Detecting Metaphorical Phrases in the Polish Language

Aleksander Wawer **Agnieszka Mykowiecka**
Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5
01-248 Warsaw, Poland
{*axw*, *agn*}@ipipan.waw.pl

Abstract

In this paper we describe experiments with automated detection of metaphors in the Polish language. We focus our analysis on noun phrases composed of an adjective and a noun, and distinguish three types of expressions: with literal sense, with metaphorical sense, and expressions both literal and methaphorical (context-dependent). We propose a method of automatically recognizing expression type using word embeddings and neural networks. We evaluate multiple neural network architectures and demonstrate that the method significantly outperforms strong baselines.

1 Introduction

Language expressions can be interpreted literally or metaphorically, e.g. *round table* is just a table which is round, but it can also describe a way of organizing a discussion. The chances of these two interpretations are not equal for all expressions. With some of them, e.g. *zielony długopis* 'green pen' it is hard to imagine when they get figurative meaning – they are strictly compositional – while others, e.g. *biały szum* 'white noise' are used only in figurative meaning. There is also a third group of phrases (to which *round table* belongs) used both literally and metaphorically. Identification of potentially figurative usage may improve the performance of many NLP applications. It is crucial for information extraction task as literal interpretation of metaphors can lead to incorrect results (Patwardhan and Riloff, 2007), machine translation (Shutova, 2011) and textual entailment (Agerri, 2008). In particular, Thibodeau and Boroditsky (2011) even analyze the role of metaphor in reasoning about social policy on crime. Although the ultimate goal is to decide on every phrase occurrence whether it could be interpreted compositionally (literally) or not, such task requires annotated data which are quite hard to prepare. In this work we concentrate on the initial classification of isolated phrases – we try to categorize Polish phrases build up from a noun and a modifying adjective into these three categories, i.e. phrases which are nearly for sure interpreted literally (L), phrases which have only metaphorical meaning (M) and phrases which occur in both interpretations (B). We test our approach on a set of about 500 phrases selected (mainly) from the top frequent phrases of Polish National Corpus, NKJP, (Przepiórkowski et al., 2012), and manually categorized into these three classes. We tested methods which do not require to build vectors of the analyzed phrases. Our models use only vectors representing phrase constituents.

2 Existing Work

Discrimination of literal and metaphorical (compositional and non-compositional) phrases is not a new idea. First attempts to solve this problem use different type of measures. Lin (1999) compared the mutual-information measures of the constituents with the mutual information of similar expressions obtained by substituting one of the elements with a related word, while Schone and Jurafsky (2001) evaluated a number of co-occurrence based measures. Both approaches did not give satisfactory results. In many later works distributional models were used. Baldwin et al. (2003) showed that LSA-based similarity between the multiword expression and each of its components is indicative for compositionality. Katz and Giesbrecht (2006) compared the actual phrase vector to the estimated compositional meaning vector calculated as a sum of the meaning vectors of the

Proceedings of Recent Advances in Natural Language Processing, pages 772–777,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_099

parts. The hypothesis was that the similarity between these two vectors should be larger in case of phrases which are not used non-compositionally. The test set contained 81 potential German multi word (preposition-noun-verb) collocation candidates) from a database described in Krenn (2000). When the threshold of similarity value was set to 0.2, the method achieved F-measure of 0.48. The results of (Baldwin et al., 2003) method on the same set were 0.16 for verbs and 0.51 for nouns. Several other solutions consist in training classifiers: Tsvetkov et al. (2014) used abstractness, imageability (whether or not a word can be easily associated with image), wordnet-like, but specifically defined, supersenses, and word vectors, in random forest classifier to distinguish metaphoric and literal adjective-noun pairs. On the set of 884 metaphoric and 884 literal phrases and 360 features, they obtained results with 0.86 accuracy (in 10-fold cross validation). Hovy et al. (2013) employed the idea of selectional preference violation as the indicator of metaphor and trained an SVM classifier with tree kernels to capture compositional properties of metaphorical language. Their hypothesis is that unusual semantic compositions in the data may be indicative of the use of metaphor. They trained the model on labeled examples of literal and metaphorical uses of 329 words (3872 sentences) using word vectors, part of speech tags and WordNet supersenses introduced into dependency trees as features and obtained F-score=0.75. The detailed comparison of these earlier methods can be found in (Shutova, 2015).

