Recognition of Lexico-semantic Relations in Word Embeddings for Polish

anonymised ...

anonymised ...

Abstract

Word embeddings were used for the extraction of hyponymy relation in several approaches, but also it was recently shown that they should work. In our work we verified both claims using a very large wordnet of Polish as a gold standard for lexico-semantic relations and word embeddings extracted from a very large corpus of Polish. We showed that a hyponymy extraction method based on linear regression classifiers trained on clusters of vectors can be successfully applied on large scale. We presented a possible explanation for contradictory findings. Moreover, we extended the method to the recognition of meronymy.

1. Introduction

A very large wordnet, e.g. plWordNet (Maziarz et al., 2014) describes many lexico-semantic relations, linking lexical units (or word senses) by thousands of relation instances. However, even in a very large wordnet some relation instances can be omitted and typically wordnets are very biased towards only a few relations, e.g. hypernymy/hyponymy for nouns, with much smaller coverage for the other. Measures of semantic relatedness constructed on the basis of word embeddings (Mikolov et al., 2013b) are known to express many different lexico-semantic relations, e.g. on a list of the k most related words to a word x we can typically find words associated with x by different relations. However, word embeddings are very heterogeneous with respect to the types of semantic associations expressed, also including syntactic and pragmatic relations. What is worse, word embeddings have problems with representing different senses of a word (typically only a few most frequent can be spotted on the lists) and with proper representation of less frequent words (even words with frequency 100–200 per 1G can be erroneously described, not mentioning those < 100). The question is whether we can successfully recognise among the associations suggested by word embeddings those that correspond to lexico-semantic relations, i.e. whether we can interpret word embeddings in a meaningful way for humans. The works presented for English are contradictory even in the case of the relatively simplest relation, namely *hypernymy*: from successful extraction (Fu et al., 2014) till denial of the feasibility (Levy et al., 2015). We want to reapproach this intriguing issues, first checking the contradictory points of view on large corpora and comprehensive wordnet for Polish, second by expanding this research with one more relation, a more difficult one, namely meronymy. This a part of broader work on the automated extraction of lexico-semantic relations that are under-represented relation in wordnets, e.g. in order to improve wordnet-based WSD.

2. Related Works

ClassHyp system of (Piasecki et al., 2008) used a representation based on Distributional Semantics expanded with several statistical knowledge source extracted from corpora in supervised learning applied to the extraction of selected wordnet relations from text corpora.

(Fu et al., 2014) assumed that as a hyponym extends features of its hypernym (i.e. shares features with its hypernym and adds more specific ones), the hyponym's and hypernym's word embedding vectors should be related in some characteristic way, i.e. we can find some aspect of semantic inclusion when comparing both vectors. They proposed to use offsets between embedding vectors as representation of the projection (or "mapping") of a hyponym on its hypernym. Offsets were simply calculated by subtracting vectors representing a hypernym y and a hyponym x : $\mathbf{y} - \mathbf{x}$. As (Fu et al., 2014) observed that the hypernymy relation can be varied beyond one uniform projection, they proposed to cluster the offset vectors for know training parts into a number of groups and next to train a separate classifier based on linear regression for each group. Training examples were taken from a large Chinese thesaurus (5 level hierarchy, quite shallow, with coarse grained sense distinction). Vector offsets for direct and indirect hypernymy pairs were clustered into separate groups, but the indirect pairs represented quite close relations (max. length 3). The number of clusters was established experimentally on a separate development set. A test example was classified as a positive if it was classified positively by at least one classifier.

(Levy et al., 2015) analysed several methods for the extraction of relations can that express different forms of lexical inference, e.g. hypernymy, entailment or causation. They tested four different ways for representing pairs of words by feature vectors based on word embeddings, namely: *concatenation*, *subtraction* (called *difference*) and representations by single vectors of one of the words. Several different tests proposed in literature were used. In most of the cases supervised methods based on two-vector representation were only slightly better than single vector representation of the more general word. (Levy et al., 2015) proposed also an evaluation experiment in which negative pairs are artificially built from words included in the positive pairs. By using a SVM classifier they showed that the correlation between match error and recall (positive) is close to the perfect correlation in a series of experiments. As a result, (Levy et al., 2015) concluded that the supervised classifiers proposed in literature, including (Fu et al., 2014), are learning whether "y is a prototypical hypernym (i.e. a category) regardless of x, rather than learning a concrete relation between x and y." They called this potential effect "lexical memorizing". They also claimed that "contextual features might lack the necessary information to deduce how one

word relates to another". However, it is worth to notice that (Levy et al., 2015) did not apply the original approach of (Fu et al., 2014) in their key test (sic!). Moreover, all evaluations were done only for English and for quite limited test data.

