Literal, Metphorical or Both? Detecting Metaphoricity in Isolated Adjective-Noun Phrases

Agnieszka Mykowiecka
ICS PAS
Jana Kazimierza 5
Warsaw, Poland
agn@ipipan.waw.pl

Małgorzata Marciniak
ICS PAS
Jana Kazimierza 5
Warsaw, Poland
mm@ipipan.waw.pl

Aleksander Wawer
ICS PAS
Jana Kazimierza 5
Warsaw, Poland
axw@ipipan.waw.pl

Abstract

The paper addresses the classification of isolated Polish adjective-noun phrases according to their metaphoricity. We tested neural networks to predict if a phrase has a literal or metaphorical sense or can have both senses depending on usage. The input to the neural network consists of word embeddings, but we also tested the impact of information about the domain of the adjective and about the abstractness of the noun. We applied our solution to English data available on the Internet and compared it to results published in papers. We found that the solution based on word embeddings only can achieve results comparable with complex solutions requiring additional information.

1 Introduction

One of the essential features of every natural language is its ambiguity. And apart from the homonymy and polysemy of words, the phenomenon which makes automatic text understanding difficult is the possible metaphorical usage of both simple and more complex phrases. Identification of potentially figurative usage is crucial for language processing efficiency and may improve the performance of many NLP applications. It is crucial for information extraction tasks, as the lack of figurative meaning detection may lead to false identification of a particular object or event (Patwardhan and Riloff, 2007). For example, we do not want to extract a mention of some kind of pastry in the phrase *These vegan recipes are a piece of cake*. In machine translation (Shutova, 2011) and textual entailment (Agerri, 2008) tasks, similar examples can easily be given as well. Tasks which can potentially be solved better when metaphors are correctly recognized are numerous. In particular, (Thibodeau and Boroditsky, 2011) even analyze the role of metaphor in reasoning about social policy on crime.

Our research problem results directly from the very well-known fact that language expressions can be interpreted literally i.e. their meaning can be a composition of the meaning of their parts; or metaphorically, when either the meaning of some words or combination of them is not interpreted literally.

Let us illustrate this in the Polish language on multiple phrases with an adjective *żelazny* '(to be made of) iron'. The expression e.g. *żelazny uchwyt* 'iron grip' can denote just a grip/handle which is made of iron, but it can also describe a feeling of fear and intimidation. The chances of these two interpretations are not equal for all expressions. With some of them, e.g. *żelazna krata* 'iron grille' it is hard to imagine when they get a figurative, non-literal meaning – they are strictly compositional – while others, e.g. *żelazne nerwy* 'iron nerves' are only used in the figurative, non-literal meaning. Identification of potentially figurative usages may improve the performance of many NLP applications. Although the ultimate goal is to decide whether each phrase occurrence could be interpreted compositionally (literally) or not, such task requires annotated data which is quite hard to prepare. In this work, we concentrate on the initial classification of isolated adjective-noun (AN) phrases – we try to categorize Polish phrases built up from a noun and a modifying adjective into these three categories, i.e. phrases which are almost certainly interpreted literally (L), phrases which only have a metaphorical meaning (M) and phrases which occur in both interpretations (B).

Although we apply this categorization in Polish, it may as well be used for other languages. For example, in English the phrases 'dirty hands' may be used literally and figuratively and qualify as B.

Proceedings of the Workshop on Figurative Language Processing, pages 27–33
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

2 Related Work

The problem of recognizing the metaphoricity of isolated phrases has been considered as a research topic in several papers. Almost all authors focus on phrases which are only literal or metaphorical and neglect phrases that represent both senses.