A recent attempt in metaphor analysis using compositional semantic method is presented in Gutierrez et al. (2016). The authors classify 8592 adjective-noun pairs as either metaphoric or literal. They build compositional DS models in which adjectives are treated as linear maps from nouns to AN phrases. For each adjective they split the phrases involving that adjective into two subsets, the literal subset and the metaphorical subset and build three models representing literal, metaphorical and both types of usage of every adjective from the traininig phrases. An unseen phrase involving a known adjective is classified as metaphorical if the vector representing the phrase is more similar to the vector representing the same phrase with the adjective replaced by its metaphoric usage representation than to the vector obtained for its literal usage. The resulting F-measure and accuracy are 0.79 and 0.81 respectively.

As the state-of-the-art methods of contextual recognition of figurative phrases we can cite Peng and Feldman (2017) who explored the idea that idioms and their non-idiomatic counterparts do not appear in the same contexts and that the words which are representatives of the local context are likely to associate strongly with a literal expression. The association was measured as in terms of projection (inner product) of word vectors onto the vector representing the literal expression. For different data sets they achieved the accuracy of 0.57 to 0.87. However, in their work, phrases are recognized in sentential context and the analyze concerns verb-noun pairs.

3 Training and Test Data

Annotation of phrases with the type of their usage was done specially for this experiment. A list of 437 figurative expressions of the form adjective-noun were composed from two different sources. About 100 phrases were proposed by several project members being native speakers of Polish. This list was then enriched with the examples manually selected from the top of the frequency list of the adjective+noun phrases occurring in NKJP. Next, all phrases were verified by a linguist who classified them into two categories: M – phrases which are only used metaphorically, e.g. *barwna historia* ‘colourful story’ and B – phrases which can be used both in literal and metaphorical senses, like *biała karta* ‘white page’ which in Polish can mean that we start from the beginning without any judgments on the previous work or behaviour or just a page which is not covered with text. Only one phrase from this set was eventually classified as used only literally (*linia autobusowa*) ‘bus line’. The phrases are also annotated with information on the domain which is described by an adjective, e.g. for *chłodne oko* ‘cool eye’ the adjective domain is *temperature*, and the type of a noun (abstract or specific). The second list contain adjective-noun phrases which have only literal meaning (at least in not very awkward situations) and they include the same adjectives as phrases on the first list. These phrases were also manually selected from the top of the frequency list of NKJP. Their type was verified by the second annotator and only phrases on which both of them

agreed are included in the final list. In total, our data comprises 282 phrases with only metaphorical meaning, 1041 with literal meaning and 154 phases which can have both types of usage.

4 Data Analysis

In order to obtain better understanding of our data set, we analyzed properties of adjectives and nouns included in phrases of all types.

In the case of nouns, we manually annotated whether the noun is concrete or abstract (two possible values). In the case of adjectives, we manually annotated its domain. By *domain* we understand the root of hypernymy tree for an adjective (eg. for "red", the root hypernym is "color"). However, we did not use any existing typology, but the annotators (linguists) proposed semantic groupings into which words with more specific meanings fall, not necessarily based on hyperonymy in the strict sense (eg. "sensual experience" for taste such as "bitter", "dimension" for "high"). The number of the topmost classes in this hierarchy is 47.

In the two sections below we analyse whether adjective and noun types are related to metaphorical use of a phrase (each constituted of a verb and a noun).

4.1 Adjective Types

Adjective type appears to have an influence whether the phrase is metaphorical. To confirm this, we compute $\tilde{\chi}^2$ statistics ($\tilde{\chi}^2$=151 with 96 degrees of freedom and p=0.0002), therefore we conclude that the relationship between adjective domain and metaphorical character of the phrase is significant. Table 1 illustrates frequencies of selected adjective types in literal (L), metaphorical (M) and both metaphorical and literal (B) phrases.

4.2 Noun Types

Turney et al. (2011) prove that metaphorical word usage is correlated with the degree of abstractness of the word's context. As we analyse isolated phrases we cannot use information about the context, but we observe similar relation between the abstractness of the noun and the metaphorical usage of the phrase to which it belongs. Table 2 shows frequencies of phrase types and noun types. As could be expected, the observation to be made is that in metaphorical phrases nouns tend to be abstract, while those used in both metaphorical and literal and only literal phrases are more often concrete.

