FACULTY
OF INFORMATICS
Masaryk University

DisSim: the Disambiguation of Words for Computing the Similarity of Documents

Project MUNI 33 / 10 2016

https://gitlab.fi.muni.cz/xnovot32/dissim-jupyter

Vít Novotný

Contents

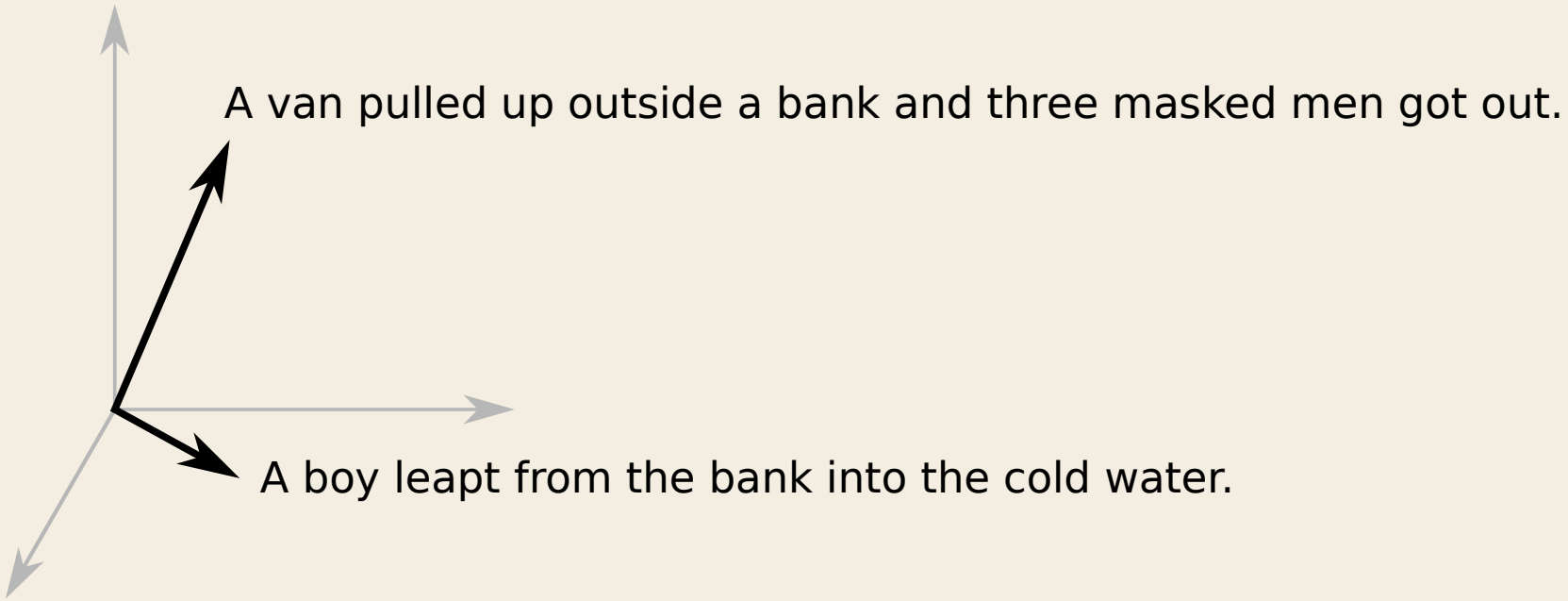
Section 1

Introduction

Introduction

Vector Space Document Model

A class of document representations that encode documents in a high-dimensional vector space:



The classic IR vector space model is due to Salton et al. (1975).

Introduction

Word Sense Disambiguation

In computational linguistics, word-sense disambiguation (WSD) refers to the problem of identifying the sense in which a word was used:

- A van pulled up outside a bank and three masked men got out.
- A boy leapt from the bank into the cold water.

