An evaluation of Czech word embeddings

Karolína Hořeňovská
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
horenovska@ufal.mff.cuni.cz

Abstract

We present an evaluation of Czech low-dimensional distributed word representations, also known as word embeddings. We describe five different approaches to training the models and three different corpora used in training. We evaluate the resulting models on five different datasets, report the results and provide their further analysis.

1 Introduction

Distributed word representations, often referred to as word embeddings, have received a lot of attention in recent years, and they have been used to improve results in many NLP tasks. The term itself refers to representing words as low-dimensional real-valued vectors (usually with dimensionality of 50-1000), and is opposed to explicit sparse representations, i.e. representing words as high-dimensional vectors of 0s and 1s (usually with dimensionality in the tens of thousands).

Many different models have been proposed (see section 2). By their nature, these models are language-independent (given the language can be tokenized) but usually the reported results are measured using only English. This is encouraged not only by English being the standard scientific language, but also by the availability of English text corpora and, even more importantly, English datasets to evaluate the models on.

We have decided to perform an intrinsic evaluation of embedding models on Czech. We have identified several successful models to evaluate, collected existing datasets to evaluate them on and designed two more datasets to extend the evaluation. We should note that we do not perform downstream-task evaluation, even though it might not correlate well with the intrinsic evaluation (Tsvetkov et al., 2015). We also use the models with their default parameters and only try changing the corpus they are trained on.

The rest of the paper is organized as follows: first, we describe related work (section 2). We continue with a description of selected models (section 3), corpora used in training (section 4) and the datasets (section 5). Finally, we present the results (section 6).

2 Related work

Related work could be clustered into three groups of papers.

First, we should mention papers performing evaluation of Czech word embeddings. Such evaluation exists for Word2Vec and GloVe using analogy corpus (Svoboda and Brychcín, 2016), however we are not aware of any more recent evaluation (which would cover also more recent models). Still, some papers evaluate some word embeddings in the context of a new dataset, as is the case of Czech similarity-relatedness dataset (Konopik et al., 2017).

Second, there are intrinsic evaluations of embeddings. These are usually part of new model proposals but there are exceptions. A notable one is a comparison by Baroni et al (2014), and also the work by Levy and Goldbert (2014), though this paper proposes another objective to solving analogies. Tsvetkov (2015) should also be mentioned for showing that intrinsic evaluation of embeddings need not correlate with performance on downstream tasks, and Nayak et al (2016) proposed a suite to test word embeddings.

Finally, there are model proposals. The most famous one is probably Word2Vec (Mikolov et al., 2013a), which has later been extended to fastText (Bojanowski et al., 2017). Despite being so well known, Word2Vec is neither the older (cf. e.g. the work by Schütze (1993)) nor the only one. Another famous models include GloVe (Pennington et al., 2014), LexVec (Salle et al., 2016b), ELMo

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 65–75
Turku, Finland, 30 September – 2 October,

(Peters et al., 2018) or a recent model BERT (Devlin et al., 2018).

In addition to models themselves, there are proposals on altering the trained model so that it better fits a purpose, e.g. by transforming the vector space to get vectors of synonyms closer to each other and increase the distance between antonym vectors (Faruqui et al., 2014; Mrkšić et al., 2016).

3 Selected models

In this section, we outline each of the selected models. We also report which implementation we use in our experiments.

Following some literature, we characterize each model as either *predictive* (trained by learning to predict a word) or *counting* (trained using co-occurences counts).

Unfortunately, it has not been feasible to train some model-corpus combinations.[1] We were not able to train fastText on Czech National Corpus using forms and LexVec on Czech National Corpus using either forms or lemmata. We have also not trained BERT on our own, instead, we only use a pre-trained model.

3.1 Word2Vec

Word2Vec (Mikolov et al., 2013a,b) is probably the most famous neural embedding model. The same name actually refers to two different architectures – called continuous bag of words (CBOW) and skip-gram (SG).

