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 *fastText*. 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 (Harris, 1954), commonly referred to as Word Embeddings, represent words as vectors of real numbers from low-dimensional space. The goal of such representations is to capture syntactic and semantic relationship between words. It was shown that the word vectors can be used for significant improving and simplifying 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 (Berardi et al., 2015; Elrazzaz et al., 2017; Köper et al., 2015; Svoboda and Brychcín, 2016), but there is still a little attention to 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(Finkelstein et al., 2002) and RG65(Rubenstein and Goodenough, 1965) 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 (Mikolov et al., 2013a). 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 (Rubenstein and Goodenough, 1965; Charles, 2000).

There is a variety of datasets for measuring semantic relatedness between English words, such as *WordSimilarity-353* (Finkelstein et al., 2002), *Rubenstein and Goodenough (RG)* (Rubenstein and Goodenough, 1965), *Rare-words* (**?**), *Word pair similarity in context* (**?**), and many others.(**?**) reported that a predict vector space trained with a simplified neural language model (Bengio et al., 2006) encodes syntactic and semantic properties of language, which can be recovered directly from space through linear translations, to solve analogies such as: $\vec{king} - \vec{man} = \vec{queen} - \vec{woman}$. Evaluation scheme based on word analogies were presented in (Mikolov et al., 2013a).

To the best of our knowledge, only small portion of recent studies attempted evaluating Croatian word embeddings. In (Zuanovic et al., 2014) 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 (Svoboda and Brychcín, 2016).

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 (Huang et al., 2012). Word Embeddings methods implemented in tool *Word2Vec* (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) 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 (Bojanowski et al., 2016) 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 (Mikolov et al., 2013a) tool and tool FastText that promises better score for morphologically rich languages.

3.1. CBOW

CBOW (Continuous Bag-of-Words) (Mikolov et al., 2013a) 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 (Bengio et al., 2006). 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. 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 (Mikolov et al., 2013b). 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. 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 (Mikolov et al., 2013a) (Mikolov et al., 2013b). Training of the Skipgram model does not involve dense matrix multiplications and that makes training also extremely efficient (Mikolov et al., 2013b).

3.3. Fast-Text

FastText(Bojanowski et al., 2016) 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 (Šnajder et al., 2013) 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 categories, each having around 20 - 100 word question pairs. Combination of question pairs gives overall 36,880 semantics questions:

- `capital-common-countries`: This group consist of 23 the most common countries. These countries were adopted from original Word2Vec analogies and having highest number of occurrences in text between all languages.

- `chemical-elements`: Represents 119 pairs of chemical elements with their shortcut symbol (i.e. O - Oxygen).

- `city-state`: Gives 20 regions (states) inside the Croatia and gives one of city example in such region.

- `city-state-USA`: 67 pairs of cities and corresponding states in USA. This category is adopted from original English word analogy test.

- `country-world`: 118 pairs of countries with main cities from all over the world. Translated from original Word2Vec analogies.

- `currency-shortcut`: 20 pairs of state currencies with its shortcut name (i.e. Switzerland - CHF).

- `currency`: 20 pairs of states with their currencies (i.e. Japan - yen). Translated from original EN analogy corpus.

- `eu-cities-states`: 40 word pairs of states from EU and their corresponding main city (i.e. Belgium - Brussels)

- `family`: 41 word pairs with family relation in masculine vs feminine form (i.e. brother - sister)

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

- `jobs`: This category is language-specific, consist of 109 pairs of job positions in masculine× feminine form.

- `adjective-to-adverb`: 32 pairs of adjectives and its representatives in adverb form.

- `opposite`: 29 pairs of adjectives with its opposites. This category collects words from which is easy to make its opposites usually with preposition ”un” or ”in”, respective preposition ”ne” in Croatian (i.e. certain - uncertain). Adopted from original EN word analogies.

- `comparative`: 77 pairs of adjectives and its comparative form (i.e. good - better).

- `superlative`: 77 pairs of adjectives and its superlative form.

- `nationality-man`: 84 pairs of states and humans representing its nationalities in masculine form. (i.e. Switzerland - Swiss)

- `nationality-female`: 84 pairs of states and its nationalities in feminine form. This is language specific.

- `past-tense`: 40 pairs of verbs and its past tense form.

- `plural`: 46 pairs of nouns and its plural form.

- `nouns-antonyms`: 100 pairs of nouns and its antonyms.

