Siamese CBOW

Optimizing Word Embeddings for Sentence Representations

Tom Kenter Alexey Borisov Maarten de Rijke

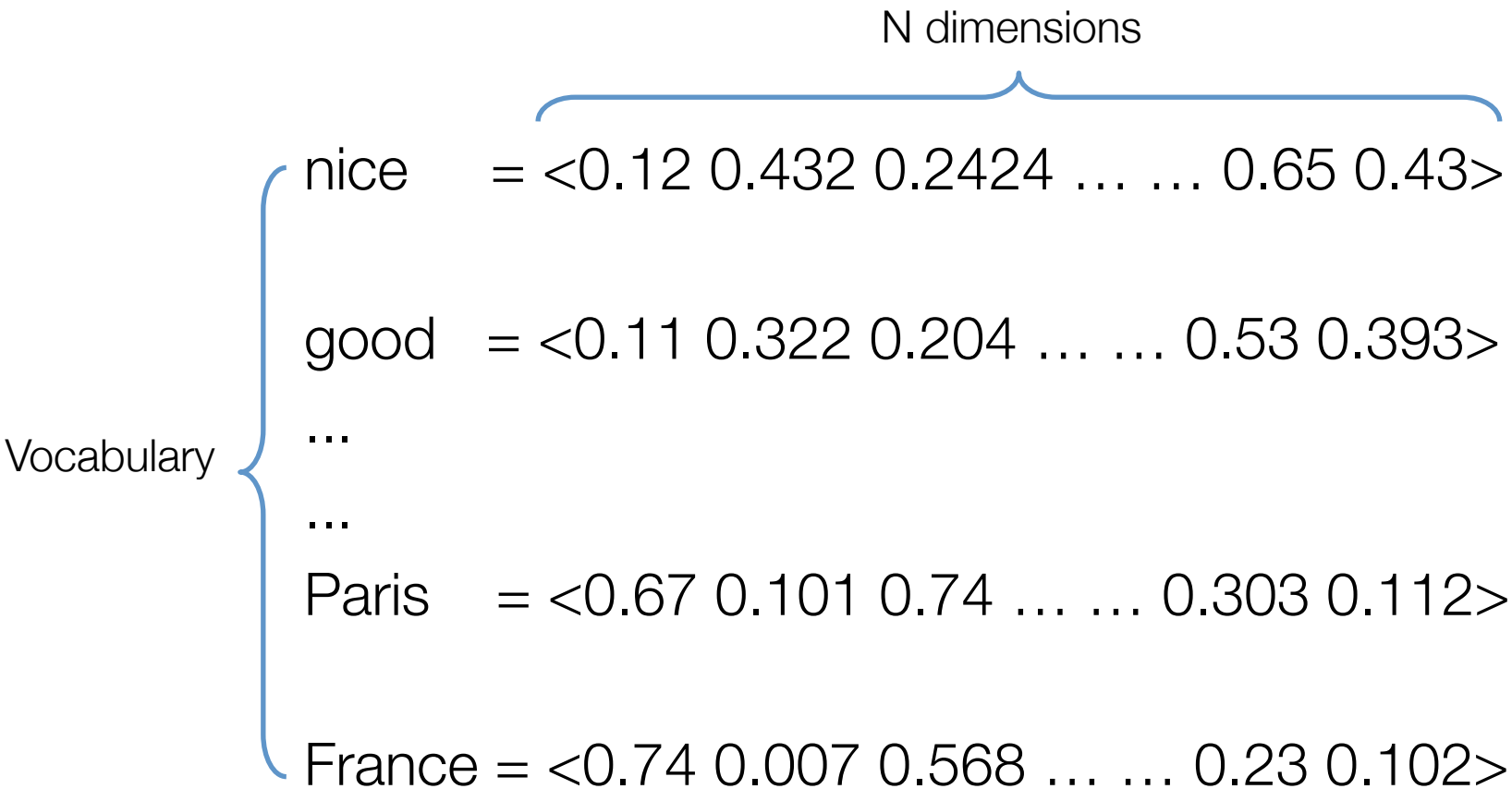
SEA – Nov 18 2016

Tom Kenter
PhD candidate
ILPS, University of Amsterdam

UNIVERSITY OF AMSTERDAM

