Evaluation of Croatian Word Embeddings

Lukáš Svoboda, Slobodan Beliga
Department of Computer Science and Engineering, Faculty of Applied Sciences, University of West Bohemia, Affiliation2, Affiliation3
Univerzitní 8, 306 14 Plzeň, Czech Republic, Address2, Address3
svobikl@kiv.zcu.cz, sbeliga@uniri.hr

Abstract

Many unsupervised learning techniques have been investigated to obtain useful word embedding representation. Research is focusing mostly on English and less on highly inflected languages from Slavic family.
We derived new corpus from the original *Word2vec* and added some of the specific linguistic aspect from Croatian language. We compared two popular word representation models, *Word2Vec* and *Glove*. Models has been trained on a new robust Croatian analogy corpus. We also translated WordSim353 and RG64 corpuses to Croatian and made basic semantic measurements.
Results show that models are able to create meaningful word representation. However, this research has shown that free word order and the higher morphological complexity of Croatian language influences the quality of resulting word embeddings.

1. Introduction

Word representation based on distributional semantics (**?**), commonly referred to as Word Embeddings, represent words as vectors of real numbers from low-dimensional space. The goal of such representation is to capture syntactic and semantic relationship between words.
It was shown that the word vectors can be used for significant improving and simplifying of many NLP applications (**?**; **?**). There are also NLP applications, where Word Embeddings does not help much (**?**).
Most of work is focused on English. Recently the community has realized that the research should focus on other languages with rich morphology and different syntax (**?**; **?**), but there is still little attention to highly inflected languages from Slavic family. These languages are highly inflected and have a relatively free word order.
In this paper, we focus on Croatian word embeddings. To be able to compare different word embeddings methods, we created two dataset based on original WordSim353(**?**) and RG64(**?**) translated to Croatian. Except the similarity between words, we would like to explore other semantic and syntactic properties hidden in word embeddings. A new evaluation scheme based on word analogies were presented in (**?**). Based on this popular evaluation scheme, we have produced a Croatian version of original Word2Vec analogy corpus in order to qualitatively compare the performance of different models.

2. Related Work

Nowadays, word embeddings are typically obtained as a product of training neural network-based language models. Language modeling is a classical NLP task of predicting the probability distribution over the "next" word. In these models a word embedding is a vector in $\mathbb{R}^n$, with the value of each dimension being a feature that weights the relation of the word with a "latent" aspect of the language. These features are jointly learned from plain unannotated text data. This principle is known as the *distributional hypothesis*. The direct implication of this hypothesis is that the word meaning is related to the context where it usually occurs and thus it is possible to compare the meanings of two words by statistical comparisons of their contexts. This implication was confirmed by empirical tests carried out on human groups in (**?**; **?**).
There is a variety of datasets for measuring semantic relatedness between English words, such as *WordSimilarity-353* (**?**), *Rubenstein and Goodenough (RG)* (**?**), *Rare-words* (**?**), *Word pair similarity in context* (**?**), and many others. Evaluation scheme based on word analogies were presented in (**?**).
To the best of our knowledge, only small portion of recent studies attempted evaluating Croatian word embeddings. In (**?**) authors translated a few questions from English analogy corpus to Croatian to be able to evaluate their Neural based model. However this translation was only made for a total of 350 questions. They used it only for their own simple tests and also did not publish such a small corpus. There is only one analogy corpus representing Slavic family language - Czech word analogy corpus presented in (**?**).
Many methods have been proposed to learn such word vector representations. One of the Neural Network based models for word vector representation which outperforms previous methods on word similarity tasks was introduced in (**?**). Word Embeddings methods implemented in tool *Word2Vec* (**?**) and GloVe (**?**) significantly outperform other methods for Word Embeddings. Word vector representations made by these methods have been successfully adapted on variety of core NLP tasks. Recent library *FastText* (**?**) tool is derived from Word2Vec and enriches word embeddings vectors with subword information.

3. Models

We experimented with state-of-the-art models used for generating word embeddings. Neural network based models CBOW and Skipgram from Word2Vec (**?**) tool and model GloVe that focuses more on the global statistics of the trained data. We have also tested the most recent Fast-Text tool that promises better score for morphologically rich languages.