Both architectures are basically feed-forward networks. CBOW's inputs are words (tokens) within another word's context and its golden output is the surrounded word. Often, context window of size 5 is used, i.e. five words preceding and five words following the predicted word form the inputs. There's one projection layeyr between input and output layers. For skip-gram, it is the other way round, i.e. one word forms the input and words surrounding it are predicted. In both architectures, all words share the projection layer, which reduces the number of parameters to train, and thus the training time.

When using Word2Vec without specifying architecture, skip-gram is usually the default as it performs better in most evaluation. However, since Svoboda and Brychcín (2016) found out CBOW performed better in their experiments on Czech, we experiment with both architectures.

We use Word2Vec implementation provided in the gensim library (Řehůřek and Sojka, 2010).

An important concept introduced in the second paper (Mikolov et al., 2013b) is negative sampling: when training a word vector, other words are randomly sampled from the corpus and the model is penalized for high similarity of their vectors.

3.2 FastText

FastText (Bojanowski et al., 2017) is an extension of Word2Vec skip-gram which incorporates subword information in resulting vectors. Words are prefixed and suffixed with boundary symbols and vectors are then trained not only for all words but also for all n-grams appearing in any of the words. Boundary symbols are important to distinguish short words from n-grams appearing inside words. Using n-gram embeddings, even vectors for out-of-vocabulary words (i.e. words not present in training corpus) can be generated.

Please note that even though Bojanowski et al. (2017) describe the model as using skip-gram, it can integrate with CBOW architecture. We have tried using both architectures.

We use the implementation provided in gensim library (Řehůřek and Sojka, 2010).

3.3 GloVe

GloVe (Pennington et al., 2014) is a counting model which utilizes co-occurence matrix, i.e. numbers of times a word occurs within the context of another word. The basic idea is that if some words are related to the same concept, the probability of appearing in their context is much higher for these words than for any other word. This ratios need to be captured by the resulting model. The formulae to capture these similarities/ratios are further weighted so that rarely seen co-occurences contribute little to the resulting vectors (through loss function) and there's a limit to which frequent co-occurencies might contribute.

We use the original implementation provided by authors.[2]

3.4 LexVec

LexVec (Salle et al., 2016b,a; Salle and Villavicencio, 2018) is, like GloVe, a counting model. It again utilizes co-occurence matrix and weights the

[1] We hope to overcome this limitation in our future works.

[2] https://github.com/stanfordnlp/GloVe

errors so that more frequent co-occurences contribute more. It however also employs the negative sampling (originally introduced as an extension to skip-gram Word2Vec (Mikolov et al., 2013b)) to force scattering vectors of unrelated words.

Since the first paper, LexVec has been extended with positional context (i.e. it is not important only whether word a appeared in the context of word b but also whether it was to the left or to the right and how many words there were in between), the ability to use external memory for storing the co-occurences (which allows to train on a huge corpus), and finally with subword information (which allows deriving vectors even for out-of-vocabulary words).

We use the original implementation provided by authors.[3]

3.5 BERT

BERT (Devlin et al., 2018), which stands for bidirectional encoder representations from transformer, is a neural predictive model. It is trained on sentences rather than on words themselves (actually, its inputs are sentence pairs) but it does produce word embeddings. It's training can be viewed as a two-step process, the model is first pre-trained using specific tasks and then fine-tuned using downstream tasks.

The two tasks used to pretrain the model are next sentence prediction (i.e. deciding whether the second sentence really followed the first one in original text or if it was picked at random) and something the authors call masked language model, which is very close to a cloze test (Taylor, 1953). The idea is that some amount of randomly chosen words is masked (i.e. replaced with a special token), and the model has to correctly predict them.

We do not train the model, we use the distributed multilingual model [4] and our department wrapper around it.

Because of its different nature, we evaluate this model only on similarity datasets (those described in subsections 5.2 and 5.3).