Gutierrez et al. (2016) address recognition of the metaphorical and literal meaning of adjective-noun phrases on the basis of metaphorical or literal senses of the adjective. Their approach was based on the model proposed in (Baroni and Zamparelli, 2010) to represent the vector of an unseen adjective-noun phrase **p** as a linear transformation given by a matrix $\mathbf{A}_{(a)}$ of an adjective a over a noun vector **n**:

$$\mathbf{A}_{(a)}\ \mathbf{n} = \mathbf{p}$$

They represent various (literal or metaphorical) senses of an adjective as two different matrixes: $\mathbf{A}_{LIT(a)}$ and $\mathbf{A}_{MET(a)}$, as in (Kartsaklis and Sadrzadeh, 2013). Gutierrez et al. (2016) assume that the literal or metaphorical meaning of the adjective, that is part of an AN phrase, makes the phrase literal or metaphorical, so they represent each literal adjective-noun phrase p_i containing adjective a as:

$$\mathbf{A}_{LIT(a)}\ \mathbf{n}_i = \mathbf{p}_i$$

and each metaphorical phrase i as:

$$\mathbf{A}_{MET(a)}\ \mathbf{n}_i = \mathbf{p}_i$$

The vectors of whole phrases and nouns can be extracted from a corpus, so the goal is to learn adjective matrices: literal ($\hat{\mathbf{A}}_{LIT(a)}$) and metaphorical ($\hat{\mathbf{A}}_{MET(a)}$) separately. To test the method, they prepared a very peculiar dataset consisting of 3991 literal and 4601 metaphorical AN phrases for only 23 adjectives, so it contained an average 370 phrases per each adjective. The requirement of many examples per adjective is crucial in this method and simultaneously difficult to obtain — at least if we want to take phrases with more than a dozen occurrences in texts used for creating vector representation into account. The best result reported by the authors was 0.809 accuracy (ACC).

Tsvetkov et al. (2014) applied a random forest classifier to detect metaphorical and literal AN phrases. Classifiers included in the ensemble were trained on the basis of three features, abstractness and imageability of nouns, supersenses, and vector-space word representation. Information about abstractness and imageability originated from the MRC psycholinguistic database (Wilson, 1988); as the database is not big, they propagated this information to other words based on vector representation. Supersenses for a noun were obtained from the WordNet as a combination of the supersenses of all synsets to which the noun belongs. Adjectives are classified into 13 supersenses adapted from GermaNet, but the information necessary for it was taken from the WordNet. To prepare vector space representation the authors used a variation of latent semantic analysis. To evaluate the method, they prepared training data consisting of 884 metaphorical AN phrases and the same number of literal phrases. The data contains phrases with 654 adjectives, so an average of 2.7 phrases per adjective. Furthermore, they collected a test set consisting of 200 phrases (100 phrases per each type) with 167 adjectives from the train set and 33 new ones. The data does not include weak metaphors and phrases which can have both interpretations. The method achieved ACC = 0.86.

Shutova et al. (2016) used word and visual embeddings to represent phrases and their components in order to detect metaphorical usage. They adopted the cosine similarity of embedding vectors as the measure of metaphoricity and postulated that the similarity is lower for metaphorical expressions. A threshold needed for classification was fixed on the basis of development data. For data from (Tsvetkov et al., 2014), the authors reported F1-measure equal to 0.79 (an accuracy is not given). A similar approach is described in the paper (Rei et al., 2017), where the authors improved the idea of Shutova et al. (2016) applying deep learning to establish the threshold. The evaluation performed on the same data indicated an accuracy of 0.829 and the F1-measure equal to 0.811, which is better than the original solution.

Bizzoni et al. (2017) proposed detecting the metaphoricity of AN phrases on the basis of word vectors only. They tested several configurations of single-layered neural networks to classify AN phrases into two groups: metaphorical and literal. They didn't use any additional knowledge except Word2Vec trained on Google News (Mikolov et al., 2013). The different configuration of neural networks was tested on the data from (Gutierrez et al., 2016), described above. The method achieved an accuracy of 0.915 when trained on 500 phrases and 0.985 when trained on 8000 phrases. Simultaneously, Wawer and Mykowiecka

(2017) proposed a similar approach to the problem of metaphoricity detection for Polish data. The authors noticed that detection of metaphorical and literal senses of phrases is not enough, and proposed classification into three types of AN phrases: literal metaphorical and phrases which occur in both interpretations (B). For this task, they reported an accuracy of 0.7, but the task is more difficult.

3 Polish Data

We prepared data containing Polish adjective-noun phrases divided into three classes. We distinguished literal (L) and metaphorical (M) phrases as in the English experiments mentioned in Section 2. Similar datasets for English excluded weak metaphors and phrases with both literal and metaphorical senses like *drowning students* (Tsvetkov et al., 2014). In our data, phrases with both meaning (B) made up the third class, we excluded only phrases that may have both senses but a literal (or metaphorical) one is not represented in NKJP (National Corpus of Polish, (Przepiórkowski et al., 2012)). An example of such phrase is *dobry pasterz* 'good shepherd' for which we were not able to find literal meaning in the corpus.