3.1. CBOW

CBOW (Continuous Bag-of-Words) (**?**) tries to predict the current word according to the small context window around the word. The architecture is similar to the feed-forward

NNLP (Neural Network Language Model) which has been proposed in (**?**). The NNLM is computationally expensive between the projection and the hidden layer. Thus, CBOW proposed architecture, where the (non-linear) hidden layer is removed and projection layer is shared between all words. The word order in the context does not influence the projection (see Figure **??**). This architecture also proved low computational complexity.

3.2. Skip-gram

Skip-gram architecture is similar to CBOW. Although instead of predicting the current word based on the context, it tries to predict a words context based on the word itself (**?**). Thus, intention of the Skip-gram model is to find word patterns that are useful for predicting the surrounding words within a certain range in a sentence (see Figure **??**). Skip-gram model estimates the syntactic properties of words slightly worse than the CBOW model, but it is much better for modeling the word semantics on English test set (**?**) (**?**). Training of the Skipgram model does not involve dense matrix multiplications **??** and that makes training also extremely efficient (**?**).

3.3. GloVe

GloVe (Global Vectors) (**?**) model focuses more on the global statistics of the trained data. The main concept of this model is the observation that ratios of word-word co-occurrence probabilities have the potential for encoding meaning of words. This approach sequentially analyzes word contexts iterating on word windows across the corpus. The authors define $P_{ij} = p(j|i)$ as the probability that the word w_j appears in the context of word w_i. The authors

3.4. Fast-Text

FastText(**?**) combines concepts of CBOW (resp. Skip-Gram) architectures introduced earlier in Section 3.1. and 3.2.. These include representing sentences with bag of words and bag of n-grams, as well as using subword information, and sharing information across classes through a hidden representation.

3.5. Training data

We trained our models on two datasets in the Croatian language. We made the entire dump of Croatian Wikipedia (dated 08-2017) and ...
Fo One Billion Word Benchmark (**?**)
Number of documents,

	Vocabulary	Vocabulary tf¿5	Tokens
EN corpus	793,471		829,250,940
CR WIKI	793,471		829,250,940

Table 1: The caption of the table

3.6. Parameters

4. Corpus

Original Word2Vec analogy corpus is composed by 19,558 questions divided in two tested group : semantic and syntactic questions, e.g. king : man = woman : queen. Fourth word in question is typically predicted one).

Level	Tools
Morphology	Pitrat Analyser
Syntax	LFG Analyser (C-Structure)
Semantics	LFG F-Structures + Sowa's Conceptual Graphs

Table 2: The caption of the table

5. Experiments

Level	Tools
Morphology	Pitrat Analyser
Syntax	LFG Analyser (C-Structure)
Semantics	LFG F-Structures + Sowa's Conceptual Graphs

Table 3: Syntactic questions

Level	Tools
Morphology	Pitrat Analyser
Syntax	LFG Analyser (C-Structure)
Semantics	LFG F-Structures + Sowa's Conceptual Graphs

Table 4: Semantic questions

Level	Tools
Morphology	Pitrat Analyser
Syntax	LFG Analyser (C-Structure)
Semantics	LFG F-Structures + Sowa's Conceptual Graphs

Table 5: Total score

6. Conclusion

7. Page Numbering

Please do not include page numbers in your article. The definitive page numbering of articles published in the proceedings will be decided by the organising committee.

8. Headings / Level 1 Headings

Headings should be capitalised in the same way as the main title, and centred within the column. The font used is Times New Roman 12 bold. There should also be a space of 12 pt between the title and the preceding section, and a space of 3 pt between the title and the text following it.

8.1. Level 2 Headings

The format for level 2 headings is the same as for level 1 Headings, with the font Times New Roman 11, and the heading is justified to the left of the column. There should also be a space of 6 pt between the title and the preceding section, and a space of 3 pt between the title and the text following it.

8.1.1. Level 3 Headings

The format for level 3 headings is the same as for level 2 headings, except that the font is Times New Roman 10, and there should be no space left between the heading and the text. There should also be a space of 6 pt between the title and the preceding section, and a space of 3 pt between the title and the text following it.