4 Corpora

In this section, we briefly describe the corpora we use to train the models on.

Apart from using different models and corpora to train on, we have also experimented with two more settings: token form (i.e. training either on forms as they appear in the corpora, or on lemmata) and keeping/substituting numbers. The second idea is rather a concept, though concretized in numbers – some words have similar function but come in many different forms (and remain distinct when lemmatized) so their token counts are low and they do not take big part in the training. However, their similar function suggests that they could still be useful in defining contexts/concepts. We therefore tried also substituting all numbers (i.e. tokens tagged as `C=` by MorphoDiTa tool) with a meta-word.

For lemmatization of both corpora and datasets, we have used the MorphoDiTa tool (Straková et al., 2014).

4.1 Czech Wikipedia dump

As is common in natural language processing, we use Wikipedia dump as a training corpus. This corpus consists of short documents (hundreds to thousands words), the style is encyclopedic but not really expert. No shuffling has to happen, all co-occurences are kept as they are. Unfortunately, Czech Wikipedia is rather small when compared to the English version, thus we expect it to produce worse results.

We have used the dump from 1st May 2019. We have processed the dump with wikiextractor[5] and tokenized it using MorphoDiTa (Straková et al., 2014) tool.

4.2 CzEng

CzEng (Bojar et al., 2016) is a parallel Czech-English corpus. The texts in it are of varied domains, including news, fiction, laws, movie subtitles and tweets. It is shuffled at block level, i.e. only a few consecutive sentences are kept together each time. Each sentence is associated with its domain but it is not possible to reconstruct the original documents.

We have used version 1.7 and only extracted the Czech part of CzEng, keeping the tokenization and lemmatization provided in CzEng. We have reordered the sentences so that all sentences which share the domain are grouped together.

[3] https://github.com/alexandres/lexvec

[4] https://github.com/google-research/bert

[5] http://attardi.github.io/wikiextractor/

4.3 Czech National Corpus

Czech National Corpus (Křen et al., 2016) is a large corpus of written Czech. Version SYN v4, which we have used, contains texts of varying types, however news are by far the most common. (This version is not considered representative because of the prevailing news, but it is much larger than representative CNC subcorpora.) The corpus is again shuffled at block level, sentences are linked with the exact document they come from but their order cannot be reconstructed.

5 Datasets

We describe the existing (as well as no-longer-existing) datasets suitable for the evaluation of Czech word embeddings.

5.1 RG-65 Czech (unavailable)

RG-65 Czech (Krčmář et al., 2011) is (or perhaps used to be) a Czech version of the famous Rubenstein-Goodenough set (Rubenstein and Goodenough, 1965), a set on word relatedness.

In the original set, the data are triplets: two words and a mean relatedness score as annotated by human annotators, there are 65 word pairs.

The authors decided to translate the word pairs, using a reference on the original meanings. Pair relatedness was annotated by 24 human annotators of varying age, gender and education. During the translation and annotation process, a total of 10 pairs was omitted since one of the words could not be easily translated (or it would be translated to exactly the same word as the other).

Unfortunately, this dataset seems not to be available any more. The URL provided in the paper does not work, neither does the first author's email. We have tried contacting another author of the paper but they did not have the data.

We therefore do not evaluate on this dataset, however we think it should be listed when discussing all relevant datasets.

5.2 WordSim353-CZ

WordSim353-CZ (Cinková, 2016) is a Czech version of WordSim (Finkelstein et al., 2002), which is another dataset on word relatedness. The data are again triplets, word pair and a score (though technically the Czech dataset contains other information for each pair).

The author decided to create a dataset as similar as possible to the original, which especially means she encouraged the annotators to annotate relatedness, even though the name refers to similarity.

During the process of creating the dataset, four candidate translations were suggested for each of the original pairs, and 25 annotators annotated all reasonable pairs. The authors then selected the pairs so that the correlation between Czech and English rankings is maximal.

