Detecting word level metaphors in Polish

Aleksander Wawer, Małgorzata Marciniak, Agnieszka Mykowiecka

Institute of Computer Science PAS, Jana Kazimierza 5, 01-248 Warszawa, Poland
(axw,mm, agn)@ipipan.waw.pl

Abstract

The paper addresses an experiment in detecting metaphorical usage of adjectives and nouns in Polish data. First, we describe the data developed for the experiment. The corpus consists of 1833 excerpts containing adjective-noun phrases which can have both metaphorical and literal senses. Annotators assign literal or metaphorical senses to all adjectives and nouns in the data. Then, we describe a method for literal/metaphorical sense classification. We use Bi-LSTM neural network architecture and word embeddings of both token- and character-level. We examine the influence of adversarial training and perform analysis by part-of-speech. Our approach proved successful and an F1 score that exceeded 0.81 was achieved.

1. Introduction

Understanding natural language utterances requires solving very many problems on very different levels. In spite of many attempts to solve NLP problems as an end-to-end task, there are still many contexts in which we want to understand words, to combine their meanings into larger schemes, and to add context constraints to sentence meaning. At every step, there is a need to resolve ambiguities which are an inherent feature of natural language understanding. Starting at the word level, many of them have several different meanings, like *bat* which can mean either a kind of solid stick or a flying mammal. To make the process of understanding more complicated, and the communication in natural language at the same time more interesting and challenging, people "invent" meanings resembling but different from canonical senses, e.g. *blue* means one of the colours but also *sad*. These meanings that have become very popular are listed in language dictionaries. We nevertheless often use a non-literal combination of words whose listing in dictionaries is difficult and not necessary, as the mechanism used to formulate such expressions is both predictable and highly productive. For example, there are many meanings of *raise* noted in the Oxford dictionary, but all of them are somehow connected with changing a position in physical space or in some sorts of lists. And then, we have the phrase *raise a question* which transfers *raise* from concrete space to an abstract one. Such word usages are generally called non-literal, and in this particular case – metaphorical (Lakoff and Johnson, 2008).

An efficient application capable of distinguishing literal from non-literal word occurrences can be very useful in many situations as in web search engines, information extraction modules and document clustering. Technically, the task can be treated as a word sense distinguishing one, but as unsupervised methods are still much less efficient than supervised ones, it is treated more as a classification or a sequence labelling task. We adapted a slightly modified approach in our paper – we identify all occurrences of words but only for the nominal and adjectival classes.

2. Related Work

Over the last decade quite a lot of work was done on metaphor detection, see (Shutova, 2015). In these many approaches, the metaphor identification task was defined variously. One group of papers concerned the classification of selected types of phrases (taken in isolation) into those which nearly always have a literal meaning, like *brown pencil* and those which have only figurative usage, e.g. *dark mood*. In this type of task adjective-noun phrases for English ((Tsvetkov et al., 2014), (Gutierrez et al., 2016)) and Polish (Wawer and Mykowiecka, 2017) were explored as well as verb constructions for English (Beigman Klebanov et al., 2016). Phrases which can have different usage can be clasiffy only in the wider context. In this field of research, some papers present experiments with identification of the type of a particular phrase occurrence in text, while in other approaches, all words from a given text are classified into literal or figurative use.

At first, mostly supervised machine learning approaches were used in which apart from features derived directly from the data, many additional data resources have been used. Among others, these features included, concreteness, imageability, WordNet relations, SUMO ontology concepts, sectional preference information, and syntactic patterns. Solutions based on neural nets training were then published. Several new approaches were elaborated and compared due to the shared task on metaphor identification on the VU Amsterdam Metaphor Corpus (Steen et al., 2010) conducted at the NAACL 2018 Workshop on Figurative Language Processing (Beigman Klebanov et al., 2018). Participants were given two tasks: the ALL_POS task, in which they had to repeat annotation at word level of every token in the presented test data, and the Verbs task, in which only verb annotation was taken into account. The best performing solution (Wu et al., 2018) used pretrained word2vec embeddings, embedding clusterings and POS tags as input to CNN and Bi-LSTM layers. In our approach, we tested adversarial training with Bi-LSTM layers.