We collected 2380 adjective-noun phrases containing 259 different adjectives, so, an average 9.18 phrases per adjective. The adjectives were manually assigned to 55 classes (typology designed for this experiment) which represent such notions as: emotions, quantity, dimension, shape, colour, etc. Among the nouns we distinguished only two classes: abstract and concrete. We did not follow WordNet typology here (e.g. hyperonymy) as too elaborate and difficult to apply.

The dataset is an extension of the resource described in (Wawer and Mykowiecka, 2017). The process of data collecting was carried out in several steps. First, we prepared a list of 440 metaphorical phrases and collected literal and more metaphorical phrases containing the same adjectives from the frequent phrases in NKJP (National Corpus of Polish, (Przepiórkowski et al., 2012)). It resulted in the collection of many phrases for each adjective. The most numerous group, 79 phrases, was collected for the adjective *czarny* 'black', it consists of 45 literal, 27 metaphorical phrases and only 7 phrases of both types (phrases of B type are rarer then literal and metaphorical ones). In order to improve the participation of B phrases in our data we looked for them in dictionaries and added them if they occurred a dozen times in our texts. Moreover we added literal and metaphorical phrases for adjectives included in the new B phrases. The obtained list of phrases was evaluated by two annotators and inconsistencies were discussed in a larger group of annotators. Table 1 contains detailed information about numbers of different types of phrases for adjective domains for which more than 20 examples were collected.

phrase type	adjectives	M	L	B
all phrases	259	1034	1018	328
physical feature	21	185	115	36
dimension	11	147	131	38
color	12	61	182	36
material	16	42	79	15
luminosity	5	48	42	15
sense	18	71	20	13
temperature	4	40	49	13
tidiness	4	56	21	7
empty/full	2	58	22	2
animal	22	32	27	23
emotion	13	28	25	11
good/bad	2	17	24	15
society	24	23	23	8
sequence	2	1	41	11
body/mind f.	7	32	12	0
space orientation	5	0	29	12
sound	5	22	10	4
life/death	4	20	8	1
strength/weakness	2	18	9	1
civilization	8	10	17	1
weather	5	18	3	6
truth false	2	4	20	3
condition	4	2	9	14
easy/difficult	1	6	16	1
freedom	2	11	8	3
terrain stability	3	10	6	5
...				
other 29 domains	55	72	70	34

Table 1: Number of phrases

In order to implement experiments, we used distributional semantic models (DSM) created by Word2vec from the gensim package (Řehůřek and Sojka, 2010) and described in (Mykowiecka et al., 2017) and avilable from `http://zil.ipipan.waw.pl/CoDeS`. As Polish is a highly inflectional language, we decided to use models based on lemmas. We used the Continuous Bag of Words (CBOW) architecture. As a learning strategy, we selected negative sampling in the standard configuration of 5 positive examples and 1 negative. Models were prepared on the basis of NKJP (general corpus of Polish) and a dump of Polish Wikipedia from 2016. Two models based on 300 or 100 dimensions were used in our experiments; one consisted of all data, while the second was limited to words occurring no fewer than 50 times for NKJP data or no fewer than 30 times for Wikipedia data.

4 Experiments Description

In our experiments, we adopted the method described in (Wawer and Mykowiecka, 2017) as a starting point. The authors applied neural networks to predict if a phrase has a literal or metaphorical sense or can have both senses depending on its usage. Word embeddings of phrase components are the input to the network. The task consists in classifying of the input phrases into three groups: L, M, and B types. Our aim was to test the method on bigger and better balanced data. We also tested not only dense neural architecture but also a sequential one, namely LSTM. The sequence in our case is a short one, consisting of two words.

Moreover, we wanted to test the impact of the type of adjective and noun on the results. To compare the results for Polish with similar experiments for English, we also performed experiments on the literal and metaphorical phrases alone. In the latter case, we eliminated B type phrases from the input data. The architecture of the network is given in Figure 1. In the task of classification into L, M, B types, the output layer consists of three instances referring to three labels.

