Evaluation of Croatian Word Embeddings

Svoboda, Lukáš; Beliga, Slobodan

Evaluation of Croatian Word Embeddings

Lukáš Svoboda[1], **Slobodan Beliga**[2]
1) Department of Computer Science and Engineering, University of West Bohemia
Univerzitní 22, 306 14 Plzeň, Czech Republic
2) Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
svobikl@kiv.zcu.cz[1], sbeliga@inf.uniri.hr[2]

Abstract

Croatian is poorly resourced and highly inflected language from Slavic language family. Nowadays, research is focusing mostly on English. We created a new word analogy dataset based on the original English *Word2vec* word analogy dataset and added some of the specific linguistic aspects from the Croatian language. Next, we created Croatian WordSim353 and RG65 datasets for a basic evaluation of word similarities. We compared created datasets on two popular word representation models, based on *Word2Vec* tool and *fastText* tool. Models have been trained on 1.37B tokens training data corpus and tested on a new robust Croatian word analogy dataset. Results show that models are able to create meaningful word representation. This research has shown that free word order and the higher morphological complexity of Croatian language influences the quality of resulting word embeddings.

Keywords: Croatian word embeddings, Croatian word analogy, Croatian language, Slavic language family, Word2Vec, FastText, Croatian word similarity dataset, WordSim353, RG65

1. Introduction

Word2Vec and FastText are tools that create models representing words as vectors of real numbers from high-dimensional space. Word representations are based on Distributional Hypothesis (Harris, 1954), where the context for each word is given by its nearby words. The goal of such representations is to capture the syntactic and semantic relationship between words.

It was shown that the word vectors can be successfully used in order to improve and/or simplify many NLP applications (Collobert and Weston, 2008; Collobert et al., 2011). There are also NLP tasks, where word embeddings does not help much (Andreas and Klein, 2014).

Most of the 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. Since there are open questions related to the embeddings in the Slavic language family, in this paper, we will focus mainly on Croatian word embeddings, from the South Slavic language family. With the aim of expanding existing findings about Croatian word embeddings, in this paper we will:

1. Compare different word embeddings methods on Croatian language which is not deeply explored, and according to its features, belongs to highly inflected language (words can have seven different cases for singular and seven for plural, genders, and numbers).

2. For the purposes of the word embeddings experiments, we will create three new datasets. Two basic word similarity datasets based on original WordSim353 (Finkelstein et al., 2002) and RG65 (Rubenstein and Goodenough, 1965) will be translated to the Croatian. Except for the similarity between words, we would like to explore other semantic and syntactic properties which are 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 will create a Croatian version of original *Word2Vec* analogy dataset in order to qualitatively compare the performance of different models.

3. Empirically compare the results obtained from the Croatian language, which belongs to the group of Balto-Slavic (subgroup: Slavic) languages, to the results obtained from the English – the most commonly studied language, which belongs to the group of Germanic language family (subgroup: West).

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* (Harris, 1954). 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 evaluating semantic relatedness between English words, such as *WordSimilarity-353* (Finkelstein et al., 2002), *Rubenstein and Goodenough (RG)* (Rubenstein and Goodenough, 1965), *Rare-words* (Luong et al., 2013), *Word pair similarity in con-*

text (Huang et al., 2012), and many others. Mikolov et al. reported in (Mikolov et al., 2013a) that word vectors 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 a review of works that evaluate syntactic and semantic analogies, we have encountered only a few datasets. In (Zuanović et al., 2014) authors translated small portion from English analogy dataset to Croatian in order to evaluate their Neural based model. However, this translation of syntactic analogy reasoning dataset was only made for a total of 350 questions based on positive-comparative form relationship in adjectives. In addition to syntactic, they also prepare semantic analogy reasoning dataset. It was based on countries and their capitals, originally proposed for the English by (Mikolov et al., 2013a) and translated into Croatian. The dataset comprises 506 entries. Recently Mrkšić et al. also trained word embeddings (Mrkšić et al., 2017), and produced a translation in the Croatian language and reannotation of gold standard resource SimLex-999 (Hill et al., 2015) with 999 word pairs.

