A review of word embedding and document similarity algorithms applied to academic text

Computer Science

Bachelor's Thesis

Author:
Jon Ezeiza Alvarez

Supervisor:
Prof. Dr. Hannah Bast

Motivation

- **A consequence of two projects:**
  - IXA group practicum
  - SCITODATE
- **A realization:**
  - There is no human endevour as well documented as science.
  - With faster progress and increased publication rate it is getting hard for humans to keep a global grasp of science.
- **A long-term goal:** An AI toolbox for automatic understanding of large amounts of academic literature.

Scope

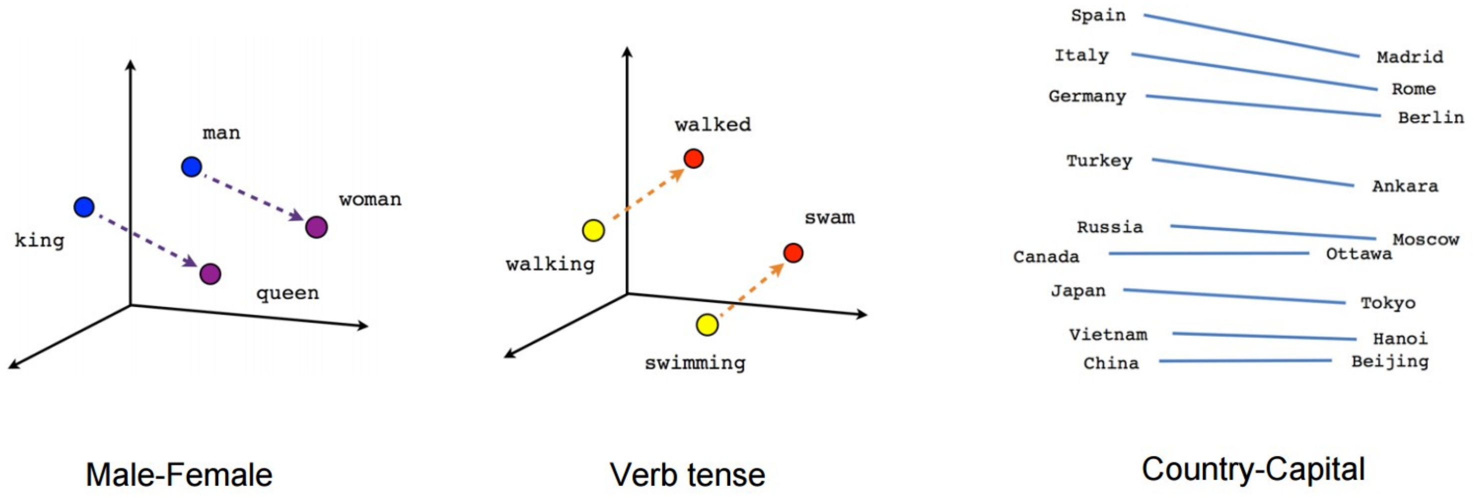
- **A small first step**
  - **Literature review** of the state-of-the-art in word embeddings and semantic textual similarity.
  - **Empirical review** of the algorithms on academic literature.

What are word embeddings?

- **Dense algebraic representations of semantic content.**
- **Trained on large corpora or knowledge graphs.**
- **Why?**
  - An alternative to knowledge graphs.
  - Input for Machine Learning.

What are word embeddings?

- Words are placed in a high dimensional vector space such that their distances equate similarity or relatedness.
- **Side effect:** Analogy, real-world knowledge



Semantic Textual Similarity (STS)

- **Task:** approximate similarity between pairs of text.
  - Phrases
  - Sentences
  - Paragraphs
  - Documents
- **Document embeddings**
  - Word embedding compositionality.

Training dataset

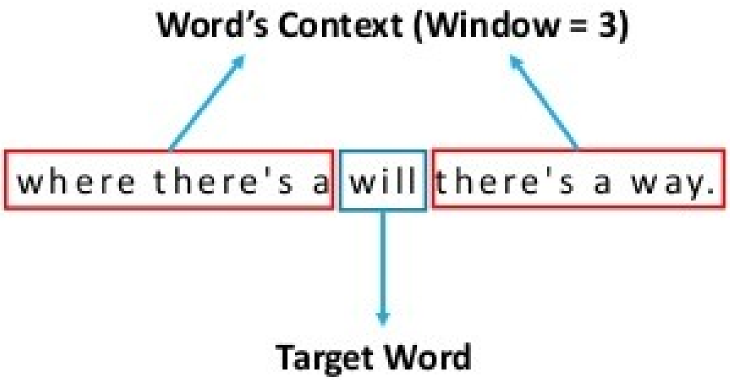
- **A corpus to learn from**
  - Bio-medical articles from PubMed
  - 3 billion tokens
  - Separate titles, abstracts and bodies.
  - Cleaned and normalized:
    - Tokenization
    - Stemming

Testing datasets

- **Triplets:** distinguish similarity from noise.
  - The first two elements are related.
  - The third element is non-related.
  - **Goal:** $\text{sim}(1, 2) > \text{sim}(1, 3)$
- **Word embeddings:** UMLS synonyms.
- **Document similarity:** ORCID author linking.