The impact of the type of adjectives and nouns was tested by extending appropriate word embeddings with additional features.

5 Results for Polish

In this section, we describe the results obtained for Polish phrases for different parameters. In all experiments, we performed 10-fold cross-validation (shuffling each time the entire set, the standard sklearn procedure resulted in a slightly different total number of phrases tested). The results were collected and the average results are given for precision, recall, F1-measure and accuracy.

Although the classification of adjective-noun phrases into M, L, B types is consistent with the linguistics reality, similar studies relating to English neglect phrases which may have both literal and metaphorical meanings. So, initially, we removed phrases annotated as B types from the data and performed the experiments with classification into two types only.

	2 dense layers, vec. size 100					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.88	0.88	0.88	
		20	0.89	0.87	0.88	
L	1017	10	0.88	0.88	0.88	
		20	0.87	0.89	0.88	
avg.	2047	10	0.88	0.88	0.88	0.879
		20	0.88	0.88	0.88	0.878

	3 dense layers, vec. size 100					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.89	0.86	0.87	
		20	0.90	0.87	0.88	
L	1017	10	0.86	0.89	0.88	
		20	0.87	0.90	0.89	
avg.	2047	10	0.88	0.88	0.88	0.876
		20	0.88	0.88	0.88	0.884

Table 2: Input: only embeddings, vectors 100

In Tables 2 and 3, we can see that the size of vectors, the tested number of epochs and choosing either 2 or 3 dense layers do not seem to have a great influence on the results. Thus, we tested the influence of a separate addition of domain of adjectives and type of noun only for models with a vector of size 300 and 3 dense layers (Table 4). Next, we tested adding both noun type and adjective domain again on all the variants as used in experiments reported in Tables 2 and 3, the results are given in Tables 5 and 6. In all these cases, we see only very small differences in F1 and accuracy. It turned out that on average, the simplest model with embeddings of size 100, 2 dense layers and no additional information is almost identically good as the model with embeddings of size 300, 3 dense layers and additional information consisting of adjective domain and binary noun type. Training nets for an additional 10 epochs did not im-

	2 dense layers, vec. size 300					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.90	0.85	0.87	
		20	0.90	0.87	0.88	
L	1017	10	0.85	0.91	0.88	
		20	0.87	0.90	0.88	
avg.	2047	10	0.88	0.88	0.88	0.888
		20	0.88	0.88	0.88	0.884

	3 dense layers, vec. size 300					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.90	0.85	0.87	
		20	0.90	0.87	0.89	
L	1017	10	0.85	0.91	0.88	
		20	0.88	0.91	0.89	
avg.	2047	10	0.88	0.88	0.88	0.876
		20	0.89	0.89	0.89	0.889

Table 3: Input: only embeddings, size of vectors 300

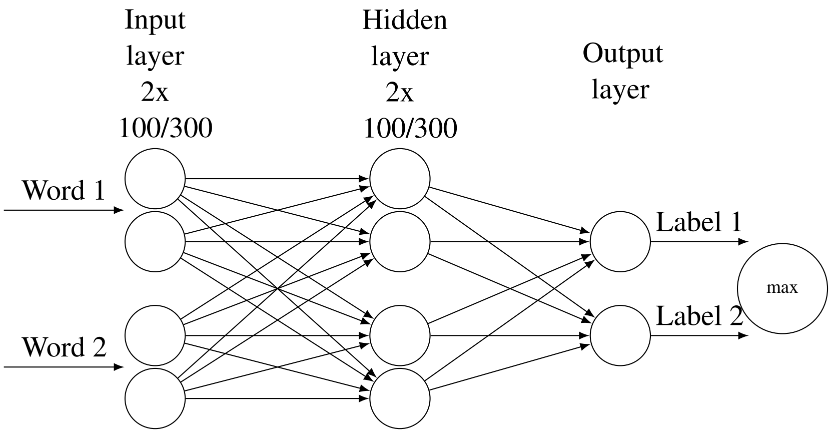

Figure 1: Net architecture for L and M phrases classification

prove the results significantly.