There is only one analogy dataset representing Slavic language family – Czech word analogy dataset presented in (Svoboda and Brychcín, 2016).

In general, 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 a variety of core NLP tasks. Recent library *fastText* (Bojanowski et al., 2017) tool is derived from *Word2Vec* and enriches word embeddings vectors with subword information.

2. Proposed Datasets

Original *Word2Vec* analogy dataset is composed of 19,558 questions divided into two tested group: semantic and syntactic questions, e.g. king : man = queen : woman. The fourth word in question is typically the predicted one.

Our Croatian analogy dataset has 115,085 question divided in the same manner as for English into two tested group: semantic and syntactic questions. Dataset has been created by three annotators (two native speakers).

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 the highest number of occurrences in the text between all languages.
- `chemical-elements`: Represents 119 pairs of chemical elements with their shortcut symbol (e.g. O – Oxygen).
- `city-state`: Gives 20 regions (states) inside Croatia and gives one of city example in such region.
- `city-state-USA`: 67 pairs of cities and corresponding states in the 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 (e.g. Switzerland – CHF).
- `currency`: 20 pairs of states with their currencies (e.g. Japan – yen). Translated from original EN analogy dataset.
- `eu-cities-states`: 40 word pairs of states from EU and their corresponding main city (e.g. Belgium – Brussels).
- `family`: 41 word pairs with family relation in masculine vs. feminine form (e.g. brother – sister).

Syntactic part of the dataset 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 $\times$ 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", the respective preposition "ne" in Croatian (e.g. certain – uncertain). Adopted from original EN word analogies.
- `comparative`: 77 pairs of adjectives and its comparative form (e.g. 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. (e.g. 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 to *opposite* category, it consists of 96 word pairs of adjectives and their antonyms. However, words are much more complex (e.g. 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. This category is the same as *past-tense*, only its extended variation to be comparable with category *verbs-pastToFemale*.

2.1. Word Similarities Corpora

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

3. Distributional Semantic Models

We experimented with state-of-the-art models used for generating word embeddings. Neural network based models *CBOW* and *Skip-gram* 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 NNLM (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 an 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 word's context based on the word itself (Mikolov et al., 2013b). Thus, the 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 *Skip-gram* model does not involve dense matrix multiplications and that makes training also extremely efficient (Mikolov et al., 2013b).

3.3. FastText

FastText (Bojanowski et al., 2017) 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.

4. Experimental Results

4.1. Training Data

We trained our models on two datasets in the Croatian language. Initially, we made the entire dump of Croatian Wikipedia – dated from August 2017 with approximately 275,000 articles. We have tokenized the text, removed non-alphanumeric tokens and extracted only sentences with at least 5 tokens. Resulting corpus has 92,446,973 tokens. Secondly, we merged data from Wikipedia with Croatian corpus presented in (Šnajder et al., 2013), and originally proposed in (Ljubešić and Erjavec, 2011), that has over 1.2B tokens. Resulting corpus has 1.37 tokens and 56,623,398 sentences. Such corpus has a vocabulary of 955,905 words with at least 10 occurrences.
For the English version of data, we used Wikipedia dump from June 2016. This dump was made of 5,164,793 articles and has 2.2B tokens. We tested analogies and similarity corpora for both languages with most frequent 300,000 words.

	Vocabulary $tf > 10$	Tokens	
EN corpus	3,234,907	2,201,735,114	
HR corpus	955,905	1,370,836,176	

Table 1: Properties of Croatian (HR corpus) and English (EN corpus) training data.

In total, we tested models on 68,986 out of 115,085 questions. It means that almost 40% question was unknown by the model. All question contained OOV words were discarded from the testing process. We tested *semantic* group on 16,968 known questions and part of corpus testing *syntactic* properties were 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 a vocabulary.

