Teasing Apart Russian Idioms And Homonymic Compositional Expressions

A Word Embedding Approach

Marina Pchelina¹ and Jae-Woong Choe²

^{1,2} Korea University, Seoul 02841, Republic of Korea
{marina_p, jchoe}@korea.ac.kr

Abstract. We test and evaluate a context-based method for MWEs compositionality detection that utilizes word embeddings. Embeddings for individual words are used to get representations of target expressions and their context. In making a judgement on compositionality/idiomaticity of an expression, our algorithm relies on the expectation that when a MWE is used literally constituents retain their original meanings and are semantically related to surrounding context words, which is not normally true of idiomatic usage. Context is the only factor that decision about compositionality of an expression is based on, which adds to simplicity and universality of the method. We test the recently introduced idea of applying Principal Component Analysis to represent semantic composition and argue that its performance is at least as good as the state of the art.

Keywords: Compositionality Detection, Word Embeddings, PCA.

1 Introduction

Non-literal uses of language pose a significant problem for many NLP tasks, especially when they look exactly the same as literal, consider expressions in bold in sentences 1) and 2).

- 1) *Jinny was so startled that she nearly **kicked the bucket** over.*
- 2) *Chatterton and Fagg and a few more like them who've since **kicked the bucket**.*

For efficient natural language understanding, there should be a way to automatically tell apart those two cases. Great attention has been paid to this issue and considerable results have been achieved, but the majority of proposed methods rely on hand-crafted

lexicons and databases. Such approaches tend to be restricted in terms of language and range of expressions they can detect.

We believe that it is local context that should be the main cue in deciding on compositionality of a given expression. We focus on solutions that are simple and general enough to cover broader types of non-compositional uses of expressions and, theoretically, to work on any language.

2 Method

The basic intuition behind our method is that when an expression is meant literally it is compositional in the sense that its constituents a) retain their original meaning and b) are usually semantically related to context words [1]. Therefore, if representation of a phrase used literally is compared to representation of its local context, they are expected to be similar, and in case of non-compositional usage, otherwise.

While representing individual words with embeddings has proved to be highly effective, computationally representing larger pieces of text is an ongoing issue, simple vector averaging being currently the state of the art. The newest idea is to draw on geometrical aspects of word vector spaces and apply Principle Component Analysis to represent context as a linear subspace [2].

The method, which we adopt here, consists in applying PCA to the linear space constituted by word embeddings and finding a linear subspace that the first few principle components create, so that the original data is represented in a compact way with minimal loss. The target expression's vector is then projected into the linear subspace, and that projection is compared with the original vector [5]. In all cases, we use cosine similarity to compare vectors and calculate threshold as the mean of all similarity scores in our data set.

3 Experiment Setup

It was important to look specifically at MWEs that could be used literally and idiomatically without notable bias toward any of the two and regardless of a particular grammatical form, for it seems those pose the most difficulty for NLP tasks. That is why they had to be picked out manually¹. We disregarded syntactic structure and length of the MWEs to check universality of the method, so the list included expressions of the form ADJ+N, PR+N, V+PR+N and others².

We had 50 target expressions, with one instance of compositional and one of non-compositional use for each, making it 100 cases overall. For context, 10-15 content words were considered from both sides of a MWE in question³. We obtained original 300-dimensional word embeddings from rusvectors.org, specifically, from the model

¹[We used wikitionary.org, russkiyyazik.ru.

²[Because only content words got embeddings, the number of embeddings per expression was 1-3.

³[Russian National Corpus was queried for expressions to obtain contexts [6].

trained with word2vec's CBOW algorithm on 900 mln words web corpus, because it performed best on simple word similarity tests [3].

4 Results and Discussion

In the first set of experiments, target expressions were represented as compound vectors calculated over all constituents with simple averaging, multiplication and PCA; context was represented⁴ with either averaging or PCA. Results are reported in Table 1.

Table 1. Accuracy with MWEs represented as compound vectors.

