Neural Architectures for Detecting Metaphorical Phrases

Anonymous RANLP **submission**

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* it is hard to imagine then 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 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. 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 and manually categorised into these three classes. We tested methods which do not require to build vectors of the analysed phrases. Out 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 a 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 similarity. 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 parts. The hypothesis was that the similarity between these two vectors should be larger in case of phrases which are not used non-compositionally. A 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.

To compare these results with 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 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 set they achieved the accuracy of 0.57 to 0.87.

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 the Polish National Corpus. 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 judgements on the previous work or behaviour or just a page which is not covered with text. One phrase was 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 the Polish National Corpus. 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 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 metaphorical phrases of types M and B.

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"). We did not use a WordNet. Instead, 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").

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). We conduct our analysis on all 436 phrases of types B and M in our data set. Those phrases were annotated with extra information about adjective and noun types.

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 metaphorical (M) and both metaphorical and literal (B) phrases.

4.2 Noun Types

Table 2 shows frequencies of phrase types and noun types. As could be expected, the observation to be made is that metaphorical nouns tend to be abstract, while those used in both metaphorical and literal phrases are more often concrete.

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

	B	M
good/bad	5	0
sound	3	7
emotions	4	5
order	6	0
colour	22	29
material	13	13
state of body/mind	0	12
temperature	11	22
dimension	10	27
physical property	14	24
supernatural phenomenon	8	3
weather phenomena	1	9
sensual experience	7	49

Table 1: Selected adjective types and metaphorical phrases

	abstract	concrete
M	198	84
B	41	113

Table 2: Noun types and metaphorical phrases

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 the National Corpus of Polish (Przepiórkowski et al., 2012). 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 $\rightarrow$ noun-v $\rightarrow$ softmax
- M2: adj-v $\rightarrow$ w $\rightarrow$ noun-v $\rightarrow$ softmax
- M3: adj-v $\rightarrow$ w-1 $\rightarrow$ noun-v $\rightarrow$ w-2 $\rightarrow$ softmax

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 $\rightarrow$ softmax
- C2: dense $\rightarrow$ dense $\rightarrow$ softmax

In the case of these architectures, D1 was implemented in TensorFlow, while D2 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 0.8109. We perform the experiments on all 1457 phrases in our data set, evaluating each combination of parameters in a 10-fold cross-validation.

Table 3 contains results of evaluations of each neural network architecture as average micro accuracy value over 10 folds and three consecutive runs for each parameter combination. Micro accuracy gives each observation (phrase) an equal contribution to the overall metric and is often preferred in multilabel settings as in our case. Average standard deviations have been reported in

parentheses, after accuracy values. We experimented with multiple numbers of training epochs and embedding vector sizes. In each case batch size was equal to one.

	embedding size		epochs
	50	100	
M2	N/A	0.818 (0.021)	4000
M3	N/A	0.795 (0.022)	4000
C1	0.786 (0.339)	0.825 (0.027)	1000
C1	N/A	0.841 (0.294)	2000
C1	N/A	0.842 (0.185)	4000
C2	N/A	0.873 (0.436)	2000

Table 3: Average accuracy (micro) in 10-fold cross-validation.

Generally, multiplicative models were not successful as their results did not differ significantly from the most frequent class baseline. We did not report M1 architecture as none of the models could 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. We report accuracy and standard deviation for the remaning ones. For multiplicative models, we had to set the number of epochs to higher values than in the case of concatenative models.

Concatenative models proved more promising. The double-layer D2 model reached as much as 0.87 accuracy, this hovewer with relatively higher variation between the folds.

7 Analysis of the results

About 14% of incorrectly assigned labels are either L or M (there is no incorrect B tags). Mistakes are made in all other directions. In Table 7 there are examples of all test phrases with two selected 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*. Two other errors are smaller as *heavy hand* is very rarely used in the literal sense and the second phrase *heavy lecture* should be probably rather classified as M. The second group of two phrases are both wrongly tagged.

		corr.	ass.
ciężka atmosfera	heavy atmosphere	B	+
ciężki bagaż	heavy luggage	L	+
ciężka bitwa	heavy battle	M	L
ciężka doniczka	heavy pot	L	+
ciężka próba	ordeal	B	+
ciężka kłódka	heavy paddlock	L	+
ciężki konar	heavy bough	L	+
ciężka ręka	hard hand	B	M
ciężki wykład	heavy lecture	B	M
wściekły lis	rabid fox	L	M
wściekły upał	furious heat	M	L

Table 4: Sample correct (+) and incorrect results. Phrases with two selected adjectives.

8 Conclusions and Future Work

We tested multiple experiments with automatic recognition of metaphorical expressions, noun phrases composed of a noun and a verb. 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.

References

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 Linguis-

tics, 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.

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.

Jing Peng and Anna Feldman. 2017. *Automatic Idiom Recognition with Word Embeddings*, 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.