However, there are also many other works that report on successful extraction of hypernymy from contextual features, e.g. (Shwartz et al., 2016).

3. Search for Relations in Word Embeddings

(Fu et al., 2014) intuitively seems to be correct: elements of the word embeddings are derived from the occurrence contexts and correspond to the semantic features of words, while the similarity of features of two words correspond to the amount of overlapping in the values. Thus, we wanted to revisit the method of (Fu et al., 2014) in a new setting and confront it with the objections of (Levy et al., 2015).

As a gold standard for lexico-semantic relations we used plWordNet – a very large wordnet of Polish (Maziarz et al., 2016). It is substantially bigger then Princeton WordNet (Fellbaum, 1998), and was constructed from scratch using a corpus-based wordnet development method. As a result plWordNet has much better coverage of words in large corpora. As source of text data, we utilised plWordNet Corpus (henceforth plWNC) which includes ≈4 billion words and combines all publicly available Polish corpora, and a very large number of Polish texts collected from the Web[1].

Using *word2vec* tool (Mikolov et al., 2013a) we built embedding vectors as representations for all words from plWNC with the frequency ≥ 8 (min_count=8). We tested several different settings of *word2vec* (to be presented in the full paper), and finally, we selected the Skip-gram model and two vector sizes: 100 and 300, as best performing.

Following (Fu et al., 2014), we represent a hyponymy instance (link): $\langle x, y \rangle$, where x – a hyponym, and y – a hypernym, by the difference of two word embedding vectors: $\mathbf{x} - \mathbf{y}$. It is also assumed that comparison of the difference vectors should reveal a *projection* corresponding to the hyponymy feature sharing pattern. This hypernymy projection, reducing the hyponym specific features, can be expressed by a linear projection of the vector $\mathbf{x}$, i.e. $\Phi\mathbf{x}$, on a vector $\mathbf{y}'$. However, both vectors can represent other semantic aspects beyond the feature sharing (e.g. polysemy, differences in context of occurrences etc.). In order to obtain a more regular picture, difference vectors for the training hyponymy instances are automatically clustered and for each cluster a different classifier is trained. The *k-means* algorithm was used for clustering and for each cluster a separated classifier was trained by the linear regression method. In a similar way, negative examples of non-hyponymic pairs constructed on the basis of plWordNet are clustered and negative classifiers are built. A test difference vector for a pair $\langle x, y \rangle$ is classified as representing hyponymy if it is positively classified by at least one of the created classifiers.

Semantic representation based on word embeddings has several limitations, e.g.: (i) the whole model can be biased by the particular selection of texts, (ii) senses of polysemous words are merged together, i.e. represented by a single vector, and (iii) the representation of less frequent words and senses can be blurred by the statistical noise. This problem was not explicitly and well enough treated in both contradictory works, namely: (Fu et al., 2014) and (Levy et al., 2015). In order to decrease the potential bias, the point (i), we used as large and diversified corpus as possible. To understand the influence of uneven representation of different senses, (ii), we divided experiments into two groups of: monosemous words only, and all words. We used also a large number of words in the experiments. To avoid noise caused by low frequency of data, we took into account in all experiments only words with more than 1,000 occurrences[2]. (Fu et al., 2014) tested their method for several different number of clusters achieving the best results with a larger numbers. We establish the value of this parameter by automated optimisation on a *development* subset. In each experiment the data were randomly divided into three subsets: *training*, *testing* and *development* in the ratio 6:2:2.

As noun hypernymy in plWordNet forms quite deep hierarchical structures (in some cases beyond 20 levels), testing indirect hyponyms with longer hypernymy paths could also make the analysis of the results more difficult. We limited positive examples only to direct hyponymic pairs.