Semantic tests reveal overall poor performance on all tested models, as we can see in Table 2. The opposite is true for English, where semantic tests give 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

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 dataset (the results of the semantic test at the top of the table; the results of the syntactic test in the middle part of the table; total results for English and Croatian at the bottom of the table).

	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.

impact on the 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 the topic in training data has the significant impact on the performance of a model. We wanted to show differences between two similar categories in case we have an insufficient amount of training data covering a 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, the 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 the 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 reveal better performance than tests oriented to semantic, 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 give surprisingly high score – similar findings were presented for the Czech language in (Svoboda and Brychcín, 2016). We could say, that languages from Slavic family tend 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 the masculine form, these results also correspond with the fact that there are more articles written in the masculine form in the training data.

4.2. Testing Data

As previously mentioned, in our experiments two word similarity datasets were used WordSim353 (Finkelstein et al., 2002) and RG65 (Rubenstein and Goodenough, 1965). In this subsection, we will discuss the issues we have encountered in the process of creating datasets by translating the original English versions into Croatian. The translation process can bring noise to data very insensibly due to difference and specificity contained in two languages (Croatian and English). Some specific cases are as follows:

- **mapping M:1 (the problem caused by lack of context)**: two (or more) words in the English language can have slightly different meanings in different contexts. In the translation process of such words to the Croatian language without taking into account the context in which the word appears, it is difficult to determine which translation is correct. To be quite clear, the context is missing because of the word-tuples form preserved in the dataset instead of the plain text form. For example, words *coast* and *shore* both appear in the dataset but in translation to the Croatian (without the context) both have one common meaning and unique translation *obala*. This problem could be entitled as 2:1 mapping (2 original words and 1 target). The problem of such mapping can be solved easily by using synonyms *priobalje* or *kopno*. The situation is unfavorable in case we use a synonym for translation as well as in the case we use the same Croatian word in the translation of two different English words. In both cases, we are not sure how much noise is introduced into the data.

- **the absence of a synonym pair**: this is a special case of the problem mentioned in the previous point as mapping M:1. In particular, two or more words have the same translation into the target language. However, the problem can not be solved by a synonym pair because it does not exist. For example, both *midday* and *noon* appear in our English version of the dataset,

but in the Croatian, they can be translated only as *podne*. However, *midday* could be translated differently (as *podnevni*), but in such translation, it becomes an adjective rather than a noun. Such a scenario certainly may allow noise introduction into the dataset, and therefore is not desirable.

- **mapping 1:M (the problem caused by lack of context)**: a problem is similar to the one mentioned above, with the difference in mapping between source and target words (1 original word from the English dataset can be translated into two or more Croatian words), and translation again depends on the context. For example, English word *drug* can be mapped into the Croatian words *lijek* (drug with the positive connotation; used for medical purposes, i.e. medicament) or *droga* (drug with the negative connotation; causes addiction, i.e. narcotic), depends again on missing context.

- **the problem of using multiple words in the translation**: for some English terms, there are no translations consist of just one word in Croatian, instead, two or more words must be used (phrase; set of words). For example, *seafood* can be translated with syntagma of two words *morski plodovi* or *plodovi mora*.

- **the problem caused by cultural differences**: In the geographical area of the Croatian, established word for *football* and *soccer* is *nogomet*, unlike the US area where there is a clear difference between these two words and sports. In the example of our dataset, both words *football* and *soccer* are present. Again, we can use a different Croatian translation *ragbi*, but we use the risk of introducing the noise into the dataset.

- **the problem of non-standard words (slang)**: some words in the slang may have different meanings than those in the standard language. For example, word *cock* can belong to the standard language but also to the slang, and depending on it has two meanings and two corresponding translations into the Croatian language: *pjetao* (kind of a bird) or *penis* (genitalia).

Besides, it is important to emphasize that although the invested efforts and high linguistic expertise, the created dataset may have unintentionally included noise into the data, which is inevitable due to restrictions in translation caused by specificities in different languages.