A version with all annotated pairs is available but we stick to the selected subset in our experiments.

5.3 Czech similarity and relatedness

The dataset for Czech similarity and relatedness (Konopik et al., 2017) not only enables another evaluation of word similarity, it also addresses the problem of scoring words which are closely related but not really similar (those may be e.g. antonyms or pairs like beach and sand). This dataset contains 953 words.

The authors decided to build the dataset from several different resources. They translated random pairs from several English datasets, RG-65 (Rubenstein and Goodenough, 1965), WordSim (Finkelstein et al., 2002), MTurk (Radinsky et al., 2011), Rare words (Luong et al., 2013) and MEN (Bruni et al., 2014). They also mined translational data using Moses (Koehn et al., 2007) and CzEng (Bojar et al., 2016), language part of Czech general study tests SCIO, and they invented a few pairs on their own.

Word pairs were annotated by 5 annotators and each of them annotated both similarity and relatedness, the annotators achieved Spearman correlation of 0.81.

The dataset itself does not contain only the word pairs and their scores but also examples of their usage, examples of ambiguities (sentences containing the same word with different meaning) and examples of the two words co-occuring; all examples were taken from the Czech National Corpus.

5.4 Czech analogy corpus

Czech analogy corpus was presented as a part of embedding-related experiments by Svoboda and Brychcín (2016), and it mimics the Google analogy test set.[6]

[6] http://download.tensorflow.org/data/questions-words.txt

It contains 11 relationship categories. Of those, 4 are purely semantic (capital cities and three groups of antonym relations, further divided based on part-of-speech), 3 are purely syntactic (noun plural, verb past tense, adjective gradation), 3 are rather syntactic (gender variation of job names and of nationalities, grammatic number variation of pronouns, including pairs like I and we) and 1 is rather semantic (family relations, i.e. he-cousin to she-cousin as father to mother). While family relations is in fact gender variation of family roles, feminime variants usually cannot be derived from masculine ones.

There is also a phrase version which contains some additional categories but we use the version containing single words only.

5.5 Extended semantic analogies

We have developed four additional analogy categories and word pairs representing those categories. These categories are: old or even archaic words and more modern words with the same / close enough meaning, e.g. *biograf* and *kino* 'cinema'; diminutives, e.g. *máma* 'mum' and *maminka* 'mummy'; more foreign-sounding (often expert) words and their more Czech-sounding variants, e.g. *akceptovat* and *přijmout* '(to) accept'; and synonyms.

While we understand and acknowledge the ambiguity of listed relations, we believe some ambiguity accompanies also antonyms and family relations, and we are curious about the model performance.

5.6 Synonym retrieval

We propose evaluating word embeddings also on synonym retrieval. Using our department thesaurus, we have randomly selected 500 words known to have at least 5 synonyms. (No two tested words are synonyms of each other.)

For each tested word, we find 10 words having the most similar vectors and we evaluate the top-1, top-3 and top-5 precision. We do it both with respect to the answer really given and with respect to an oracle which would move true synonyms to top positions whenever they would appear within the 10 candidates.

Please note that even though Leeuwenberg et al (2016) have shown that relative cosine similarity is a better approach to synonym extraction, it does not make a great difference in our case because we do not need to set a similarity threshold between synonyms and non-synonyms.

6 Results

We evaluated all trained models on all available datasets, with the exception of BERT embeddings which were only evaluated on WordSim353-CZ and Czech similarity and relatedness dataset. Please keep in mind that we were not able to train fastText on Czech National Corpus using forms and LexVec on Czech National Corpus using either forms or lemmata.

When evaluating analogies, we have tried using both 3CosAdd suggested by Mikolov et al (2013b) and 3CosMul suggested by Levy and Goldbert (2014) as similarity objective. To evaluate similarity, we use cosine similarity in all tasks.