- `adjectives-antonyms`: Similar category to *opposite*, it consists of 96 word pairs of adjectives and their antonyms. However, words are much more complex (i.e. good - bad).

- `verbs-antonyms`: 51 pairs of verbs and its antonyms.

- `verbs-pastToFemale`: 83 pairs of verbs and its past tense in feminine form. This category is extended from category *past-tense* and is language-specific.

- `verbs-pastToMale`: 83 pairs of verbs and its past tense masculine form. Category is same as past-tense, only its extended variation to be comparable with category *verbs-pastToFemale*.

For basic comparison with English, we have translated state-of-the-art English word similarity data test sets WordSim353 (Finkelstein et al., 2002) and RG65 (Rubenstein and Goodenough, 1965). These corpuses have 353 (respespective 65) word pairs. Each word pair is manually annotated with similarity. We kept similarities untouched. The words in WordSim are assessed on a scale from 0 to 10, in RG65 from 0 to 5.

Model	**CBOW**	**Skip-gram**	**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_EN	73.63	83.64	68.77	68.27
SYNTACTIC_EN	67.55	66.8	67.94	76.58
SEMANTICS_HR	16.60	28.54	25.94	7.76
SYNTACTIC_HR	37.06	35.63	49.60	54.56
ALL_HR	32.03	33.89	43.83	43.13

Table 2: Detailed results of Croatian word analogy corpus.

	English		
Models	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	37.61	52.01	32.03 (19.19)
Skip-gram	52.16	58.47	33.89 (20.31)
fastText-Skip	52.98	64.31	43.83 (25.79)
fastText-CBOW	30.41	51.06	43.14 (25.79)

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

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* were in vocabulary. Unknown words in *WordSim353* were represented as word vector averaged from 10 least common words in vocabulary.

Semantic tests gives overall poor performance on all tested models, as we can see in table 2, the opposite is true for English, where semantic tests gives usually similar score as syntactic tests. This behavior we already saw on Czech corpus presented in (Svoboda and Brychcín, 2016). It seems that free word order and other properties of highly inflected languages from Slavic family have a big impact on performance of current state-of-the-art word embeddings methods.

From results of *City-state* and *City-state-USA* category it can be seen that knowledge of topic in training data has significant impact on performance of a model. We wanted to show differences between two similar categories in case we have insufficient amount of training data covering particular topic. Category *City-state* is showing that model is able to carry such knowledge - if the topic is sufficiently represented in a training data, than model is able to carry this type of information. This behavior is seen in regions from Croatia mentioned in many articles on Croatian Wikipedia, but this was not a case with states from USA. All questions of *City-state* were covered, but only around 50% of questions in category *City-state-USA* were in vocabulary. On categories *Country-world* and *EU-cities-states* it can be seen that there is no difference between knowledge about states and main cities from EU again state-city pairs from all over the world. Another very poor performance gives group *Currency*, but this group is usually weak across all languages and shows the weaknesses of the model.

Syntactic tests gives better performance than semantic oriented tests, but they still have significantly worse performance rather than on English. This part of corpus includes language-specific group of tests - such as *Verbs-pastMale/Female*, *Nationality-man/female*. Simple *Past-tense* tests gives surprisingly high score - similarly it was also with Czech language in (Svoboda and Brychcín, 2016). We could say, that languages from Slavic family tends to have easier patterns for past tense. From language-specific groups we see that slightly better score is given in categories with word pairs in masculine form, these results also corresponds with the fact that there are more articles written in masculine form in the training data.

6. Conclusion

We have shown that similarly to Czech, another representative of Slavic language also Croatian (todo) `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

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer.

Berardi, G., Esuli, A., and Marcheggiani, D. (2015). Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, 21(04):505–524.

Elrazzaz, M., Elbassuoni, S., Shaban, K., and Helwe, C. (2017). Methodical evaluation of arabic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 454–458.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Köper, M., Scheible, C., and im Walde, S. S. (2015). Multilingual reliability and" semantic" structure of continuous word spaces. In *IWCS*, pages 40–45.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, October.

Šnajder, J., Padó, S., and Agić, Ž. (2013). Building and evaluating a distributional memory for croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 784–789.

Svoboda, L. and Brychcín, T. (2016). New word analogy corpus for exploring embeddings of czech words. *CoRR*, abs/1608.00789.

Zuanovic, L., Karan, M., and Šnajder, J. (2014). Experiments with neural word embeddings for croatian. In *Proceedings of the 9th Language Technologies Conference*, pages 69–72.