Robust Word Vectors for Russian Language

V. Malykh[1]

[1]Laboratory of Neural Systems and Deep learning,
Moscow Institute of Physics and Technology (State University)
http://www.mipt.ru/

Artificial Intelligence
and Natural Language Conference, 2016

Outline

Word Vectors

- Vector representations of words is a basic idea to enable computers to work with words in more convenient manner than with simple categorical features.

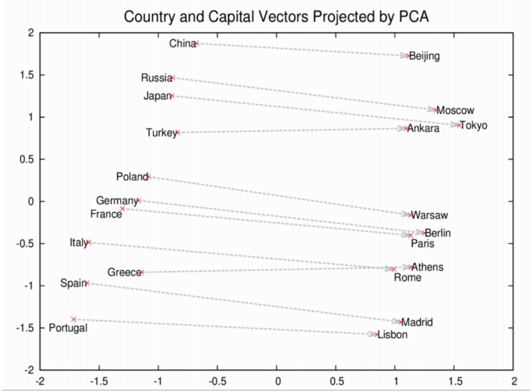


Figure is adopted from `deeplearning4j.com`

Existing approaches

Two main word-level approaches:

- Local context (e.g. Word2Vec, [**Mikolov2013**])
- Co-occurence matrix decomposition (e.g. GloVe, [**Pennington2014**])

Drawbacks:

- A model need to recognize the word exactly.
- Out of vocabulary words.

Existing approaches (2)

Char-level approach:

- Read the letters and try to predict the word, which it represents. E.g. [**Pennington2015**]

Drawback:

- Again out of vocabulary words.

Outline

Our Approach Description

- In our architecture we do not use any co-occurrence matrices.

Our Approach Description

- In our architecture we do not use any co-occurrence matrices.
- There is no entity of vocabulary in the model.

Our Approach Description

- In our architecture we do not use any co-occurrence matrices.
- There is no entity of vocabulary in the model.
- Context is handled by memorising in weights of a neural net.

Our Approach Description

- In our architecture we do not use any co-occurrence matrices.
- There is no entity of vocabulary in the model.
- Context is handled by memorising in weights of a neural net.
- We use recurrent layers to achieve this property.

Outline

1. Word Vectors
2. Our Approach
   - Our Approach Description
   - LSTM
   - BME Representation
   - Architecture
3. Experiments
   - 1st Corpus
   - 2nd Corpus

Long Short-Term Memory

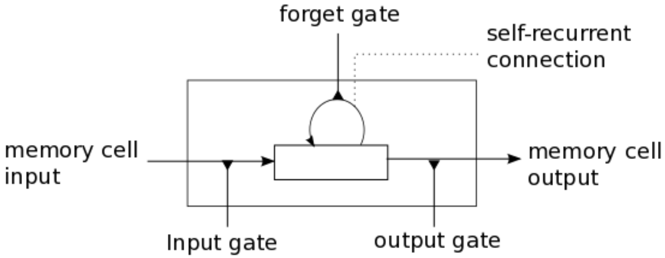


Figure is adopted from `http://deeplearning.net/tutorial/lstm.html`

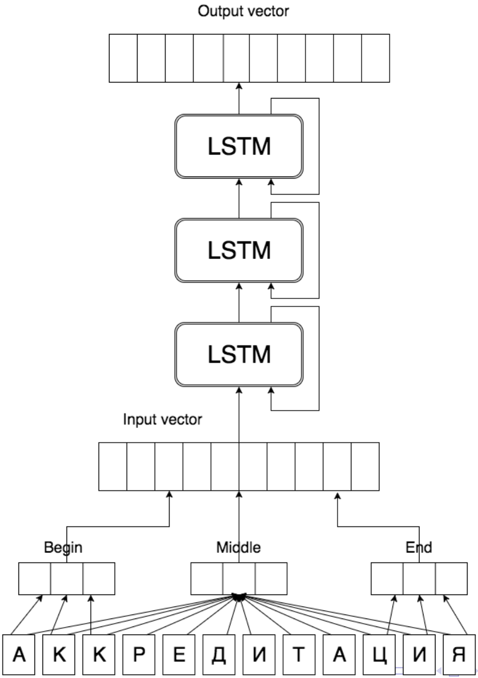
Outline

BME Representation

- **B** - begin, first 3 letters in one-hot form.
- **M** - middle, all letters in alphabet counters form.
- **E** - end, last 3 letters in one-hot form.