Phrase/context	PCA	Average
PCA	0.87	0.55
Average	0.49	0.89
Multiplication	0.79	0.52

It is immediately clear that better results are obtained when both expressions and context are represented in the same way, which apparently has to do with the fact that different composition methods provide different resulting vectors – averaging introduced negative similarity values, while PCA did not.

In the second set of experiments, we aimed to see if MWEs could be represented as an embedding of their single constituent with a minimum, maximum or furthest from the mean (extreme) score. Since individual words can have different degrees of idiomaticity within an expression, it would allow for fuller coverage of potentially idiomatic language, e.g. partly compositional expressions. Results are reported in Table 2.

Table 2. Accuracy with MWEs represented as single constituent embeddings.

Phrase/context	PCA	Average
Min	0.76	0.79
Max	0.88	0.86
Extreme	0.89	0.87

This time, scores seem to be more stable, with PCA being slightly more accurate than averaging. Overall, 17 phrases out of 50 (34%) were labelled correctly in all tests, a few examples are: *закинуть удочку* “throw the fishing rod”, *сидеть на чемодане* “sit on a suitcase”, *прижаться к стене* “press against the wall”⁵. As for the rest, where there were false predictions, overwhelmingly, they formed patterns.

⁴□ Multiplication was found not suitable for representation of larger chunks of language.

⁵□ Idiomatic meanings are, respectively, “make a cautious inquiry”, “be ready to leave at any moment”, “expose sb.”.

Firstly, relying on minimal similarity scores did not provide accurate enough results, which might be due to fact that it is fairly easy for two randomly taken words to have similarity score close to zero, which just means irrelevance. Judging by maximum scores gave better results. Secondly, scores obtained with PCA for context representation are higher than with averaging: mean accuracy scores after all experiments are 78 and 74.6, respectively. Finally, error analysis showed that wrong predictions were given in cases where target expressions either were used in a very peripheral role and did not have enough similar words around, or contained very frequent or general words that could not get distinctive embedding representations.

As for the limitations, however, precision scores in our experiments did not vary drastically, which might indicate that it is the embeddings themselves that should be improved, possibly with exemplar-based models, which allow multiple embeddings per word and have a potential to improve accuracy for compositionality detection [4].

5 Conclusion

We found the context-based test for telling apart compositional and non-compositional uses of same MWEs to be fairly effective while being potentially independent of a particular language. Additionally, our experiments hint that PCA could be considered as a suitable way of representing context, because it can take into account tens and hundreds of factors (here, words) and, at least on our dataset, it outperformed the state-of-the-art averaging. Another finding is that when potentially idiomatic expressions are not cherry-picked to be of the same kind, it might be reasonable to judge them by single constituent embeddings.

References

1. Doumas, L. A. A., Hummel, J. E.: Modeling human mental representations: What works and what doesn't and why. In: Holyoak, K. J., Morrison, R. G. (eds.) *The Cambridge handbook of thinking and reasoning*, pp. 73–91. Cambridge University Press, Cambridge, UK (2005).
2. Gong, H., Bhat, S., Viswanath, P.: Geometry of Compositionality. In: *AAAI Conference on Artificial Intelligence*, <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14699>, last accessed 2017/06/21.
3. Kutuzov, A., Kuzmenko, E.: WebVectors: a toolkit for building web interfaces for vector semantic models. In: Ignatov, D. et al. (eds.) *AIST 2016, CCIS*, vol 661, pp. 155-161. Springer, Cham (2017).
4. Reddy, S., McCarthy, D., Manandhar, S., Gella, S.: Exemplar-based word-space model for compositionality detection: Shared task system description. In: *ACL 2011 Workshop on Distributional Semantics and Compositionality*, pp. 54-60. ACL, Portland, USA (2011).
5. Widdows, D.: Semantic vector products: Some initial investigations. In: *Proceedings of QI 2008, 2nd International Symposium on Quantum Interaction*. College Publications, Oxford, UK (2008).
6. Russian National Corpus Homepage, <http://www.ruscorpora.ru>, last accessed 2017/06/15.