Word2Vec (Mikolov, K. Chen, et al., 2013)

- **Mayor breakthrough**
  - Key to success: **deep vs shallow** models
- **Window scanning method:**
  - **Assumption:** words that appear in similar contexts have similar meaning (Harris, 1954).



GloVe (Pennington, Socher, and C. Manning, 2014)

- **Formalization of window scanning method:** implicit factorization of word-word global statistics matrix.

- **Alternative:**

  – Explicit factorization of co-occurrene matrix.

FastText (Bojanowski et al., 2016)

- **Word2Vec with subword components.**
  - Modular word embeddings.
  - N-gram embeddings.
  - Compositon of subword structures.
  - Robustness to language inconsistencies and morphological variations.

WordRank (Ji et al., 2015)

- **Optimizes Nearest Neighbour ranking**
  - Instead of target-context pairwise distance.
  - Ranking tuned to have more resolution at the top.
  - Similar results to state-of-the-art with smaller corpora.
    - Not reflected in our experiments.

Results and conclusions

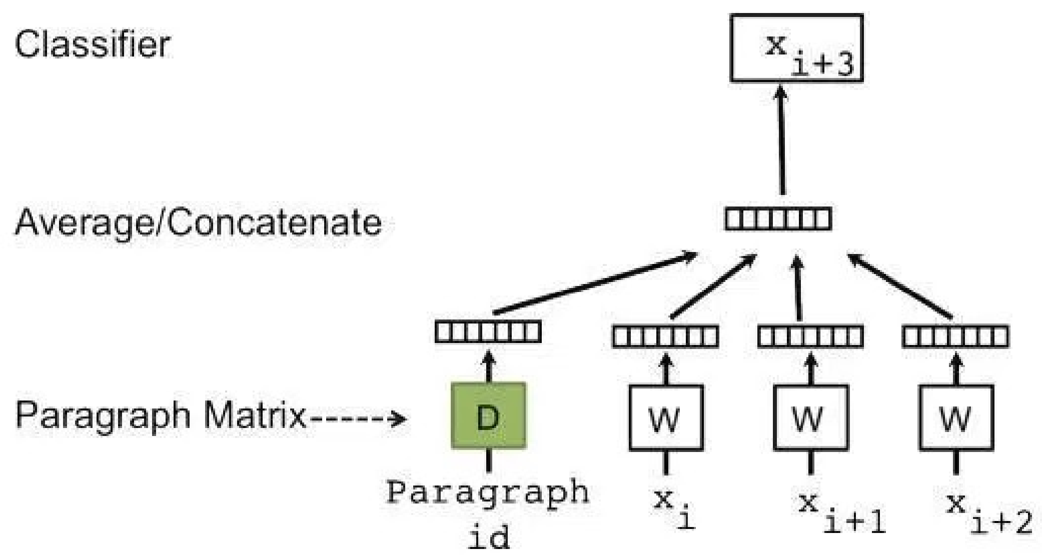
Word embeddings accuracy	1M	10M	100M	1B	2B
W2V CBow - Total	0.03	0.17	0.46	0.83	0.89
W2V Skip-gram - Total	0.04	0.18	0.46	0.83	0.89
W2V CBow - Known	0.67	0.73	0.80	0.85	**0.90**
W2V Skip-gram - Known	0.67	0.79	0.80	0.88	**0.90**
GloVe - Total	0.04	0.17	0.45	0.80	0.87
GloVe - Known	0.71	0.73	0.78	0.85	0.88
FastText - Total	**0.81**	**0.88**	**0.90**	**0.93**	-
WordRank - Total	0.02	0.21	0.45	0.78	0.89
Wordrank - Known	0.69	0.75	0.77	0.84	**0.90**

STS Baseline

- **It is early days for STS**
  - Make sure that the state-of-the-art beats naive methods.
- **Baseline:**
  - VSM similarity: BoW, Tf-Idf, BM25
  - Weighted word embedding centroids

Doc2Vec (Quoc V. Le and Mikolov, 2014)

- **Adaptation of Word2Vec**
  - Add global document vector to the context.