3. Data Description

The experiment was performed on a corpus consisting of 1833 short pieces of text selected from the NKJP (National Corpus of Polish, (Przepiórkowski et al., 2012)). The corpus is built from over 45,000 tokens including punctuation marks and excerpt delimiters. Each excerpt consists of one to three sentences and the average length is 24.5 tokens. The part-of-speech annotation is done with

phrase	All	L	M
pełne garście 'handful'	216	52	164
gorzki smak 'bitter taste'	136	68	68
głęboka rana 'deep wound', deeply wounded'	91	65	26
cierpki smak 'sour taste', 'sour grapes'	57	35	22
fałszywa nuta 'false note''deceitfully'	56	23	33
czyste ręce 'clean hands'	33	10	23

Table 1: Most popular AN phrases

	adj		ppas		adj+ppas	subst		ger		subst+ger	total	
	nb	%	nb	%	nb	nb	%	nb	%	nb	nb	%
M	1184	19.1	106	19.0	1290	1306	11.0	81	21.2	1397	2687	14.2
L	5004	80.9	453	81.0	5457	10520	89.0	301	78.8	10821	16278	85.8

Table 2: Statistics of *M/L* annotations in the corpus

the help of the Concraft2 tagger (Waszczuk, 2012).

Each excerpt contains at least one adjective-noun (AN) phrase which could have an (*L*) literal or a (*M*) metaphorical meaning depending on the context. The corpus was collected to perform experiments in recognition of *M/L* senses of 165 different AN phrases. Table 1 shows phrases with the most numerous examples. Quite often, only one element of a metaphorical AN phrase has a metaphorical meaning. For example, in the phrase *gorzka prawda* 'bitter truth', which always has a metaphorical sense, the noun *truth* usually has a literal sense and only *gorzka* 'bitter' has a metaphorical sense. The label *L* is assigned to an AN phrase if both elements are annotated as literal, while *M* is assigned if any of two elements (or both) has a metaphorical sense.

In the experiment described in the paper, we decided to annotate all adjectives and nouns in the whole corpus and to detect the meaning of separate adjectives and nouns instead of the whole phrase.

The annotation was done by two researchers specialising in metaphors in Polish: Joanna Marchula and Maciej Rosiński (the corpus is available from: http://zil.ipipan.waw.pl/CoDeS) The annotators adapted the procedure for recognition of metaphorical usage of individual words developed for the VU Amsterdam Metaphor Corpus (Steen et al., 2010). A discussion about difficulties that arise when the method is applied to Polish is given in (Marhula and Rosiński, 2017), while a modified procedure for Polish is described in (Marhula and Rosiński, 2018). An inter-annotator agreement was tested on 51 excerpts consisting of 1246 tokens. In this fragment, there are 555 adjectives and nouns which were annotated by two people. 14 words were differently annotated, and the Cohen's kappa was equal to 0.899, so the annotators obtained very good agreement. As the kappa was high and the procedure for annotation was very time-consuming, we divided the corpus into two parts which were annotated separately by one person. The final annotation was reviewed by removing minor inconsistencies and omissions which was done by one of the annotators. 180 decisions were changed, the label *M* was changed into *L* in 54 cases, and in the opposite way 126 cases. Table 2 contains information regarding how many adjectives (regular adjectives and past participles fulfilling adjective roles), nouns and gerunds are annotated as having a literal and metaphorical meaning in the final annotation.

4. Experiment Description

The basic architecture in our experiment is the BiLSTM-CRF model similar to (Lample et al., 2016). In this model, word representation is concatenated from token-level and character-level embeddings. The latter are computed using character-level Bi-LSTMs, by combining final states of each directional network. Thus, generated word embeddings are then used as input to a token-level bidirectional LSTM deep neural network. Finally, inference is carried by a CRF layer instead of traditional softmax. The structure of the model is given in Figure 1. The number of hidden units in LSTM is set to 150, the initial learning rate to 0.01, and the batch size to 10.

We also experimented with adversarial training, a technique employed in the field of machine learning which, in its original variant, attempts to fool models through malicious inputs. Recent advancements of this technique, introduced in the area of natural language processing, focus on modifying word embeddings in a malicious manner to make the problem more difficult: the worst-case perturbation coefficient η is computed and added to the embeddings. The expected effect is regularisation. This method was found effective in POS tagging as described by (Yasunaga et al., 2018). We test this approach using three η values:

- 0 (adversarial component turned off),
- 0.05 (mild adversarial setting),
- 0.1 (aggressive adversarial setting).

In our experiments, we used Wikipedia-trained Polyglot (Al-Rfou et al., 2013) word embeddings for the Polish language.