One of the standard approaches is to perform clustering on the vector-space representation of words (Anaya-Sánchez et al., 2006; Rafael Berlanga LLavori, 2012).

Introduction

The DisSim Algorithm

The following steps are repeated until a termination condition is met:

- A vector-space document model is constructed from a corpus.
- Individual words are represented by the vectors of their parent documents. After clustering, each word is replaced with the name of its cluster:
  - A van pulled up outside a bank_1 and three masked men got out.
  - A boy leapt from the bank_2 into the cold water.
- A new vector-space document model is constructed from the updated documents.

It is expected that the diambiguated model will yield better results.

Section 2

Experimental Setup

Experimental Setup

Doc2vec, K-means, and Agglomerative Clustering

As the vector-space document model, the doc2vec algorithm by Le et al. (2014) in the skip-gram variant as implemented in the Gensim Python library was used. Semantically similar documents have been empirically shown to form clusters in this model.
For clustering, we used:

- several variants of *K*-means, where *K* was the number of senses of a word given by WordNet,
- agglomerative clustering using a linkage that minimizes the variance of the clusters being merged (Ward Jr, 1963),
- a novel graph-theoretic algorithm based on the authority-propagation algorithm of PageRank (Page et al., 1999).

Experimental Setup

Training Corpus, Gold Standard, and Evaluation

As the training corpus, we used various a mixture of various corpora: Aesop's Fables, BBC news and sport news, book reviews from amazon.com, Q&A from Yahoo! Answers, and articles from the English Wikipedia.

As the gold standard, we used test data from the SemEval competition. The test data consisted of 500 paragraph-sentence pairs with human-assigned semantic similarity score.

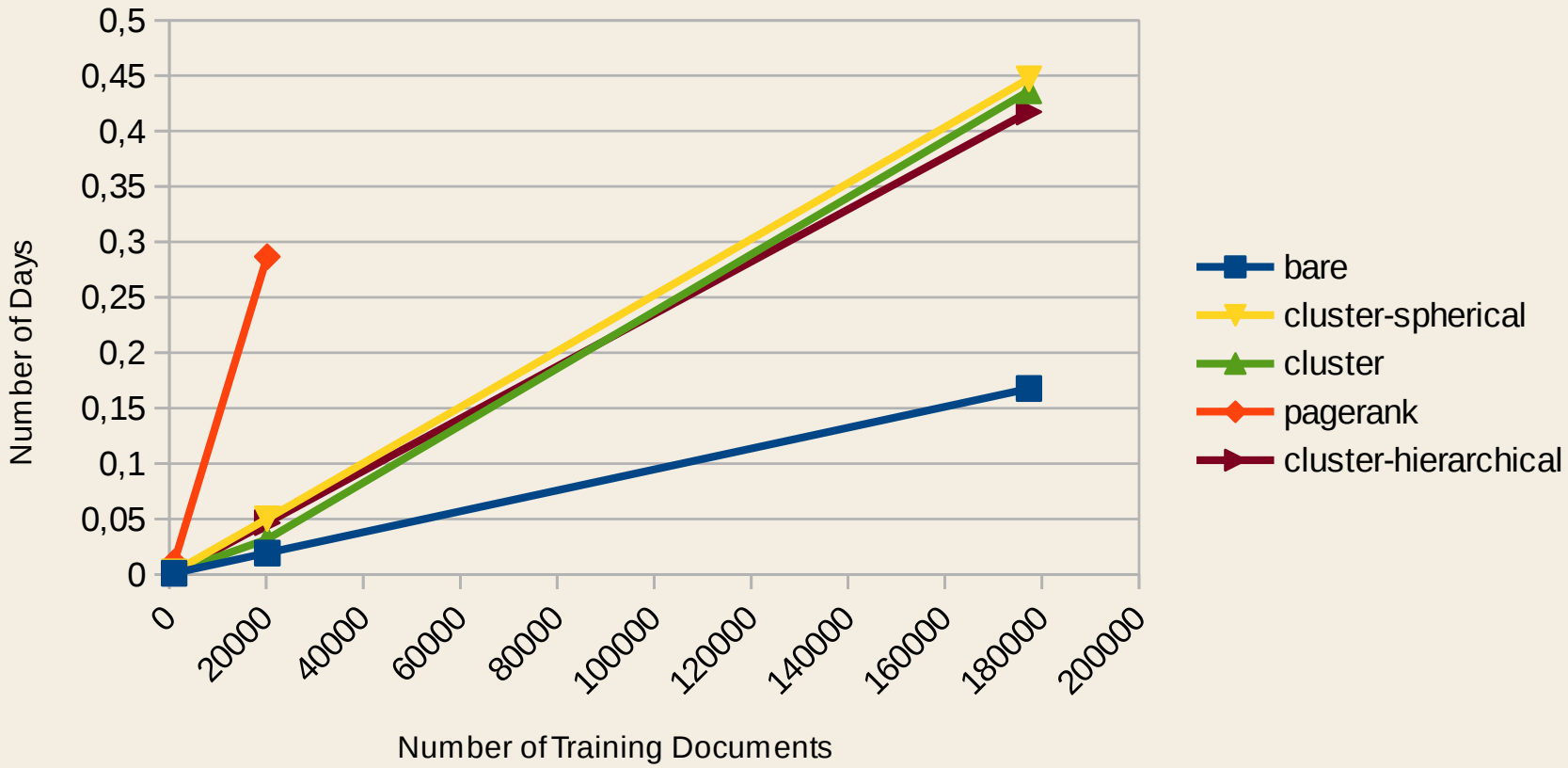
To evaluate a model, we predicted the semantic similarity of every test paragraph-sentence pair by computing cosine similarity between the vector-space representations of the paragraph and the sentence. Then we computed Pearson's correlation coefficient F between the human-assigned semantic similarity scores and the predictions.