5. Conclusion

In this paper, evaluation of Croatian word embeddings is performed. New word embeddings are derived using different models. Additionally, some of the specific linguistic aspects of the Croatian language was added. Two popular word representation models were compared, *Word2Vec* and *fastText*. Models have been trained on a new robust Croatian analogy dataset. WordSim353 and RG65 datasets were translated from the English to the Croatian, in order to perform basic semantic measurements. Results show that models are able to create meaningful word representation.

However, it is important to note that this paper presents the first comparative study of word embeddings for Croatian and English, and therefore, new insights for NLP community according to the behavior of the Croatian word embeddings. The Croatian language belongs to the group of Slavic languages and has only preliminary and basic knowledge insights from word embeddings. In addition, another contribution of this work is certainly new datasets for the Croatian language, which are publicly available from: `https://github.com/Svobikl/cr-analogy`. It is also worth mentioning that these are also the first parallel English-Croatian word embeddings datasets.

Finally, we can figure out from experiments that models for Croatian does not achieve such good results as for English. In fact, results are mostly lower for the Croatian than for the English, with the exception of one case: fastText-Skip for WordSim-353. Such results can be explained theoretically through two perspectives – technical and linguistic.

1) Firstly, **the technical one** rests on the fact related to the corpus statistics used in the experiments (i.e. the size of the training corpus). It is evident that there is more English training data than Croatian. Therefore, expectations could be higher for English than for Croatian.

2) Secondly, **the linguistic one** withdraws its arguments from two sources.

   a) **Testing data** is the first one. In the process of translating datasets (testing data) from the English to the Croatian, there are possibilities of unintentional entering of the noise into the data (for example, by using synonyms) which hance make the task harder. Due to this fact it is reasonable to expect slightly worse results for Croatian.

   b) **Training data** is the second one. Croatian an English corpuses used for the training have serious differences in morphological complexity according to regularities of the Croatian and the English language. In particular, the difference in English (Germanic language) and Croatian (Slavic language) morphology is huge. Compared to the Croatian language, English language morphology is considerably poor. The Croatian language is a highly inflected language with mostly free word ordering in sentence structure, unlike the English, which is inflectional language and has a strict word ordering in a sentence (*subject-verb-object*). For example, three Croatian words are enough to construct two different sentence constructions with the same meaning: "*Ana voli Milovana.*" and "*Milovana voli Ana.*". Unlike the English, which requires up to 5 words for the same language construction: "*Ana loves Milovan.*" and "*Milovan is loved by Ana.*". These differences are reflected in the results of embeddings modeling. It is plausible that higher degree of inflection leads to higher data sparsity, which could reduce performance. Models presented in

this paper give good approximations to the English, they are better tailored to the English language morphology and better match the structure of such a language.

The most important conclusion of this research suggests that models for the Croatian do not achieve such good results as for the English. According to (Svoboda and Brychcín, 2016), this is also true for the Czech language, another one from Slavic language family. Following this, we would like to point out that future research should be focused on model improvements for Slavic languages. It would be worth to explore which Slavic languages specificities would be advisable to incorporate into models, in order to achieve better modeling of complex morphological structures. On the other hand, corpora preprocessing which simplifies morphological variations (and reduces data sparsity), such as stemming or lemmatization procedures, could also have an effect on word embeddings and should be one of the future research directions. Besides, we would also like to further investigate properties of other models for word embeddings and try to use external sources of information (such as part-of-speech tags, referenced information on Wikipedia, etc.) and experiment with the tree structure of sentence during the training process.

6. 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.

7. Bibliographical References

Andreas, J. and Klein, D. (2014). How much do word embeddings encode about syntax? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 822–827, Baltimore, Maryland, June. Association for Computational Linguistics.

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. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398.

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.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with genuine similarity estimation. *Comput. Linguist.*, 41(4):665–695, December.

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.

Ljubešić, N. and Erjavec, T. (2011). hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal et al., editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer.

Luong, M.-T., Socher, R., and Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. In *CoNLL*, Sofia, Bulgaria.

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.

Mrkšić, N., Vulić, I., Ó Séaghdha, D., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., and Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

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.

Zuanović, 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.