We divided the data into three classes of tokens: *L* (literal), *M* (metaphorical), *O* (outside, this class covers every other token type).

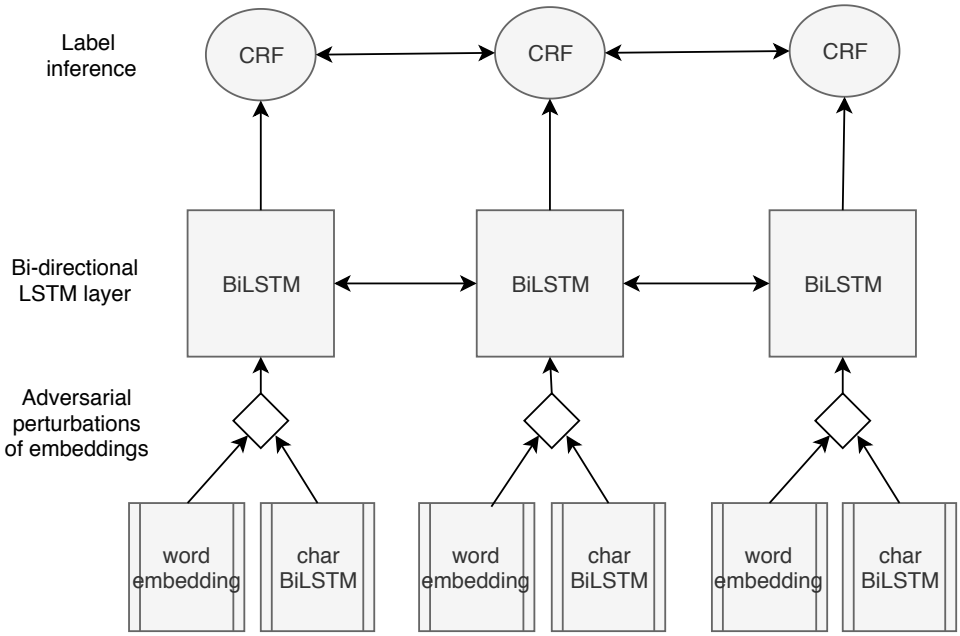

Figure 1: Diagram of the neural network.

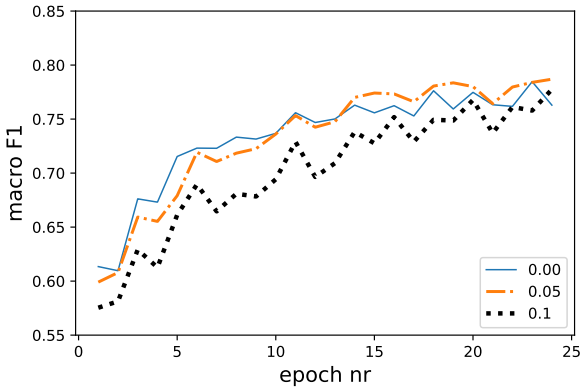

Figure 2: Macro F1 measure for various adversarial training rates - initial annotation.

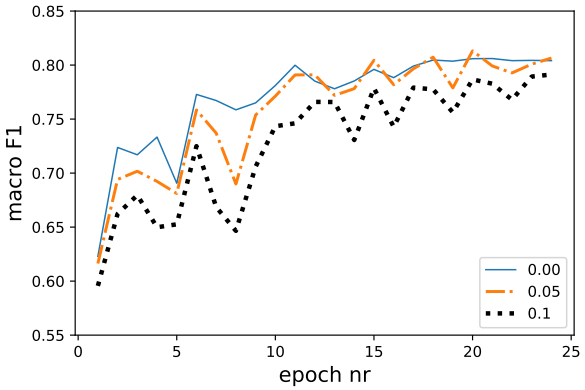

Figure 3: Macro F1 measure for various adversarial training rates - final annotation.

		precision	recall	*F1*
initial data	*M*	0.70	0.58	0.63
	L	0.94	0.94	0.94
	macro avg	0.82	0.76	0.79
final data	*M*	0.78	0.61	0.68
	L	0.94	0.92	0.93
	macro avg	0.86	0.77	0.81

Table 3: Detailed results of systems on the test data set, 25th epoch.

manual	auto	adj	ppas	subst	ger	total
M	*M*	101	37	92	0	198
L	*L*	485	5	1070	24	1616
M	*L*	31	7	50	5	93
L	*M*	38	1	41	1	81
M	*O*	1	2	1	4	8
L	*O*	16	2	40	1	59
total		672	54	1294	35	2055

Table 4: Results by POS in numbers

5. Results

We split the data randomly into three partitions: training (80%), development (10%) and test (10%). For each training epoch, results are reported for the test data set, for tokens that are either *L* or *M* according to the manual annotation. Figures 2 and 3 contain the macro *F1*-measure computed for various adversarial rates for two versions of the data set, respectively initial annotation and final annotation.