Since the number of trained and evaluated models is high, we do not report results for each of the models. Instead, we do the following:

- Divide the tasks into five groups: syntactic analogies, semantic analogies, extended analogies, similarity/relatedness assignment and synonym retrieval. For each group, we identify all models which achieve the best result on any task within this group, and for all such models, we report results on all tasks within this group. We also report the performace of BERT embeddings on the similarity/relatedness group.

- Report a basic approximation of parameter volatility, given by differences in performance when only the parameter in question is changed.

- Discuss the patterns we have noticed during our examination of all results.

Table 1 shows the results on syntactic analogies. Please note that the dominance of models trained on forms is expected since models trained on lemmata are not able to solve purely syntactic tasks (plural, past tense, pronouns, gradation). The non-zero accuracy of lemmatized models on plural is only possible due to a dataset lemmatization error.

The achieved accuracies are pleasing, with a notable exception of pronoun analogies. We suspect this could be because the pronoun analogies in fact mix several aspects, i.e. there are pairs like *já* 'I' - *my* 'we' but also *mého* 'my (sg., i.e. my thing)' - *mých* 'my (pl., i.e. my things)', instead of *mého*

	fastText Wiki forms 3CosMul	Word2Vec CNC forms 3CosAdd	Word2Vec CNC forms 3CosMul	LexVec CzEng forms 3CosMul	fastText CNC lemmata 3CosMul
Plural	**71.85**	64.11	64.04	68.17	2.70
Jobs	83.92	**87.54**	85.35	80.72	75.00
Past tense	**89.02**	66.58	67.84	87.79	0.00
Pronouns	7.54	9.79	**10.98**	10.45	0.00
Gradation	60.00	62.50	60.00	**70.00**	0.00
Nations	43.18	25.19	28.03	40.15	**67.52**

Table 1: Results on syntactic analogies; numbers were always kept in place; both Word2Vec and fastText were trained using CBOW architecture

	LexVec CzEng lemmata meta 3CosAdd	Word2Vec Wiki lemmata meta 3CosAdd	LexVec Wiki forms numbers 3CosAdd	Word2Vec CNC lemmata meta 3CosAdd	LexVec CzEng lemmata numbers 3CosMul
Anto-nouns	**23.33**	13.44	17.28	14.72	18.56
Anto-adj	20.96	**31.71**	3.54	23.17	20.15
Anto-verbs	6.79	5.27	**13.66**	7.68	6.70
City/state	5.35	41.62	3.03	**54.72**	5.08
Family	45.99	41.98	8.03	43.83	**48.61**

Table 2: Results on semantic analogies;

	CNC numbers 3CosMul	CNC meta 3CosAdd	CNC meta 3CosMul	CzEng numbers 3CosMul
Archaic	**18.92**	15.92	17.72	7.56
Diminutives	25.97	**27.66**	**27.66**	13.97
Expert	23.09	19.63	**23.48**	14.19
Synonyms	20.27	19.79	19.79	**26.71**
Total	22.80	21.32	**23.17**	14.83

Table 3: Results on extended analogies; all models were trained using Word2Vec with CBOW architecture, corpus was always lemmatized

'my' - *našeho* 'our'. The possessive pronouns are also given in genitive/accusative (same forms are used for both cases) while the personal pronouns are given in nominative.

Results on semantic analogies are reported in table 2, and results on extended analogies are given in table 3.

Consistently with Svoboda and Brychcín (2016), we have found that CBOW outperforms skip-gram on Czech, which is not consistent with the results observed on English (Mikolov et al., 2013b). We hypothesize this could be due to relatively free word order and strong inflection and conjugation found in Czech. For example, while the two sentences *Profesor pochválil studenta* and *Studenta pochválil profesor* 'A professor praised a student', have a different meaning with respect to topic-focus articulation, they can be both utilized in Czech to communicate roughly the same thing. In English, changing the word order would also require transforming the verb. Therefore, a single Czech word could be less predictive than a single English word, making skip-gram less effective.

