Evaluation of Croatian Word Embeddings

Lukáš Svoboda, Slobodan Beliga
Department of Computer Science and Engineering, University of West Bohemia
Department of Informatics, University of Rijeka
Univerzitní 22, 306 14 Plzeň, Czech Republic
Radmile Matejči? 2, 51000 Rijeka, Croatia
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 RG65 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 tasks, 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 RG65(**?**) 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. 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.

3.4. Training data

We trained our models on two datasets in the Croatian language. We made the entire dump of Croatian Wikipedia - dated 08-2017 with approximately 275,000 articles. We have tokenized the text, removed nonalphanumeric tokens and extracted only sentences with at least 5 tokens. Resulting corpus has xy tokens. We merged data from Wikipedia with Croatian corpus presented in (**?**) that has 1.2B tokens. Resulting corpus has xy tokens and xy sentences. Such corpus has vocabulary of xy words with at least 10 occurences.
For English version of data, we used Wikipedia dump from June 2016. This dump was made of 5,164,793 articles, has xy tokens and vocabulary of xy words.
We tested analogies and similarity corpuses for both languages with most frequent 300,000 words.

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

Table 1: Properties of Croatian training data corpus.

3.5. 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 the predicted one).
Our Croatian analogy corpus has 115,085 question divided in the same manner as for English into two tested group: semantic and syntactic questions.

Semantic questions are divided into 9 group, each having around 30 - 100 word question pairs. Combination of question pairs gives overall 36,880 semantics questions:

- capital-common-countries
- chemical-elements
- city-state
- city-state-USA
- country-world
- currency-shortcut
- currency
- eu-cities-states
- family

Syntactic part of corpus is divided into 14 group, consisting of 78,205 questions:

- jobs
- adjective-to-adverb
- opposite
- comparative
- superlative
- nationality-man
- nationality-female
- past-tense
- plural
- nouns-antonyms
- adjectives-antonyms
- verbs-antonyms
- verbs-pastToFemale
- verbs-pastToMale

Model	CBOW	Skip	fastText-Skip	fastText-CBOW
Capital	44.17	62.5	59.58	21.25
Chemical-elements	1.02	2.25	0.74	0.41
City-state	22.11	37.89	47.63	46.32
City-state-USA	5.78	8.23	4.30	0.37
Country-world	23.93	44.49	40.15	7.31
Currency	4.68	8.19	6.43	0.58
Currency-shortcut	2.08	8.19	2.50	0.42
EU-cities-states	21.59	41.95	42.33	6.16
Family	34.83	41.82	42.72	34.76
Jobs	68.94	64.06	88.54	95.45
Adj-to-adverb	18.36	21.36	35.33	62.01
Opposite	17.34	18.05	59.03	86.10
Comparative	34.90	33.57	43.22	41.46
Superlative	33.22	27.70	40.50	51.77
Nationality-man	17.01	23.87	60.05	62.13
Nationality-female	14.38	55.66	57.77	53.98
Past-tense	67.31	61.03	66.67	78.21
Plural	37.12	44.65	44.24	35.10
Nouns-ant.	12.70	10.96	10.80	21.24
Adjectives-ant.	13.39	13.11	18.59	12.59
Verbs-antonyms	9.18	6.18	7.25	9.71
Verbs-pastFemale	60.92	19.47	71.04	80.50
Verbs-pastMale	66.68	62.89	76.04	85.04
SEMANTICS	16.60	28.54	25.94	7.76
SYNTACTIC	37.06	35.63	49.60	54.56
ALL	32.03	33.89	43.83	43.13

Table 2: The caption of the table

Models	English		
	WordSim353	RG65	EN-analogies
CBOW	57.94	68.69	69.98 (44.02)
Skip-gram	64.73	78.27	73.57 (46.28)
fastText-Skip	46.13	76.31	68.27 (42.94)
fastText-CBOW	44.64	73.64	76.58 (48.17)
	Croatian		
CBOW	xy	xy	32.03 (19.19)
Skip-gram	xy	xy	33.89 (20.31)
fastText-Skip	xy	xy	43.83 (25.79)
fastText-CBOW	xy	xy	43.14 (25.79)

Table 3: Comparison with English models. Measurement in brackets gives the results including OOV 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

5. Experiments

In total we tested on 68,986 out of 115,085 questions, it means that almost 40% question was unknown by model. All unknown questions were discarded from testing process. We tested Semantic group on 16,968 known questions and part of corpus testing Syntactic properties was measured on 52,018 questions.

Only 10 out of 353 question was unknown for WordSim353 corpus and all 65 questions of RG65 corpus were in vocabulary. Unknown words in WordSim353 we have represented as a vector averaged from 10 least common words in vocabulary.

6. Conclusion

`https://github.com/Svobikl/cr-analogy`

7. Acknowledgements

This work was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

8. Bibliographical References