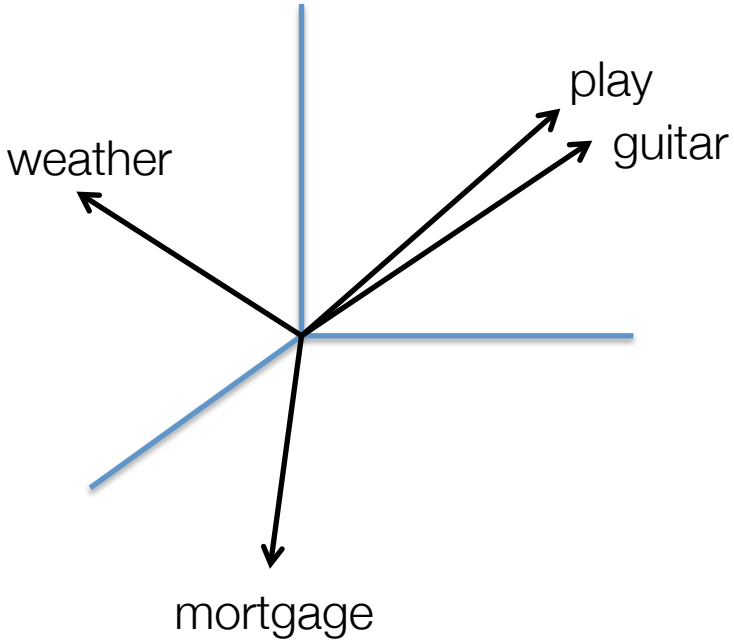
Word embeddings

Vector representations of words

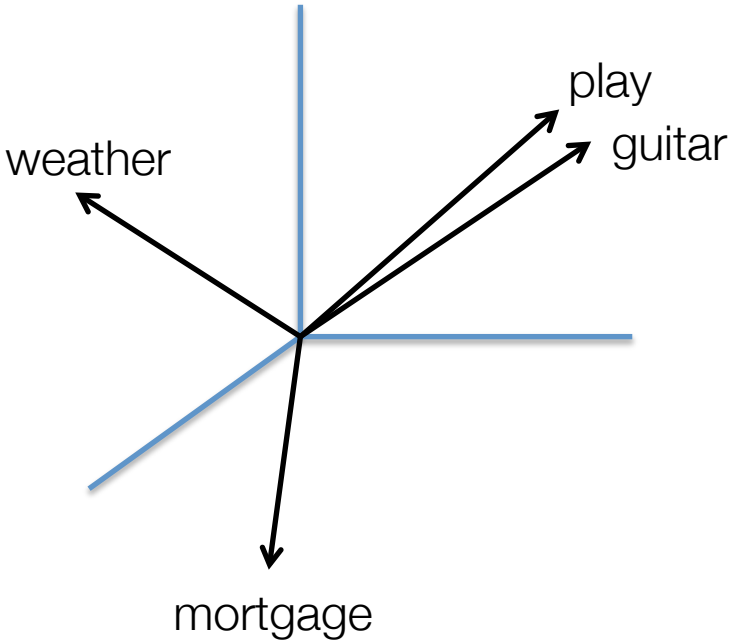


Word embeddings

Vector representations of words



Word embeddings



angle between $\overrightarrow{\text{play}}$ and $\overrightarrow{\text{guitar}}$