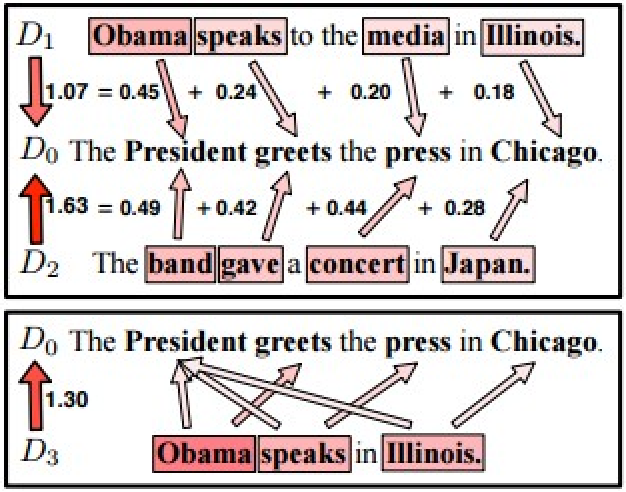


Doc2VecC (M. Chen, 2017)

- **Realization:** simple word embedding average is a hard baseline to beat.
  - Optimize word embeddings such that averaging them results in meaningful document vector representations.
  - Heavy corruption to improve generality.

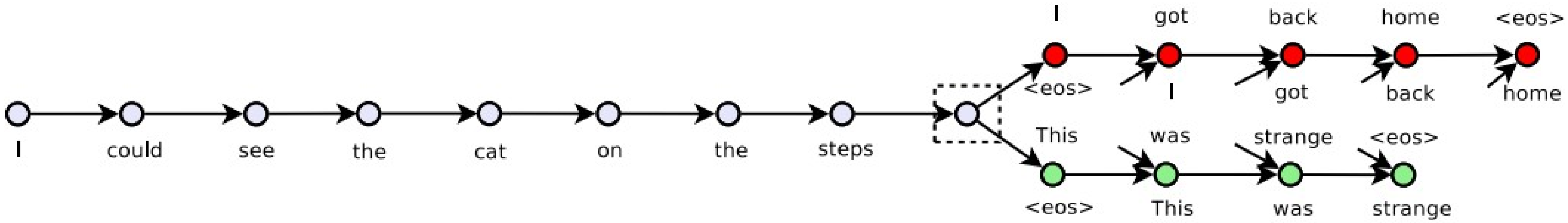
Word Mover's Distance (Kusner et al., 2015)

- A pairwise document similarity metric.
- Compares two sets of embeddings with weights (frequencies, VSM).
- Earth Mover's Distance



Skip-thoughts (Kiros et al., 2015)

- **Exploits sentence adjacency to train sentence embeddings.**
- **Encoder-decoder RNN architecture**
  - Breakthrough in machine translation



Sent2Vec (Pagliardini, Gupta, and Jaggi, 2017)

- **Shallow sentence embedding model**
  - Heavily based on Wor2Vec CBow
  - The window is a full semantic unit (sentence, paragraph, document…) instead of a few consequtive words words.

Results and conclusions

Best results of each algorithm

STS eval	Baseline	Doc2Vec	Doc2VecC	WMD	Sent2Vec
Titles	0.91 (EMB)	0.65 (1M)	0.87 (1M)	0.90	**0.91 (1M)**
Abstracts	0.93 (both)	0.86 (1M)	**0.92 (50K)**	**0.92**	0.87 (100K)
Bodies	0.96 (VSM)	**0.97 (500K)**	0.94 (10K)	-	0.83 (10K)

Summary

- **Accomplishments**
  - Thorough literature review of state-of-the-art
  - Analysed 10 algorithms:
    - Intuituion
    - The maths
    - Computational complesity
    - Empirical study
      - Computational benchmark
      - Evaluation

Conclusions

- **Word embeddings**
  - Very active field since Word2Vec
  - Most algorithms are derivative of Word2Vec, no clear advantages on evaluation.
  - Some breakthoughs: FastText.
- **Semantic Textual Similarity**
  - Active but early days.
  - Most models barely match naive baselines.
  - A lot of innovation and exploration, may lead to a breakthrough in a few years.

Future work

- **Main barrier:** lack of official datasets in the scientific domain.
  - Human scored similarity pairs in scientific domain.
  - Stronger article linkage
  - Training set for document similarity
- **SCITODATE R&D roadmap:**
  - NER for linking to BioPortal
  - Vocabulary mining
  - Fact and relationship mining
  - Named Entity prediction