The best result is not always achieved using the largest corpus available. Out of 15 analogy classes, 4 are best solved when training on

Wikipedia dump. The difference is very subtle for noun plural, rather subtle for past tense and verb antonyms (with LexVec trained on CzEng being the second in all cases) but high for adjective antonyms (the best non-wiki model achieves accuracy of 25.67). While we are not able to truly explain, we suspect several factors could be responsible: Wikipedia dump is probably more consistent in style than both other corpora (which are compilations of various sources); many pages originated as English Wikipedia translation and thus are likely to follow English stylistics, making the language more similar to English; its encyclopaedic nature could make the language more regular in general. Perhaps these properties could outweight the corpus size.

However, CNC in general gives good results on extended analogies. We suppose its size does make an advantage, though indirectly, by making the appearance of queried words in the corpus more likely and their contexts more recognizable (some words are unusual in Czech, especially words from archaic and expert analogies).

We notice that while syntactic analogies are better solved by models trained on forms (with the exception of gendered nationality analogies), most semantic analogies are better solved by models trained on lemmata. We suppose this is due to large numbers of word forms for each lemma (a prototypic Czech noun has 14 forms, adjectives and verbs have even more), further strengthened by lemmata having some basic sense disambiguation annotation.

The exception to lemmata performing better are verbal antonyms. The best lemma-based model achieves accuracy of 10.18, which is notably lower than the best result. We are not sure about the cause. However, verbs have lots of forms (which all get lemmatized to the same string) and many verbal forms contain auxiliary words, often also verbal. The combination of that could make distinguisting contexts more difficult.

Table 4 gives the results on similarity tasks. To evaluate BERT on Czech similarity and relatedness dataset, we extracted all example sentences (which are given to demonstrate the use of the word with the desired meaning) and inferred the embeddings of all words in them. We then used the embeddings of the queried words to evaluate the model.[7]

We were quite surprised to see the relatively low results of BERT, compared to other models. We suspect the elimination of accented characters could hurt BERT performance since accents may differentiate meaning in Czech and the removal of accented characters might produce the same string (as turning both *malý* 'small (masculine)' and *malá* 'small (feminime)' into *mal*; *můra* 'moth' and *míra* 'measure, rate' to *mra*) or even to valid Czech words (as turning *zeď* 'wall' into *ze* 'from'). However, this should be rather rare, except for systematic occurences as with the masculine and feminime adjectives.

We find it more likely that BERT performance is hurt by inferring embeddings of rather artificial sentences. For WordSim, the sentences had only the queried words. For the similarity and relatedness dataset, these were true sentences, but without further context.

The results on synonym retrieval are reported in table 5. We again see that CBOW architecture outperforms skip-gram, which might be because of relatively free word order in Czech. The effect could be even stronger in context of synonym retrieval, as the distinction e.g. between subject and object could also be the distinction between (near-)synonym and (near-)antonym verb.

The corpus size might be a more important factor than model selection for synonym retrieval. Even though moving from forms to lemmata helps both in general and specifically with this task, models trained on unlemmatized CNC often outperformed models trained on lemmatized CzEng. However, Word2Vec/CBOW trained on smaller lemmatized corpus still outperformed other models trained on CNC. Unfortunately, we cannot be sure about the performance of LexVec on CNC but its performance on CzEng is 30%-70% of word2vec/CBOW performance with the remaining parameters matching.

We have also noticed that while oracle precision is good, the synonyms often do not come first. The exact precisions differ but for all models, the real precision is one third to one half of oracle precision.