9. Citing References in the Text

9.1. Bibliographical References

All bibliographical references within the text should be put in between parentheses with the author's surname followed by a comma before the date of publication (**?**). If the sentence already includes the author's name, then it is only necessary to put the date in parentheses: **?**). When several authors are cited, those references should be separated with a semicolon: (**?**; **?**). When the reference has more than three authors, only cite the name of the first author followed by "et al." (e.g. (**?**)).

9.2. Language Resource References

9.2.1. When Citing Language Resources

When citing language resources, we recommend to proceed in the same way to bibliographical references, except that, in order to make them appear in a separate section, you need to use the
`citelanguageresource` tag. Thus, a language resource should be cited as (**?**).

9.2.2. When Not Citing Any Language Resource

When no language resource needs to be cited in the paper, you need to comment out a few lines in the `.tex` file:

```
% \usepackage{multibib}
% \newcites{languageresource}{}
% \section{Language Resource References}
% \bibliographystylelanguageresource
%   {lrec}
% \bibliographylanguageresource{xample}
```

10. Figures & Tables

10.1. Figures

All figures should be centred and clearly distinguishable. They should never be drawn by hand, and the lines must be very dark in order to ensure a high-quality printed version. Figures should be numbered in the text, and have a caption in Times New Roman 10 pt underneath. A space must be left between each figure and its respective caption.
Example of a figure enclosed in a box:
Figure and caption should always appear together on the same page. Large figures can be centred, using a full page.

Figure 1: The caption of the figure.

10.2. Tables

The instructions for tables are the same as for figures.

11. Footnotes

Footnotes are indicated within the text by a number in superscript[1].

12. Copyrights

The Language Resource and Evaluation Conference (LREC) proceedings are published by the European Language Resources Association (ELRA). They are available online from the conference website.
ELRA's policy is to acquire copyright for all LREC contributions. In assigning your copyright, you are not forfeiting your right to use your contribution elsewhere. This you may do without seeking permission and is subject only to normal acknowledgement to the LREC proceedings. The LREC 2018 Proceedings are licensed under CC-BY-NC, the Creative Commons Attribution-Non-Commercial 4.0 International License.

13. Conclusion

Your submission of a finalised contribution for inclusion in the LREC proceedings automatically assigns the above-mentioned copyright to ELRA.

14. Acknowledgements

Place all acknowledgements (including those concerning research grants and funding) in a separate section at the end of the article.

15. Providing References

15.1. Bibliographical References

Bibliographical references should be listed in alphabetical order at the end of the article. The title of the section, "Bibliographical References", should be a level 1 heading. The first line of each bibliographical reference should be justified to the left of the column, and the rest of the entry should be indented by 0.35 cm.

[1]Footnotes should be in Times New Roman 9 pt, and appear at the bottom of the same page as their corresponding number. Footnotes should also be separated from the rest of the text by a horizontal line 5 cm long.

The examples provided in Section 16 (some of which are fictitious references) illustrate the basic format required for articles in conference proceedings, books, journal articles, PhD theses, and chapters of books.

15.2. Language Resource References

Language resource references should be listed in alphabetical order at the end of the article, in the **Language Resource References** section, placed after the **Bibliographical References** section. The title of the "Language Resource References" section, should be a level 1 heading. The first line of each language resource reference should be justified to the left of the column, and the rest of the entry should be indented by 0.35 cm. The example in Section 17 illustrates the basic format required for language resources. In order to be able to cite a language resource, it must be added to the `.bib` file first, as a `@LanguageResource` item type, which contains the following fields:

- `author`: the builder of the resource
- `title`: the name of the resource
- `publisher`: the publisher of the resource (project, organisation etc)
- `year`: year of the resource release
- `series`: more general resource set this language resource belongs to
- `edition`: version of the resource
- `islrn`: the International Standard Language Resource Number (ISLRN) of the resource[2]

If you want the full resource author name to appear in the citation, the language resource author name should be protected by enclosing it between {...}, as shown in the model `.bib` file.

Appendix: How to Produce the `.pdf` Version

In order to generate a PDF file out of the LaTeX file herein, when citing language resources, the following steps need to be performed:

- Compile the `.tex` file once
- Invoke `bibtex` on the eponymous `.aux` file
- Invoke `bibtex` on the `languageresources.aux` file
- Compile the `.tex` file twice

16. Bibliographical References

17. Language Resource References

[2]The ISLRN number is available from `http://islrn.org`