Section 3

Results

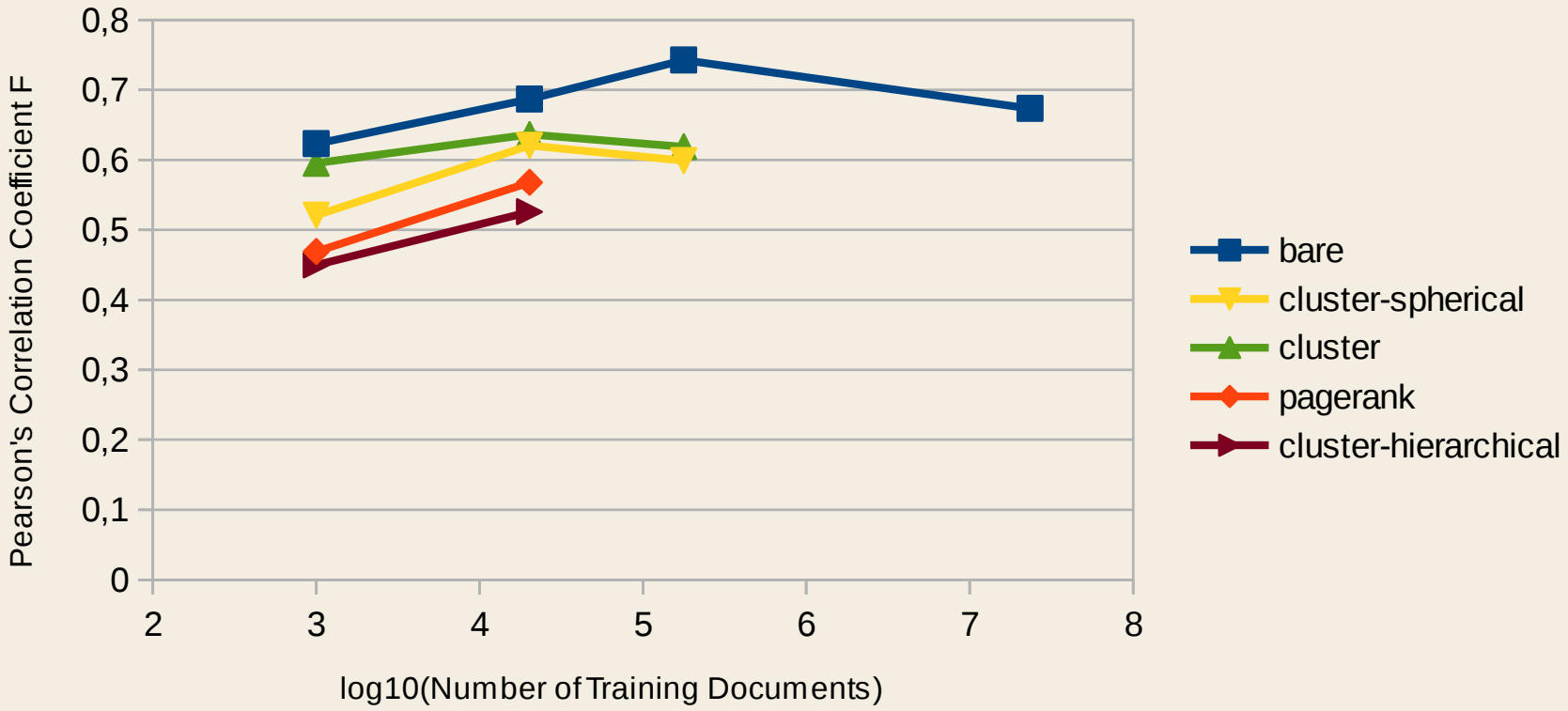
Results

Training Speed



Results

Correlation Between Predictions and Gold Scores



Section 4

Discussion

Discussion

- Compared to SemEval results, bare doc2vec model scores high.
  - A variant of this model won SemEval 2017.
- WordNet has been shown to overestimate the number of meaningful word senses, reducing the quality of clustering.
- A gold standard and an evaluation metric tailored specifically for WSD might yield better results.

References I

ANAYA-SÁNCHEZ, Henry; PONS-PORRATA, Aurora; BERLANGA-LLAVORI, Rafael, 2006. Word Sense Disambiguation Based on Word Sense Clustering. In: *Advances in Artificial Intelligence - IBERAMIA-SBIA 2006: 2nd International Joint Conference, 10th Ibero-American Conference on AI, 18th Brazilian AI Symposium, Ribeirão Preto, Brazil, October 23-27, 2006. Proceedings*. Ed. by SICHMAN, Jaime Simão; COELHO, Helder; REZENDE, Solange Oliveira. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 472–481. ISBN 978-3-540-45464-9. Available from DOI: 10.1007/11874850_51.

LE, Quoc V; MIKOLOV, Tomas, 2014. Distributed Representations of Sentences and Documents. In: *Distributed Representations of Sentences and Documents. ICML*. Vol. 14, pp. 1188–1196.

References II

PAGE, Lawrence; BRIN, Sergey; MOTWANI, Rajeev; WINOGRAD, Terry, 1999. *The PageRank citation ranking: Bringing order to the web*. Technical report. Stanford InfoLab.

RAFAEL BERLANGA LLAVORI, Tamara Martín Wanton y, 2012. A clustering-based Approach for Unsupervised Word Sense Disambiguation. *Procesamiento del Lenguaje Natural*. Vol. 49, no. 0, pp. 49–56. ISSN 1989-7553. Available also from: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4553.

SALTON, G.; WONG, A.; YANG, C. S., 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM*. Vol. 18, no. 11, pp. 613–620. ISSN 0001-0782. Available from DOI: 10.1145/361219.361220.

References III

WARD JR, Joe H, 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. Vol. 58, no. 301, pp. 236–244.