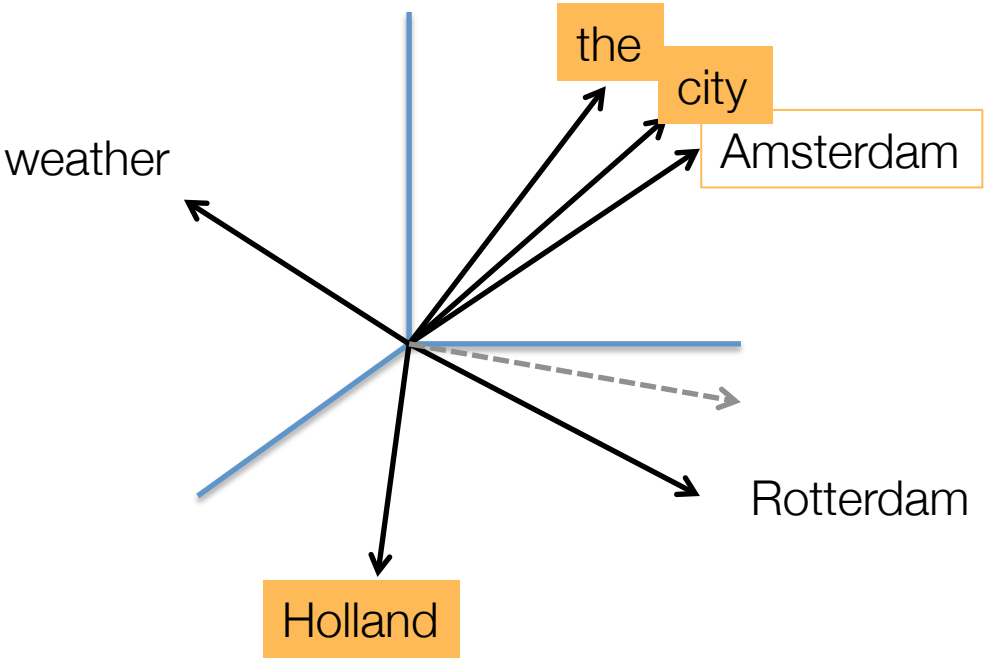
is smaller than

angle between $\overrightarrow{\text{play}}$ and $\overrightarrow{\text{weather}}$

$\downarrow$

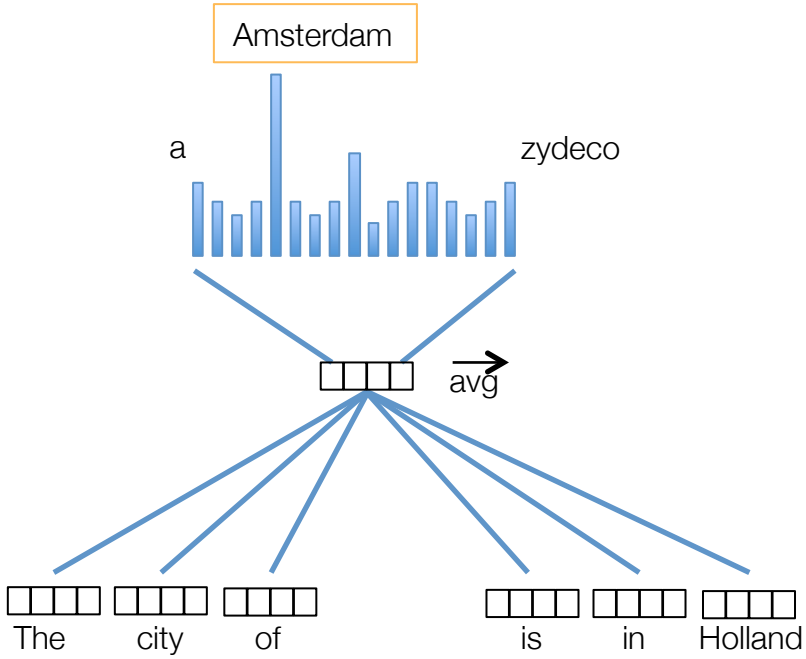
cosine($\overrightarrow{\text{play}}$, $\overrightarrow{\text{guitar}}$) < cosine($\overrightarrow{\text{play}}$, $\overrightarrow{\text{weather}}$)