	L	B	M
good/bad	25	5	0
sound	11	3	7
emotions	27	4	5
order	42	6	0
colour	185	22	29
material	77	13	13
state of body/mind	12	0	12
temperature	50	10	22
dimension	137	10	27
physical property	25	14	24
supernatural phenomenon	0	8	3
weather phenomena	1	1	9
sensual experience	21	7	50

Table 1: Selected adjective types and metaphorical phrases

	abstract		concrete
M	193	(0.68%)	89
B	41	(0.27%)	112
L	253	(0.24%)	789

Table 2: Noun types and metaphorical phrases

5 Network Architectures

The observations made in previous section allow to hypothesize that certain information influencing the metaphorical character of a phrase, namely adjective type and noun type (such as for instance relation to sensual experiences of an adjective and high abstractness of a noun), can be contained in word embeddings trained on large corpora. For this reason we try to predict metaphorical character of each phrase using word embeddings provided as input to neural network models. In the following sections we describe our experiments with this approach.

We attempt to recognize the type of a phrase (as literal, metaphorical, and both) consisting of an adjective and a noun using several types of neural network structures. As word embeddings we used word2vec vectors trained on a dump of Polish language Wikipedia and NKJP. All word2vec parameters had default values as in (Řehůřek and Sojka, 2010).

5.1 Multiplicative

This model implements Marco Baroni et al. hypothesis that adjectives act as functions on nouns (Baroni et al., 2014). In this view, computing meaning of a noun phrase is based on multiplication of noun embeddings by adjective embeddings (functions). Success of this idea might depend on training adjective and noun embeddings according to different objectives, which obviously is not the case for word2vec embeddings.

We experiment with three set-ups of this idea, each with different sets of weight vectors. Let adj-v denote the embedding of an adjective, noun-v the embedding of a noun, and w, w-1 and w-2 trainable weight layers. In each case, the presented formulas were followed by multiplication by trainable softmax weights layer with bias weights. The size of all weight vectors and softmax weight vector was equal to word embedding size.

The three implementations we tested are as follows:

- M1: adj-v * noun-v
- M2: adj-v * w * noun-v
- M3: adj-v * w-1 * noun-v * w-2

All of these architectures have been implemented in TensorFlow.

5.2 Concatenated

In this type of models, embeddings of an adjective and a noun were concatenated, and this concatenated vector was subsequently passed to neural network layers. By dense we denote a regular layer of weights (called also dense).

Implementations we tested were as follows:

- C1: dense → softmax
- C2: dense → dense → softmax

In the case of these architectures, C1 was implemented in TensorFlow, while C2 in Keras.

6 Results

Because of overwhelming majority of one class (those with literal meaning) we compare our methods to the most frequent class baseline. The accuracy of this baseline can be computed as 0.70. We perform the experiments on all 1457 phrases in our data set, evaluating each combination of parameters in a 10-fold cross-validation.

Table 6 contains results of evaluations of each neural network architecture as average micro precision (P) and recall (R) for each phrase type over 10 folds and three consecutive runs for each parameter combination. Reporting 'micro' values (e.g., precision P, recall R) means giving each observation (phrase) an equal contribution to the overall metric and is often preferred in multilabel settings, as in our case. We experimented with multiple embedding vector sizes. In each case batch size was equal to one.

		embedding size					
		50		100		200	
		P	R	P	R	P	R
M3	Both	0.32	0.16	0.27	0.25	N/A	N/A
	Lit.	0.84	0.91	0.85	0.86	N/A	N/A
	Met.	0.62	0.58	0.58	0.57	N/A	N/A
	All	0.74	0.77	0.74	0.74	N/A	N/A
C2	Both	0.29	0.27	0.29	0.29	0.31	0.31
	Lit.	0.87	0.88	0.87	0.88	0.87	0.91
	Met.	0.63	0.61	0.65	0.61	0.74	0.62
	All	0.76	0.77	0.77	0.77	0.79	0.79

Table 3: Average micro precision (P) and recall (R) in 10-fold cross-validation

In Table 6 we report results only for M3 and C2 architectures, each proven superior within their type.

The M1 architecture could not be successfully trained as the weights did not converge during learning. In the case of M1 and M2, between 10% and 30% of the models also did not converge. The same thing occurred to M3 models with embedding size 200 (marked as N/A in the table). Surprisingly, M3 model with embedding size 50 turned to have better recall than the one with embedding size 100, maintaining the same precision value. Therefore, embedding size appears to be reversely proportional to prediction quality in the case of M-type models.