Following the suggestion of (Levy et al., 2015), we constructed the data sets for experiments in two ways: (A) *random* division into subsets, (B) and *lexical train/test splits* rule proposed by them. In (B) the division is random but positive cases in the test set (i.e. true hyponymy instances) cannot include hypernyms occurring in the training set. Moreover the negative cases are also constructed in a tricky way explained below in Eq.1–3, where T^+ is a set of word pairs belonging to the given relation:

$$T_x^+ = \{x \mid (x, y) \in T^+\} \quad (1)$$

$$T_y^+ = \{y \mid (x, y) \in T^+\} \quad (2)$$

$$S = (T_x^+ \times T_y^+) \setminus T^+ \quad (3)$$

S contains false relation instances, but constructed from words included in the positive example, also hypernyms that are suspected to be the signal recognised by classifiers. Results of experiments on recognition of the hyponymy relation are presented in Tab. 1. Each experiment was performed in $k = 10$ fold cross-validation setting. Due to the limited space, only average results are presented in Tab. 1. In the following experiments we have analysed:

Hypo-Mono – hyponymy recognition for monosemous words: 6,000 hyponymy pairs including only monosemous words as positive examples, 6,000 negative examples; the two variants of the generation of negative

[1] It consists of IPI PAN Corpus (Przepiórkowski, 2004), the first annotated corpus of Polish, National Corpus of Polish (Przepiórkowski et al., 2012), Polish Wikipedia (from 2016), *Rzeczpospolita* Corpus (Weiss, 2008) – corpus of electronic editions of a Polish newspaper from the years 1993-2003, supplemented with text acquired from the Web – only text with small percentage of words unknown to a very comprehensive morphological analyser Morfeusz 2.0 (Woliński, 2014) were included; duplicates were automatically eliminated from the merged corpus.

[2] A heuristic threshold applied in many works and which seems to heuristically demarcate the area of robust representations

examples were applied: *random* and *lexical split*; the size of the embedding vectors was 100.

Hypo-Poly – 20,000 hyponymy pairs including polysemous words; 20,000 negative examples were selected using the assumed two methods; the vector size 100.

Hypo-Mono300 – as in Hypo-Mono but the vector size is 300 in order to check a more fine-grained description, only lexical split method was used for the generation of the negative examples, i.e. the more difficult one.

Hypo-Poly300 – as above, but 20,000 hyponymy pairs including polysemous words were used, 20,000 negative selected by the lexical split.

Mero-Poly – 7,900 meronymy pairs (only the main subtype *part of*), 8,000 pairs of words that are not connected in *plWordNet* at all or are connected by paths longer than 3 links were selected as negative examples by the lexical split method, the vector size was: 100.

For pairs of experiments performed using two different ways of the selection of negative examples, as well as for two different sizes of the vectors we checked the statistical significance of the differences. First we tested if the results obtained in different folds come from the normal distribution by applying Shapiro-Wilk test, e.g. for **Hypo-Mono** we obtained p value of 0.8082 for the random selection results and 0.8648 for the lexical split series, so with the confidence level of 0.05 we cannot reject the null hypothesis that the results come from a normal distribution. Having normality confirmed, we applied *t-Student* test to the differences between results, e.g. in the case of **Hypo-Mono** p value is 0.4583 and with the confidence level 0.05 we cannot reject the null hypothesis of the lack of a difference between both series of results. In the case of **Hypo-Poly** the fold results do not come from a normal distributions according to Shapiro-Wilk test, so we applied *Mann–Whitney U test* and the obtained p value of 0.008931 shows the lack of statistical significance of the differences. In a similar way we checked that the differences between results for different vector sizes are significant, namely **Hypo-Mono** vs **Hypo-Mono300** and **Hypo-Poly** vs **Hypo-Poly300**. We did not analysed differences between the results for monosemous and polysemous words, but in these differences are very visible.

In Tab. 1, we can observe that in all experiments very good results in the recognition of hyponymy relation were achieved. As (Levy et al., 2015) expected, the lexical split selection of negative samples caused the decrease of the results. However, the observed differences are small ≈ 1 for monosemous and ≈ 2 for polysemous, while e.g. (Shwartz et al., 2016) reported the difference of ≈ 20. Moreover, these differences are not statistically significant. It means that recognition methods trained on other hypernyms that those in the test set are still working, properly recognise hyponymy instances and are not simply deviating to prototype recognition as suggested (Levy et al., 2015). Moreover, the small difference between the random and lexical split selections can be also attributed to the imperfection of the linear projection based on a limited number of clusters. In all cases recall is higher than precision, but in applications, e.g. in wordnet development, this is a required property, as we do not want to loose potential hyponymy instances. Significantly lower results were obtained for longer embedding vectors of 300 elements, especially for **Hypo-Mono300** are surprising. This can be caused by insufficient number of training examples, as in the case of **Hypo-Poly300** with larger training set the results are higher. An analysis of the correlation between results and the training set size will be presented in the full paper.