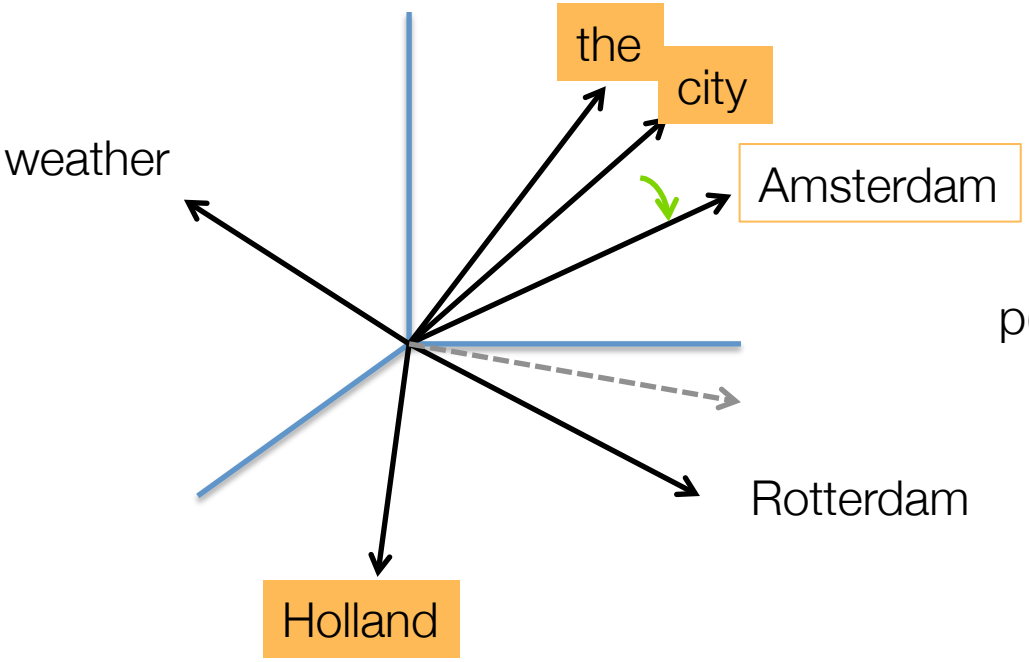
$$\text{semantic similarity}(A, B) = \text{cosine}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \cdot ||\vec{B}||} = \frac{\sum_{i=1}^{n} \vec{A}_i \times \vec{B}_i}{\sqrt{\sum_{i=1}^{n} (\vec{A}_i)^2} \times \sqrt{\sum_{i=1}^{n} (\vec{B}_i)^2}}$$



Training

$$p(\text{Amsterdam} \mid \text{city, Holland, in, is, of, The}) \propto \overrightarrow{\text{Amsterdam}} * \overrightarrow{\text{avg}}$$

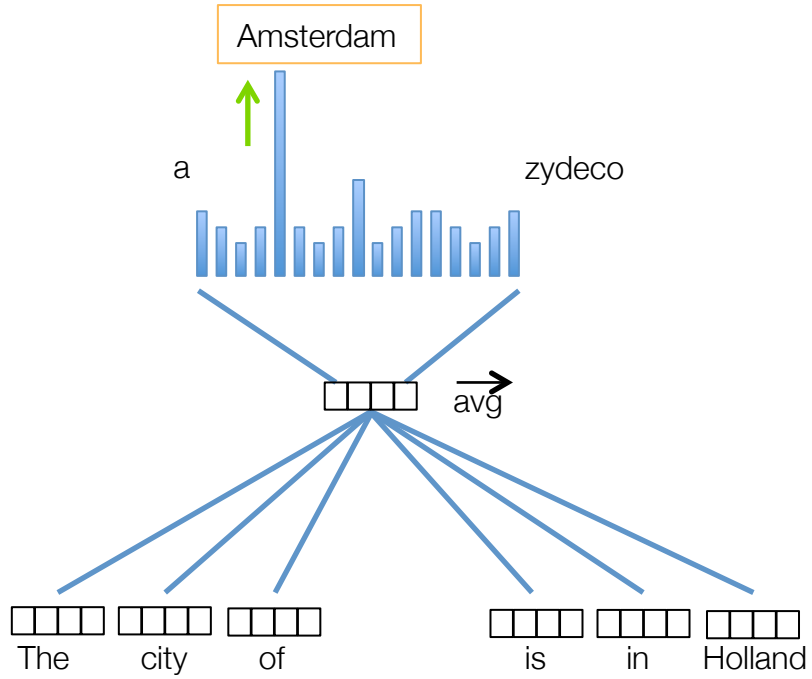




Training

p(Amsterdam | city, Holland, in, is, of, The)

$$\propto \overrightarrow{\text{Amsterdam}} * \overrightarrow{\text{avg}}$$



Sentence embeddings

From word-level semantics to sentence-level semantics

In stead of embeddings for words, how about embeddings for
- sentences?
- paragraphs?
- documents?