Outline

1. Word Vectors
2. Our Approach
   - Our Approach Description
   - LSTM
   - BME Representation
   - Architecture
3. Experiments
   - 1st Corpus
   - 2nd Corpus


Output vector
LSTM
LSTM
LSTM
Input vector
Begin
Middle
End
А К К Р Е Д И Т А Ц И Я

Math description

Negative Contrast Estimation loss:

$$NCE = e^{-s(v,c)} + e^{s(v,c')} \tag{1}$$

where v - a word vector, c - word vector from the word context, c' - word vector outside of the word context,
and $s(x, y)$ is some scoring function. We are using **cosine similarity** as scoring function:

$$cos(x, y) = \frac{x \cdot y}{|x||y|} \tag{2}$$

Outline

1. Word Vectors
2. Our Approach
   - Our Approach Description
   - LSTM
   - BME Representation
   - Architecture
3. Experiments
   - 1st Corpus
   - 2nd Corpus

Corpus Description

- News headings corpus in Russian

Corpus Description

- News headings corpus in Russian
- 3 classes: strong paraphrase, weak paraphrase, and non-paraphrase

Corpus Description

- News headings corpus in Russian
- 3 classes: strong paraphrase, weak paraphrase, and non-paraphrase
- Firstly introduced in 2015 in [**Pronoza2015**], in 2016 extended version.

Corpus Description
Statistics

Pairs[1]	7227
Strong Paraphrase Pairs	1668
Weak Paraphrase Pairs	2957
Non Paraphrase Pairs	2582

[1] Remark: in our experiments we use only 1 & -1 classes.

Experiment setup

- The metric is **ROC AUC** on **cosine similarity** interpreted as probability of the positive class (strong paraphrase).

Experiment setup

- The metric is **ROC AUC** on **cosine similarity** interpreted as probability of the positive class (strong paraphrase).
- We're adding artificial noise - additional & vanishing letters, replacement of letters.

Experiment setup

- The metric is **ROC AUC** on **cosine similarity** interpreted as probability of the positive class (strong paraphrase).
- We're adding artificial noise - additional & vanishing letters, replacement of letters.
- 10 runs with each noise level.

Results

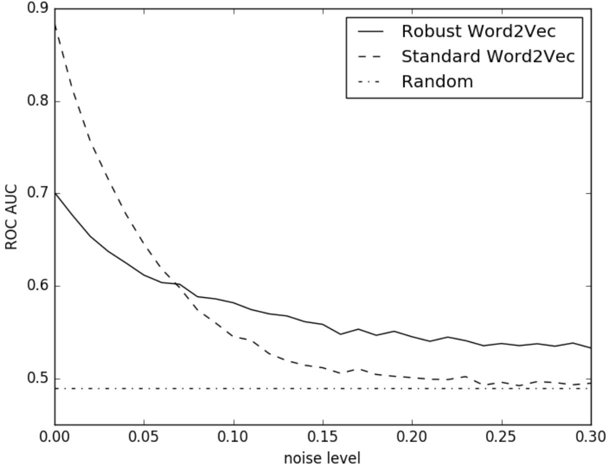


Figure: Results on Paraphraser corpus

Outline

Corpus Description

- Plagiarism detection in scientific papers.

Corpus Description

- Plagiarism detection in scientific papers.
- 150 pairs of articles' titles & descriptions in Russian.

Corpus Description

- Plagiarism detection in scientific papers.
- 150 pairs of articles' titles & descriptions in Russian.
- 3 human experts should produce their evaluation in $[0, 1]$.

Corpus Description

- Plagiarism detection in scientific papers.
- 150 pairs of articles' titles & descriptions in Russian.
- 3 human experts should produce their evaluation in $[0, 1]$.
- Was introduced in 2014 in work [**Derbenev2014**].

Results (2)

Table: Results of testing on scientific plagiarism corpus

System	Quality
Random Baseline	0.213 ± 0.025
Word2Vec Baseline	0.189
Robust Word2Vec	0.232

Summary

- We have introduced an architecture to produce word vectors, basing on characters.
- It does not store explicitly word vectors, so it has only fixed weights number, does not depending on the vocabulary size.
- The architecture does not rely on any type of pre-processing (i.e. stemming).
- The architecture outperforming the existing word vectors models in noisy environment.

Future Work

- We should find more corpora for paraphrase, maybe naturally noisy (e.g. Twitter Paraphrase Corpus for English).
- Try the architecture on other languages.
- Try to improve the quality on low noise regions, by the means of more deep architecture, attention, etc.

Thank you for your attention!
I would be happy to answer your questions.

References I

N. V. Derbenev, D. A. Kozliuk, V. V. Nikitin, V. O. Tolcheev
Experimental Research of Near-Duplicate Detection Methods for Scientific Papers.
Machine Learning and Data Analysis. Vol. 1 (7), 2014 (in Russian).

E. Pronoza, E. Yagunova, A. Pronoza
Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction.
Proceedings of the 9th Russian Summer School in Information Retrieval, August 24?28, 2015, Saint-Petersburg, Russia, (RuSSIR 2015, Young Scientist Conference), Springer CCIS

T. Mikolov et al.
Distributed representations of words and phrases and their compositionality.
Advances in neural information processing systems. 2013.

References II

W. Ling et al.
Finding function in form: Compositional character models for open vocabulary word representation.
In Proc. of EMNLP2015

J. Pennington, R. Socher, and C.D. Manning.
Glove: Global Vectors for Word Representation.
EMNLP. Vol. 14. 2014.