In all our experiments, GloVe did not perform

[7]Technically, we first associated the word with a unique identifier, added the identifier and the inferred embedding to a special model using gensim, and finally evaluated this special model against the translation of the dataset into the identifiers. The identifiers are needed since embeddings are contextualized, i.e. different for the same word in different contexts, but gensim only supports mapping one word to one embedding.

	fastText CNC meta	LexVec CzEng numbers	BERT
SimRel/Similarity	**72.45**	65.39	46.90
SimRel/Relatedness	**66.51**	62.14	38.63
WordSim353-CZ	69.17	**70.41**	13.88

Table 4: Results (Spearman correlation coefficient) on similarity tasks; All models were trained on lemmatized corpus; fastText was trained using CBOW architecture

	numbers	meta-numbers
Top-1 oracle	**78.10**	74.79
Top-3 oracle	47.80	**49.45**
Top-5 oracle	32.07	**32.48**
Top-1 precision	**35.12**	33.88
Top-3 precision	**26.86**	26.45
Top-5 precision	21.65	**22.15**

Table 5: Results on synonym retrieval; Models were trained using Word2Vec/CBOW on lemmatized CNC

well. The rank of best performing GloVe model was usually around 30, therefore being worse than about a quarter of all other models. It is however possible that GloVe would benefit from tweaking the parameters more carefully. Altering a parameter often has the opposite effect on GloVe than on other modesl, which also encourages this assumption. Still, it should be noted that this result again is consistent with the findings of Svoboda and Brychcín (2016), who discovered GloVe performed worse than Word2Vec on Czech.

Despite all the research into incorporating subword information into embeddings (which is, among other, motivated by morphologically rich languages), models trained on lemmata perform better that their counterparts trained on forms. Tasks which require form distinguishing are a natural exception to this. We suspect this gap is partially caused by some forms being quite different from its lemma (and therefore hardly connectable on form/subword level), by lots of forms being only seemingly similar (sharing a long substring but meaning a different thing), and also by some forms appearing in specific contexts only (making the model learn a relation more specific than it should be).

However, we believe performing a strictly syntactic evaluation of embeddings which would focus on deriving correctly inflected/conjugated forms would be an interesting experiment to evaluate to benefits of subword information in morphologically rich languages.

As has been already mentioned, CBOW outperforms skip-gram on Czech. The difference is bigger on syntactic analogies; CBOW advantage is less clear in fastText models than Word2Vec models and on similarity tasks (in which CBOW only outperforms skip-gram if trained on lemmata).

Word2Vec with CBOW architecture generally performs well, though there are tasks (especially similarity assignment) on which LexVec gives notably better results.

Number substitution with meta-words alters the results only slightly. Though sometimes the best result is achieved by a model trained on text with those meta-words, the substitution hurts more often than it helps.

Similarly, the difference in analogy performance between different similarity objectives is rather subtle, though it is notable that semantic analogies are generally best solved with 3CosAdd objective while syntactic analogies are generally best solved with 3CosMul. However, this pattern is not repeated in extended analogies which are mostly semantic but best solved with 3CosMul (though the results on extended analogies are low which might further reduce the effect of similarity objective).

In general, training on CzEng instead of CNC results in worse results, suggesting CNC is more appropriate for training word embeddings. The difference of size is likely to play a role, but without further investigations we cannot eliminate the possibility that the fact that CzEng is comprised of more different text types also worsens (or possibly improves) the results.

The comparison of training with CzEng and Wikipedia dump is less one-sided. In most cases, moving from CzEng to Wikipedia dump has a negative impact, however it does improve the re-

Parameter	Mean	Deviation	Maximum
model	9.61	12.62	87.54
corpus	7.22	9.38	55.61
form/lemma	13.87	19.85	89.02
form/lemma*	5.89	6.36	30.30
numbers	1.13	1.70	15.00
similarity objective	1.58	2.34	18.54

Table 6: Approximation of parameter volatility given by the distribution of performance differences (percent points) when altering the parameter with all remaining parameters fixed; Minimum difference is always 0; similarity objective is only taken into account on analogy tasks; * line refers to values when skipping noun plural, past tense, pronouns and gradation which are by nature unsolvable by lemmata-based models

sults on several task/model combinations (especially syntactic analogies). We also noticed that on similarity assignment, LexVec performs better than most models when both are trained on CzEng but worse when trained on Wikipedia dump (the comparison for CNC is not available). The effects of moving from CNC to Wikipedia dump are similar to effects of moving from CNC to CzEng (i.e. usually negative).