Evaluation task
Sentence similarity: given two sentences, do they mean the same thing?

Siamese CBOW

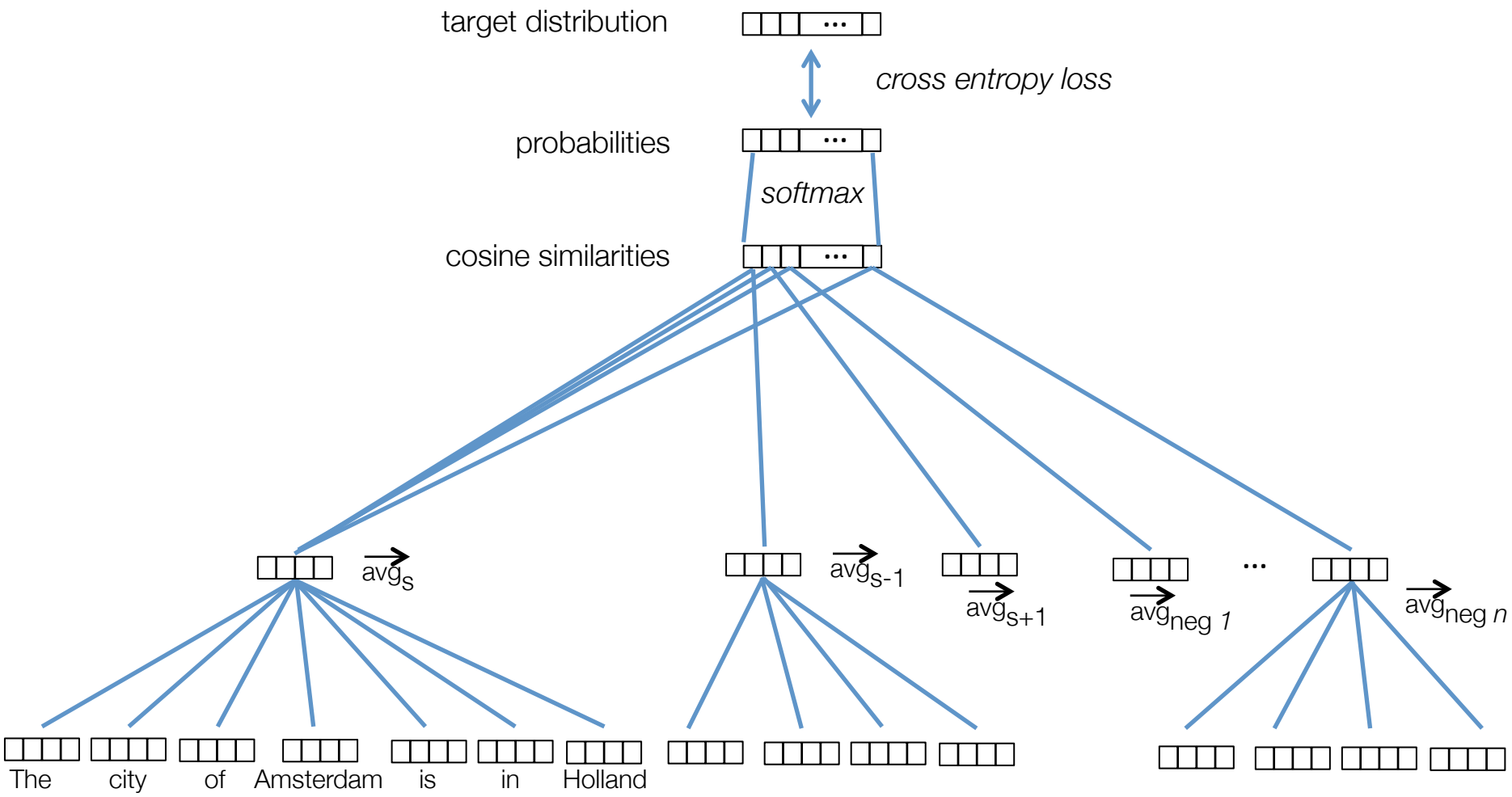
Average embeddings of words in a sentence