	3 dense layers, vec. 300, 20 epochs				
	nb	P	R	F1	acc.
	adjective domains				
M	1030	0.90	0.88	0.89	
L	1017	0.88	0.89	0.89	
avg.	2047	0.89	0.89	0.89	0.885
	noun type				
M	1030	0.91	0.87	0.89	
L	1017	0.87	0.91	0.89	
avg.	2047	0.89	0.89	0.89	0.889

Table 4: Input: word embeddings, type of noun or adjective domain

	2 dense layers, vec. size 100					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.90	0.87	0.88	
		20	0.89	0.87	0.88	
L	1017	10	0.87	0.91	0.89	
		20	0.87	0.89	0.88	
avg.	2047	10	0.89	0.89	0.89	0.886
		20	0.88	0.88	0.88	0.880

	3 dense layers, vec. size 100					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.88	0.87	0.88	
		20	0.90	0.87	0.88	
L	1017	10	0.87	0.88	0.88	
		20	0.87	0.90	0.89	
avg.	2047	10	0.88	0.88	0.88	0.876
		20	0.89	0.89	0.89	0.886

Table 5: Input: word embeddings, adjective domain, type of noun (abstract/concrete)

The same architecture was used to classify phrases into three groups. Table 7 shows the results for classification of all the data into literal, metaphorical and both type phrases; the input data consists of word embeddings of 300 dimensions (the results for 100 vectors are slightly lower – F1 for B class is equal to 0.48). The results for the B phrases are much lower than for L and M phrases. Adjective domains and abstractness do not improve the results, see Table 8.

	2 dense layers, vec. size 300					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.90	0.90	0.89	
		20	0.89	0.89	0.89	
L	1017	10	0.87	0.90	0.89	
		20	0.89	0.89	0.89	
avg.	2047	10	0.88	0.88	0.88	0.883
		20	0.89	0.89	0.89	0.890

	3 dense layers, vec. size 300					
	nb	ep.	P	R	F1	acc.
M	1030	10	0.90	0.88	0.89	
		20	0.89	0.88	0.88	
L	1017	10	0.89	0.90	0.89	
		20	0.88	0.89	0.88	
avg.	2047	10	0.89	0.89	0.89	0.890
		20	0.88	0.88	0.88	0.884

Table 6: Input: word embeddings, adjective domain and type of noun (abstract/concrete), vectors 300

6 Results for English Data

As it is difficult to compare methods applied on different data, we decided to use our method on data available on the Internet and compare it with the results reported in papers. The available resources contain only literal and metaphorical phrases. We tested two sets of such data. The first one was originally used in (Tsvetkov et al., 2014) – the solution described in Section 2 and the data is available from `https://github.com/ytsvetko/metaphor`. The train set consists of 884 metaphorical phrases and 884 literal ones, and

		3 dense layers, 20 epochs			
	nb	P	R	F1	acc.
M	1030	0.82	0.86	0.84	
L	1017	0.80	0.78	0.79	
B	328	0.52	0.47	0.49	
avg.	2374	0.77	0.77	0.77	0.773
		LSTM, 2 layers, 10 epochs			
	nb	P	R	F1	acc.
M	1030	0.84	0.86	0.85	
L	1017	0.81	0.82	0.82	
B	328	0.52	0.46	0.49	
avg.	2374	0.78	0.79	0.79	0.789

Table 7: Polish phrases classification into M, L and B; 300 dimennsions vectors

		LSTM, 2 layers, 10 epochs			
	nb	P	R	F1	acc.
M	1030	0.83	0.85	0.84	
L	1017	0.80	0.82	0.81	
B	328	0.48	0.40	0.44	
avg.	2374	0.77	0.78	0.77	0.778

Table 8: Polish phrases classification into M, L and B. Input: 300 dimensions word embeddings, adjective domain and type of noun

the test set has 100 phrases of each type. In our experiment, we used 300 element pre-trained GLoVe vectors trained on Wikipedia 2014 and Gigaword 5 (Pennington et al., 2014). We neglected to add information on adjective domains to directly test the solution based only on distributed word representation. Our results for both dense and LSTM architectures are given in Table 9. Tsvetkov et al. (2014) reported in their paper an accuracy of 0.86, which is a little higher than our result – 0.84. The same data was used in (Rei et al., 2017) where the authors reported an accuracy of 0.829 and for metaphor detection precision: 0.903, recall: 0.738 and F1-measure: 0.811. Our overall slightly better result (in comparison to (Rei et al., 2017)) is due to better recall for metaphorical phrases.