7 Conclusion

We have presented an intrinsic evaluation of Czech word embeddings. We have evaluated several models trained on three different corpora, using different strategies during the training process. We have evaluated the resulting embeddings on a variaty of tasks – analogy, similarity, synonym retrieval.

The most important of our findings, regarding model selection, are that GloVe model using the default parameter settings does not seem to work well on Czech, that CBOW architecture of Word2Vec/fastText generally outperforms the Skip-gram architecture (unlike on English) and that LexVec performs fairly well in our experiments. It is worth noting that model selection affected the results more than corpus selection.

While bigger corpus might be expected to give better results, our results regarding corpus size are mixed. In most cases, the best performing model is trained on CNC, the largest corpus we have used, and if the best result is achieved using CzEng, the model is usually LexVec (which we were not able to train on CNC). However, the best result in several tasks is achieved using Wikipedia dump. We hypothesize the encyclopaedic nature of Wikipedia and the similarity of its language to English (following from many pages being translated or based on their English counterparts) could be important factors.

We have also found that models trained on lemmatized corpus usually perform better. Given that lemmatization tools are available for Czech, we would therefore recommend lemmatizing the text even when training on models which employ subword information. We hypothesize the differences of forms as well as some basic sense disambiguation might play a role.

We have several future goals which have emerged from the described work. Obviously, overcoming the limitations and being able to train all models on any corpus is one of them. We expect to try reformumalting the analogy task so that there can be more than one correct answer (which is clearly useless for tasks like currect capitals but might be interesting for tasks like antonyms or diminutives). We would also like to create more syntactic tasks to further evaluate the benefits of subword information, train the models on corpora subsets to better evaluate the effect of using bigger corpus, and carefully evaluate analogies and synonym retrieval using contextualized embeddings.

Acknowledgement

This work has been supported by the grant No. 1704218 of the Grant Agency of Charles University and the grant No. 19-19191S of the Grant Agency of Czech Republic. It has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). The research was also partially supported by SVV project number 260 453.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Silvie Cinková. 2016. Wordsim353 for czech. In *International Conference on Text, Speech, and Dialogue*, pages 190–197. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Miloslav Konopik, Ondřej Pražák, and David Steinberger. 2017. Czech dataset for semantic similarity and relatedness. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 401–406.

Lubomír Krčmář, Miloslav Konopík, and Karel Jezek. 2011. Exploration of semantic spaces obtained from Czech corpora. In *DATESO*, pages 97–107.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Zasina. 2016. SYN v4: large corpus of written czech. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Artuur Leeuwenberg, Mihaela Vela, Jon Dehdari, and Josef van Genabith. 2016. A minimally supervised approach for synonym extraction with word embeddings. *The Prague Bulletin of Mathematical Linguistics*, 105(1):111–142.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*.

Neha Nayak, Gabor Angeli, and Christopher D Manning. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 19–23.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke

Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

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

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016a. Enhancing the lexvec distributed word representation model using positional contexts and external memory. *arXiv preprint arXiv:1606.01283*.

Alexandre Salle and Aline Villavicencio. 2018. Incorporating subword information into matrix factorization word embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 66–71.

Alexandre Salle, Aline Villavicencio, and Marco Idiart. 2016b. Matrix factorization using window sampling and negative sampling for improved word representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 419–424.

Hinrich Schütze. 1993. Word space. In *Advances in neural information processing systems*, pages 895–902.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Lukáš Svoboda and Tomáš Brychcín. 2016. New word analogy corpus for exploring embeddings of Czech words. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 103–114. Springer.

Wilson L Taylor. 1953. "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054.