Predict the current sentence from the sentences before and after it.
Similar to what word2vec does with words

Sentences that have similar surrounding sentences should get similar representations

Directly optimize word embeddings for the task of being averaged

Siamese CBOW



Solving the softmax problem

We compare sentence representations to sentence representations, so there is no vocabulary to softmax over

There is no vocabulary to sample from

In theory, we should do a softmax over all possible sentences

Instead, we consider the surrounding sentences to be positive examples: S^+
we sample random sentences as negative examples: S^-

We do a softmax over $S^+ \cup S^-$

Results on SemEval datasets in terms of Pearson's r (Spearman's r). Highest scores, in terms of Pearson's r, are displayed in bold. Siamese CBOW runs statistically significantly different from the word2vec CBOW baseline runs are marked with a †

Dataset	w2v skipgram	w2v CBOW	skip-thought	Siamese CBOW
2012				
MSRpar	.3740 (.3991)	.3419 (.3521)	.0560 (.0843)	**.4379**† (.4311)
MSRvid	.5213 (.5519)	.5099 (.5450)	**.5807** (.5829)	.4522† (.4759)
OnWN	.6040 (.6476)	.6320 (.6440)	.6045 (.6431)	**.6444**† (.6475)
SMTeuroparl	.3071 (.5238)	.3976 (.5310)	.4203 (.4999)	**.4503**† (.5449)
SMTnews	**.4487** (.3617)	.4462 (.3901)	.3911 (.3628)	.3902† (.4153)
2013				
FNWN	**.3480** (.3401)	.2736 (.2867)	.3124 (.3511)	.2322† (.2235)
OnWN	.4745 (.5509)	**.5165** (.6008)	.2418 (.2766)	.4985† (.5227)
SMT	.1838 (.2843)	.2494 (.2919)	**.3378** (.3498)	.3312† (.3356)
headlines	.5935 (.6044)	.5730 (.5766)	.3861 (.3909)	**.6534**† (.6516)
2014				
OnWN	.5848 (.6676)	.6068 (.6887)	.4682 (.5161)	**.6073**† (.6554)
deft-forum	.3193 (.3810)	.3339 (.3507)	.3736 (.3737)	**.4082**† (.4188)
deft-news	.5906 (.5678)	.5737 (.5577)	.4617 (.4762)	**.5913**† (.5754)
headlines	.5790 (.5544)	.5455 (.5095)	.4031 (.3910)	**.6364**† (.6260)
images	.5131 (.5288)	.5056 (.5213)	.4257 (.4233)	**.6497**† (.6484)
tweet-news	.6336 (.6544)	.6897 (.6615)	.5138 (.5297)	**.7315**† (.7128)
2015				
answ-forums	.1892 (.1463)	.1767 (.1294)	**.2784** (.1909)	.2181 (.1469)
answ-students	.3233 (.2654)	.3344 (.2742)	.2661 (.2068)	**.3671**† (.2824)
belief	.2435 (.2635)	.3277 (.3280)	.4584 (.3368)	**.4769** (.3184)
headlines	.1875 (.0754)	.1806 (.0765)	.1248 (.0464)	**.2151**† (.0846)
images	.2454 (.1611)	.2292 (.1438)	.2100 (.1220)	**.2560**† (.1467)

Conclusions

- Siamese CBOW is a fast and reliable algorithm for generating word embeddings optimized for being averaged across sentences
- You can learn sentence representations from unsupervised data by learning from context sentences
- Using only a small number of negative examples is sufficient, which allows for very fast training