The substantial discrepancy of our findings with the claimed inability to train supervised recognition on the basis of word embedding vectors observed in (Levy et al., 2015) can be also caused by the choice of different classification methods: so far we followed the work of (Fu et al., 2014) and we combined unsupervised clustering with the construction of supervised classifiers based simply on linear regression, while (Levy et al., 2015) used only SVM algorithm. To complete the picture we also repeated for all experiments the error analysis proposed by (Levy et al., 2015), e.g. for **Hypo-Mono** it is presented in Fig. 1, the analysis of the other experiments will be included in the final paper. In Fig. 1 the ratio of the matching error (see Tab. 1 caption), a kind of 'negative' recall, and the positive recall for different folds is presented. If a classifier recognises not relation instances, but hypernyms as prototypes, than it reacts in a similar way to both negative and positive examples as the negative ones prepared by lexical split include hypernyms from the training data. (Levy et al., 2015) showed that this ratio for different experiments is perfectly set on the diagonal. In our case all values are far way from the diagonal.

We also used the training and testing data prepared according to lexical split from **Hypo-Mono** and a SVM-based classifier. Many experiments were performed with different settings of the classifier (kernels: linear, polynomial and radial, cost $C \in \{1, 10, 100, 1000\}$, example influence $\gamma \in \{0.001, 0.0001\}$) and 10-fold cross validation. The results were compared in the ratio analysis presented in Fig. 2. There is varied distribution of the ratio values in comparison to the univocally bad situation reported in (Levy et al., 2015). However, many of the points are located on the diagonal or close to it. This suggests that the pessimistic conclusions of (Levy et al., 2015) maybe limited only to some settings of the SVM classifier when applied to the relation recognition in the word embedding vectors.

It is worth to emphasise that we achieved a very good result for meronymy, see Tab. 1 by applying exactly the same method of (Fu et al., 2014) as for hyponymy recognition, i.e. linear regression classifiers trained on clusters of difference vectors. However, meronymy is typically more difficult relation to be extracted. Its recognition in the word embeddings vectors cannot be explained by sharing a prototype, as its more complex relation, and in this particular experiment we were using lexical split technique, too.

4. Conclusions

The claim of (Levy et al., 2015) that supervised classifiers trained on combinations of word embeddings vectors are learning in fact that one of the words is a prototypical hy-

Exp.	**Acc**	**P**	**R**	**F**	**Err**	**Type**	Vec. Size
Hypo-Mono	85.22%	78.91%	96.27%	86.72%	27.91%	Rnd	100
std. dev.	0.64%	1.00%	0.65%	0.65%	1.92%	Rnd	100
Hypo-Mono	84.98%	78.90%	95.18%	86.27%	28.05%	Lex. split	100
std. dev.	0.61%	1.59%	0.79%	0.91%	2.22%	Lex. split	100
Hypo-Poly	78.94%	74.35%	88.35%	80.74%	31.63%	Rnd	100
std. dev.	0.65%	0.41%	1.70%	0.79%	1.78%	Rnd	100
Hypo-Poly	77.23%	73.83%	84.66%	78.85%	30.54%	Lex. split	100
std. dev.	0.79%	1.40%	2.39%	1.04%	2.25%	Lex. split	100
Hypo-Mono300	73.31%	65.16%	98.20%	78.32% –		Lex. split	300
std. dev.	1.11%	1.82%	0.39%	1.31%	–	Lex. split	300
Hypo-Poly300	82.54%	84.51%	94.72%	89.32%	–	Lex. split	300
std. dev.	1.01%	1.11%	0.69%	0.73%	–	Lex. split	300
Mero-Poly300	79.95%	74.66%	90.43%	81.77%	–	Lex. split	100
std. dev.	1.05%	1.71%	1.38%	0.99%	–	Lex. split	100

Table 1: Supervised recognition of lexico-semantic relations on the basis word embedding vectors, where *Acc* is the percentage of correct decisions, *P* – positive precision, *R* – positive recall, *F* – F-measure from *P* and *R*, *Err* – the match error, $2FP/(TN + FP)$, a 'reversed' recall, *Type* – the selection method for negative examples and *Vec. size* – the size of the embeddings vectors. All results are average from the 10 folds cross validation. In *std. dev.* standard deviation calculated for 10 results is provided. In the case of similar experiments only the differences between **Hypo-Mono** vs **Hypo-Mono300** and **Hypo-Poly** vs **Hypo-Poly300** are statistically significant.