Concatenative models C2 (with two dense layers) turned to be superior from C1 (one dense layer) by few percentage points in each measure. In the case of M-type models, embedding size appears to be directly proportional to prediction quality.

Generally it must be stated that concatenative models proved quite promising. Their quality increased with embedding size and the best model achieved the overall precision (and recall) of 0.79.

		corr.	ass.
biała gorączka	white fewer	M	B
biały kolor	whilte colour	L	M
ciężka atmosfera	heavy atmosphere	B	+
ciężki bagaż	heavy luggage	L	+
ciężka bitwa	heavy battle	M	B
ciężka doniczka	heavy pot	L	+
ciężka próba	ordeal	B	+
ciężka kłódka	heavy paddlock	L	+
ciężki konar	heavy bough	L	+
ciężka ręka	hard hand	B	+
ciężki wykład	heavy lecture	B	M
wściekły lis	rabid fox	L	L
wściekły upał	furious heat	M	M
zdrowy chłopiec	healthy boy	L	M
zdrowy psiak	healthy dog	L	L
zdrowy rozsądek	common sense	M	B

Table 4: Sample correct (+) and incorrect results.

7 Analysis of the Results

Incorrectly assigned labels were nearly equally frequent B, L and M tags. Mistakes are made in all directions. In Table 7 there are selected examples for test phrases with the adjectives of all test phrases with adjectives *biały* ''white', and *zdrowy* 'healthy' and all test phrases with the adjectives *ciężki* and *wściekły*. The first one means 'heavy' but is frequently used as 'difficult'. The latter is ambiguous and means in Polish both 'furious' and 'rabid'. Phrases with *ciężki* are classified quite well. The only severe error is for *heavy battle*. The other error is smaller as the phrase *heavy lecture* should be probably rather classified as M. The next group of two phrases are both correctly tagged while phrases while 'healthy dog' is labelled correctly and 'healthy boy' is not.

8 Conclusions and Future Work

We performed multiple experiments with automatic recognition of metaphorical expressions, noun phrases composed of a noun and an adjective. We divided those expressions into three types: strictly metaphorical, both metaphorical and literal (where actual meaning is determined by the context of usage), and finally strictly literal (that are not used in non-literal, metaphorical sense). The paper contains an analysis, supported by manual annotation, that demonstrates relationships between phrase type (falling into one of metaphorical classes) and types of involved nouns and adjectives.

We proposed to automatically recognize phrase types using word embeddings to represent word meaning. We described several experiments using selected neural network architectures. We predicted phrase type without using sentence contexts, only based on word embeddings of adjectives and nouns that constitute each phrase. Results significantly outperform strong baseline of the most frequent class.

In future we plan to focus on sentence-level, context-dependent detection of metaphorical phrases. This involves detecting when phrases of B type (contextually metaphorical) take their non-literal meaning. Also we plan on applying our models on large corpora to detect more phrases of B type than in our current data set.

Acknowledgments

The paper is partially supported by the Polish National Science Centre project *Compositional distributional semantic models for identification, discrimination and disambiguation of senses in Polish texts* number 2014/15/B/ST6/05186.

References

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

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*. Association for Computational Linguistics, Stroudsburg, PA, USA, MWE '03, pages 89–96.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology* 9:5–110.

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)*.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphoricalword use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*. Association for Computational Linguistics, pages 52–57.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. pages 12–19.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. DFKI-LT - Dissertation Series.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '99, pages 317–324.

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)*. Association for Computational Linguistics, Prague, Czech Republic, pages 717–727. http://www.aclweb.org/anthology/D07-1075.

Jing Peng and Anna Feldman. 2017. Automatic idiom recognition with word embeddings. In Juan Antonio Lossio-Ventura and Hugo Alatrista-Salas, editors, *Information Management and Big Data: Second Annual International Symposium, SIMBig 2015, Cusco, Peru, September 2-4, 2015, and Third Annual International Symposium, SIMBig 2016, Cusco, Peru, September 1-3, 2016, Revised Selected Papers*, Springer International Publishing, Cham, pages 17–29.

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*. ELRA, Valletta, Malta, pages 45–50.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*. Pittsburgh, PA.

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

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics* 41(4):579–623.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLOSone* 6(2). https://doi.org/10.1371/journal.pone.0016782.

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

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association of Computational Linguistics, pages 680–690.