Table 3 contains the results measured at the end of training, 25th epoch, on both versions of the data. We use

	adj		ppas		subst		ger	
	M	*L*	*M*	*L*	*M*	*L*	*M*	*L*
precision	0.727	0.998	0.833	0.841	0.692	0.955	na	0.828
recall	0.759	0.900	0.357	0.925	0.643	0.930	na	0.923
F1	0.743	0.946	0.500	0.881	0.667	0.942	na	0.873

Table 5: Evaluation by POS

the mild adversarial setting of η 0.05 as it is the one which provides the best results. Table 4 gives the best results by POS in numbers. The first two columns represent manual and automatic annotation. The next columns gives numbers of annotated adjectives, past participles, nouns and gerunds in the test set. In Table 5, results by POS are given for precision, recall and *F1*-measure. The results of recognition metaphorical meaning of words are much worse as the training data contains almost five times fewer examples. Recognition of adjectives gives better results than nouns despite two times fewer examples in the training data.

6. Conclusion

The tested architecture is well-known and has been applied in the named entity recognition task, then also for part-of-speech (in its adversarial variant). Its application to the metaphor recognition task addressed in the paper has proven successful, as the best macro *F1* score achieved 0.81. The best results have been achieved with moderate influence of adversarial training. We tested two variants of the data set and the more coherent version has proven to perform better with the deep learning models. This fact shows that the consistency in annotating the data, in spite of the quite high kappa coefficient on the test fragment, is not perfect. It is an open question whether it comes from the different understanding of the metaphoricity or the annotation task itself. We plan to check whether detecting any kind of non-literal phrase occurrence would be easier or more difficult using the chosen method.

Our results point to the likely conclusion that metaphorical senses of adjectives are easier to recognize – the model recognizes them slightly better despite lower frequency.

Acknowledgments

This work was supported by the Polish National Science Centre project 2014/15/B/ST6/05186 *Compositional distributional semantic models for identification, discrimination and disambiguation of senses in Polish texts*.

7. References

Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena, 2013. Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Sofia, Bulgaria: Association for Computational Linguistics.

Beigman Klebanov, Beata, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor, 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Beigman Klebanov, Beata, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, and Chee Wee, 2018. *Proceedings of the Workshop on Figurative Language Processing*. Association for Computational Linguistics.

Gutierrez, Dario, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen, 2016. Literal and Metaphorical Senses in Compositional Distributional Semantic Models. In *Proceedings of ACL 2016 (short papers)*.

Lakoff, George and Mark Johnson, 2008. *Metaphors we live by*. University of Chicago press.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Marhula, Joanna and Maciej Rosiński, 2017. Co oferuje MIPVU jako metoda identyfikacji metafory? *Polonica*, XXXVII (37).

Marhula, Joanna and Maciej Rosiński, 2018. Chapter 9 Linguistic metaphor identification in Polish. In *Metaphor identification in multiple languages: MIPVU around the world*. https://osf.io/phf9q/.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (eds.), 2012. *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Shutova, Ekaterina, 2015. Design and Evaluation of Metaphor Processing Systems. *Comput. Linguist.*, 41(4):579–623.

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma, 2010. *A method for linguistic metaphor identification. From MIP to MIPVU*. Number 14 in Converging Evidence in Language and Communication Research. John Benjamins.

Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer, 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association of Computational Linguistics.

Waszczuk, Jakub, 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In

Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012).

Wawer, Aleksander and Agnieszka Mykowiecka, 2017. Detecting Metaphorical Phrases in the Polish Language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd.

Wu, Chuhan, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang, 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*.

Yasunaga, Michihiro, Jungo Kasai, and Dragomir R. Radev, 2018. Robust Multilingual Part-of-Speech Tagging via Adversarial Training. In *Proceedings of NAACL*. Association for Computational Linguistics.