The second data set chosen was that prepared by (Gutierrez et al., 2016). The results of our experiments are reported in Table 10. In this case, the accuracy obtained by the network with one hidden dense layer was equal to 0.969 (between the results given in (Bizzoni et al., 2017)). This significant increase is due to the much smaller number of different adjectives and the larger number of phrases with the same adjective in this data set.

7 Conclusions

Information included in standard word embeddings makes it possible to differentiate between literal and metaphorical adjective-noun phrases, both in Polish and English. It seems that not using the cosine measure of vector similarity for metaphors detection (as discussed in Section 2), but applying a neural network to this problem is a good solution.

	nb	P	R	F1	acc.
Dense, 20 epochs, 10-times cross validation					
M	882	0.87	0.86	0.86	
L	871	0.86	0.87	0.86	
avg.	1753	0.86	0.86	0.86	0.864
LSTM, 20 epochs, 10-times cross validation					
M	882	0.86	0.86	0.85	
L	871	0.86	0.85	0.85	
avg.	0.86	0.86	0.86	0.855	
Dense, 20 epochs, test data					
M	100	0.90	0.72	0.80	
L	100	0.77	0.92	0.84	
avg.	200	0.83	0.82	0.82	0.819
GRU, 2 hidden layers, 20 epochs, test data					
M	100	0.90	0.78	0.83	
L	100	0.81	0.91	0.85	
avg.	200	0.85	0.84	0.84	0.845
LSTM, 2 hidden layers, 20 epochs, test data					
M	100	0.90	0.76	0.83	
L	100	0.79	0.92	0.85	
avg.	200	0.85	0.84	0.84	0.84

Table 9: Our results for Tsvetkov et al. (2014) data

Dense, 20 epochs, 10-times cross validation

	nb	P	R	F1	acc.
M	4596	0.96	0.97	0.97	
L	3991	0.96	0.97	0.97	
avg.	8587	0.97	0.97	0.97	0.969

Table 10: Our results for (Gutierrez et al., 2016) data

For the tested network architectures the accuracy varies between 0.81 and 0.97 depending on the character and size of the training set. The effect of using sequential architecture (GRU or LSTM units) is not straightforward: it improves results on the training/test set scenario, but not in the case of cross-validation setting.

Surprisingly, the adjective domain and the information on noun concreteness do not seem to have any significant influence on the results.

Recognizing phrases which can have either literal or metaphorical meaning (depending on the context) is much harder. The best F1 result for these phrases is at a level of 0.49. The overall results for recognition of the three labels (L, M and B) are lower by 0.11 than the results for recognition of just L and M cases. Still the result of 0.77 could be of practical use.

In the future, we plan to focus on phrases that have both literal and metaphorical usages (B) and recognize their usage on sentence level. Although the recognition of a type of phrase considered in isolation cannot be fully reliable, we think that the obtained results can be used as the additional source of information for phrases which are less frequent in text.

Acknowledgments

This work was supported by the Polish National Science Centre project 2014/15/B/ST6/05186.

References

Rodrigo Agerri. 2008. Metaphor in textual entailment. In *Coling 2008: Companion volume – Posters and Demonstrations*, pages 3–6.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yuri Bizzoni, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. 2017. "deep" learning : Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52. Association for Computational Linguistics.

Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of ACL 2016 (short papers)*.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2013. Prior Disambiguation of Word Tensors for Constructing Sentence Vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1601. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Piotr Rychlik. 2017. Testing word embeddings for Polish. *Cognitive Studies / Études Cognitives*, 17:1–19.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 717–727, Prague, Czech Republic. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

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.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546. Association for Computational Linguistics.

Ekaterina Shutova. 2011. *Computational Approaches to Figurative Language*. Ph.D. thesis.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *HLT-NAACL*. The Association for Computational Linguistics.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLOSone*, 6(2).

Yulia Tsvetkov, 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*, pages 248–258. Association of Computational Linguistics.

Aleksander Wawer 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*, pages 772–777, Varna, Bulgaria. INCOMA Ltd.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, and Computers*, 20(1):6–10.