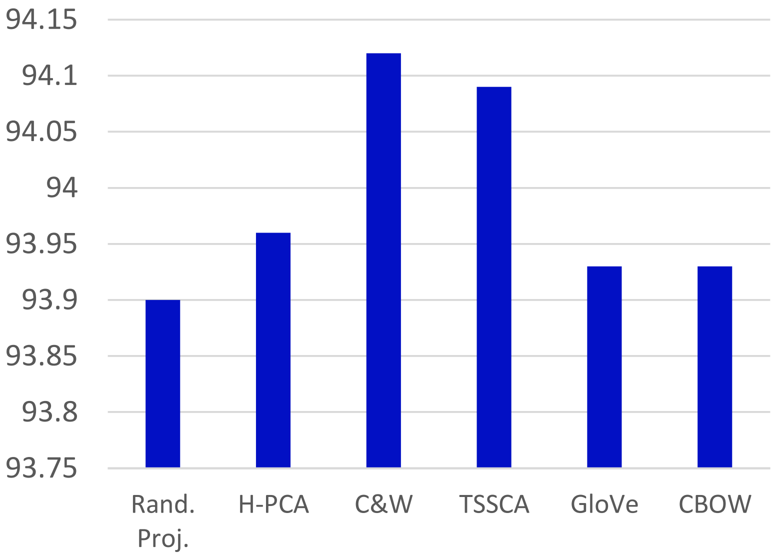
Intrinsic performance



Normalized scores by global word frequency.

$\neq$

Extrinsic performance



F1 chunking results

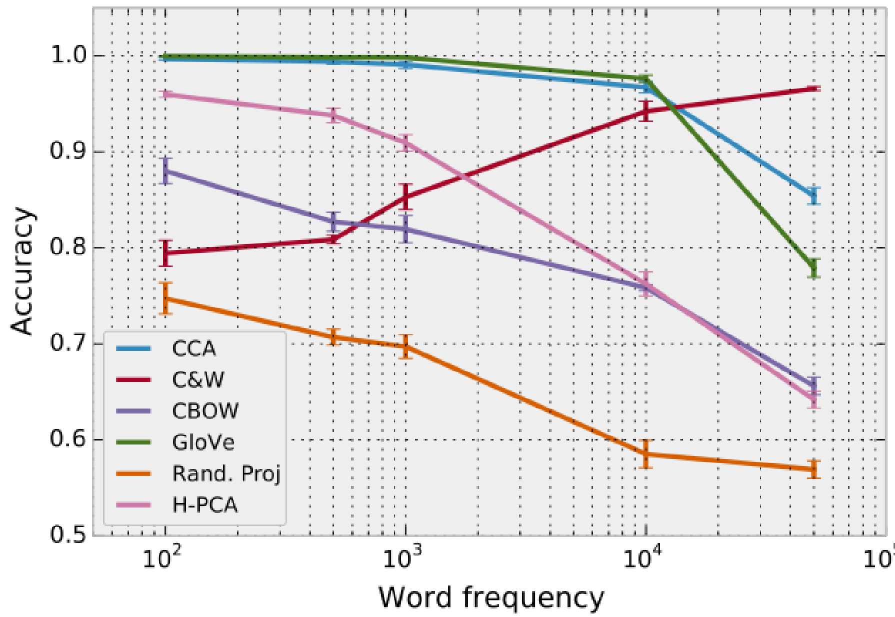
$\Rightarrow$ Intrinsic performance $\neq$ extrinsic performance

Looking back

- ✓ How can we improve absolute evaluation?
  - Comparative evaluation
- ✓ How should we pick the query inventory?
  - Strive for diversity and balancedness
- ✓ Are there other interesting properties?
  - Coherence
- ✓ Does better intrinsic performance lead to better extrinsic results?
  - No!

Discussion

- Why do we see such different behavior?
  - Hypothesis: Unwanted information encoded as well
- Embeddings can accurately predict word frequency



Discussion

- **Also:** Experiments show strong correlation of word frequency and similarity
- Further problems with cosine similarity:
  - Used in almost all intrinsic evaluation tasks – conflates different aspects
  - Not used during training: disconnect between evaluation and training
- **Better:**
  - Learn custom metric for each task (e.g., semantic relatedness, syntatic similarity, etc.)

Conclusions

- Practical recommendations:
  - Specify what the goal of an embedding method is
  - Advantage: Now able to use datasets to inform training
- Future work:
  - Improving similarity metrics
  - Use data from comparative experiments to do offline evaluation
- All data and code available at:
  - http://www.cs.cornell.edu/~schnabts/eval/