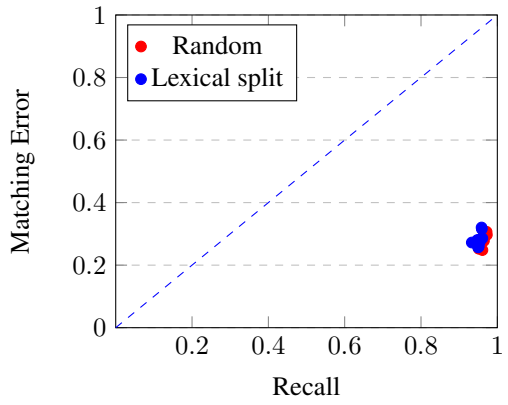

Figure 1: Ratio between the matching error and recall for different folds in **Hypo-mono** experiment and for the two methods of the selection of negative samples.

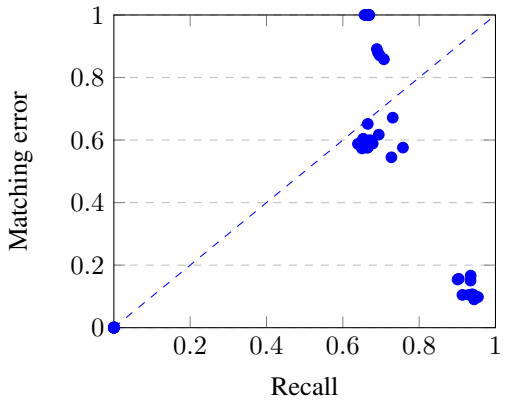

Figure 2: Ration between the matching error and recall for supervised recognition of hyponymy by using different configurations of SVM

pernym, instead recognising the pair as an instance of the hyponymy relation seemed to be well motivated, but contradictory with intuition and many results reported in literature. One of them, namely (Fu et al., 2014), presented good results, but tested on a limited scale of only 412 words. In our work we verified both claims using a very large wordnet of Polish (developed in a more linguistically oriented wayward) as a gold standard for lexico-semantic relations and word embeddings extracted from a very large, automatically pre-processed corpus of Polish. We showed that the method proposed by (Fu et al., 2014) can be successfully applied to the extraction of the hyponymy relations. In series of carefully conducted and evaluated experiments we verified negatively the objections of (Levy et al., 2015). This contradiction can be due to different languages and datasets used, but also to the fact that they concentrated their attention on the use of SVM classifiers only, while we showed that in some settings SVM classifier can produce much worse results for this particular task.

In addition we applied successfully the method of (Fu et al., 2014) also to the recognition of meronymy achieving very good results tested on a very large data sample prepared manually. We plan to expand this approach to other relations, e.g. lexico-semantic relations manifested derivationally that are quite numerous in Polish. We aim at building a semi-automated system for improving the density of relation in a wordnet. It will be also very valuable to continue the research on types of classifiers and experimental settings that make extraction methods of this types successful.

5. References

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *ACL (1)*, pages 1199–1209.

Omer Levy, Steffen Remus, Chris Biemann, Ido Dagan, and Israel Ramat-Gan. 2015. Do supervised distributional methods really learn lexical inference relations? In *HLT-NAACL*, pages 970–976.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plwordnet as the cornerstone of a toolkit of lexico-semantic resources. In *Proceedings of the Seventh Global Wordnet Conference*, pages 304–312.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Pawel Kedzia. 2016. plwordnet 3.0 - a comprehensive lexical-semantic resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Maciej Piasecki, Michał Marcińczuk annd Stanisław Szpakowicz, and Bartosz Broda. 2008. Classification-based filtering of semantic relatedness in hypernymy extraction. In Bengt Nordström and Aarne Ranta, editors, *Advances in Natural Language Processing, 6th International Conference, GoTAL 2008, Gothenburg, Sweden, August 25-27, 2008, Proceedings*, volume 5221 of *LNCS*, pages 393–404. Springer.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany, August. Association for Computational Linguistics.

Dawid Weiss. 2008. Korpus Rzeczpospolitej [Corpus of text from the online edition of "Rzeczpospolita"]. `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`.

Marcin Woliński. 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.