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 endeavour 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).

![Diagram of Word2Vec Window Scanning Method](image)
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-occurrence matrix.
FastText (Bojanowski et al., 2016)

- **Word2Vec with subword components.**
  - Modular word embeddings.
  - N-gram embeddings.
  - Composition 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

<table>
<thead>
<tr>
<th>Word embeddings accuracy</th>
<th>1M</th>
<th>10M</th>
<th>100M</th>
<th>1B</th>
<th>2B</th>
</tr>
</thead>
<tbody>
<tr>
<td>W2V CBow - Total</td>
<td>0.03</td>
<td>0.17</td>
<td>0.46</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td>W2V Skip-gram - Total</td>
<td>0.04</td>
<td>0.18</td>
<td>0.46</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td>W2V CBow - Known</td>
<td>0.67</td>
<td>0.73</td>
<td>0.80</td>
<td>0.85</td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td>W2V Skip-gram - Known</td>
<td>0.67</td>
<td>0.79</td>
<td>0.80</td>
<td>0.88</td>
<td><strong>0.90</strong></td>
</tr>
<tr>
<td>GloVe - Total</td>
<td>0.04</td>
<td>0.17</td>
<td>0.45</td>
<td>0.80</td>
<td>0.87</td>
</tr>
<tr>
<td>GloVe - Known</td>
<td>0.71</td>
<td>0.73</td>
<td>0.78</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>FastText - Total</td>
<td><strong>0.81</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.93</strong></td>
<td>-</td>
</tr>
<tr>
<td>WordRank - Total</td>
<td>0.02</td>
<td>0.21</td>
<td>0.45</td>
<td>0.78</td>
<td>0.89</td>
</tr>
<tr>
<td>Wordrank - Known</td>
<td>0.69</td>
<td>0.75</td>
<td>0.77</td>
<td>0.84</td>
<td><strong>0.90</strong></td>
</tr>
</tbody>
</table>
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.
**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
- Shallow sentence embedding model
  - Heavily based on Wor2Vec CBow
  - The window is a full semantic unit (sentence, paragraph, document...) instead of a few consecutive words.
# Results and conclusions

Best results of each algorithm

<table>
<thead>
<tr>
<th>STS eval</th>
<th>Baseline</th>
<th>Doc2Vec</th>
<th>Doc2VecC</th>
<th>WMD</th>
<th>Sent2Vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles</td>
<td>0.91 (EMB)</td>
<td>0.65 (1M)</td>
<td>0.87 (1M)</td>
<td>0.90</td>
<td>0.91 (1M)</td>
</tr>
<tr>
<td>Abstracts</td>
<td>0.93 (both)</td>
<td>0.86 (1M)</td>
<td><strong>0.92 (50K)</strong></td>
<td>0.92</td>
<td>0.87 (100K)</td>
</tr>
<tr>
<td>Bodies</td>
<td>0.96 (VSM)</td>
<td><strong>0.97 (500K)</strong></td>
<td>0.94 (10K)</td>
<td>-</td>
<td>0.83 (10K)</td>
</tr>
</tbody>
</table>
Summary

• **Accomplishments**
  - Thorough literature review of state-of-the-art
  - Analysed 10 algorithms:
    • Intuition
    • The maths
    • Computational complexity
    • 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 breakthroughs: 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