A Dataset for Noun Compositionality Detection for a Slavic Language

Dmitry Puzyrev †, **Artem Shelmanov**‡, **Alexander Panchenko**‡, and **Ekaterina Artemova**†

†National Research University Higher School of Economics, Moscow, Russia
‡Skolkovo Institute of Science and Technology, Moscow, Russia
dapuzyrev@edu.hse.ru, echernyak@hse.ru
a.{shelmanov, panchenko}@skoltech.ru

Abstract

This paper presents the first gold-standard resource for Russian annotated with compositionality information of noun compounds. The compound phrases are collected from the Universal Dependency treebanks according to part of speech patterns, such as ADJ+NOUN or NOUN+NOUN, using the gold-standard annotations. Each compound phrase is annotated by two experts and a moderator according to the following schema: the phrase can be either compositional, non-compositional, or ambiguous (i.e., depending on the context it can be interpreted both as compositional or non-compositional). We conduct an experimental evaluation of models and methods for predicting compositionality of noun compounds in unsupervised and supervised setups. We show that methods from previous work evaluated on the proposed Russian-language resource achieve the performance comparable with results on English corpora.

1 Introduction

The quality of many natural language processing applications is heavily dependent on the quality of vector representations of text elements. The streamline NLP research encompasses many works on building various distributional semantic models (DSMs), and on methods for combining vector representations of atomic elements like words into representations of bigger fragments: phrases, sentences, texts. A simple but strong baseline for this task suggests averaging word embeddings of a text fragment (sometimes weighted, e.g., according to IDF). Although the result vector representation is rough compared to results could be achieved by more elaborate neural network encoding methods, it was shown that this baseline has high performance in many tasks (Weston et al., 2013; Mikolov et al., 2013; Mitchell and Lapata, 2008; Anke and Schockaert, 2018). The main advantages of such methods are computational efficiency and an ability to use them in an unsupervised setting, while neural encoders would commonly require heavy computational power, labeled datasets, and substantial time for training.

However, simple averaging of word embeddings often is too naïve. Idiomatic noun phrases are one of the cases where the averaging of the phrase parts would yield a wrong result since the meaning of such phrases is metaphorical and could not be directly "summed up" from meanings of its components. Therefore, it would be beneficial to have a DSM that tackles this problem, by having a distinct embedding for the whole phrase.

In this work, we focus on the task of predicting compositionality of noun phrases in Russian language texts. The goal is to develop a resource and methods for distinguishing compositional compounds, which meaning could be split into parts, from non-compositional ones that have a solid meaning, and for which we would like to have a dedicated embedding. The ability to detect compositionality for noun compounds is considered beneficial for many tasks including machine translation, semantic parsing, as well as word sense disambiguation.

The **contribution** of this paper is two-fold:

1. We present the first *gold-standard dataset* for Russian annotated with compositionality information of noun compounds.[1]

2. We provide an *experimental evaluation* of models and methods for predicting compositionality of noun compounds. We show that the methods from the previous work trained on the proposed Russian-language resource achieve the performance comparable with results on English corpora.

[1] https://github.com/slangtech/ru-comps

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 56–62,
Florence, Italy, 2 August 2019.

2 Related Work

The construction of datasets presenting compositionality can be traced back to as early as the 2000s: Baldwin and Villavicencio (2002) proposed chunk-based extraction methods for English verb-prepositional combinations and gave some binary judgments on the subject of considering them as phrasal verbs. In the follow-up paper, Baldwin et al. (2003) used the same framework to retrieve 1,710 Noun-Noun compounds from 1996 Wall Street Journal corpus. The authors use LSA to calculate the similarity between a phrase and its components as one of the early compositionality prediction attempts. McCarthy et al. (2003) evaluated 116 candidates of English phrasal verbs using three annotators' predictions on a scale from 0 to 10. Venkatapathy and Joshi (2005) used 800 verb-object collocations obtained from British National Corpus to give annotations from 1 to 6 where one stands for total non-compositionality and 6 for complete compositionality.

The dataset developed by Reddy et al. (2011) contained 90 English noun compounds and used an average of 30 judgments to give each phrase compositionality scores. This work provided compositionality assessments for both the phrase and its constituents enabling the use of various operations with corresponding embeddings of a compound and its distinctive parts in the context of linking human validations with measurements of semantic distance.

Ramisch et al. (2016) extended this dataset to 180 phrases presenting two parallel sets for French and Portuguese languages. English Noun-Noun compounds were mapped with Noun-Prep-Noun and Noun-Adj constructions according to the grammar equivalents. Farahmand et al. (2015) presented considerably larger dataset, which has 1,042 Noun-Noun compounds annotated with the help of 4 experts.

We also should note some works on compositionality detection datasets for non-English languages. Gurrutxaga and Alegria (2013) studied 1,200 Basque Noun-Verb collocations and resolve classification task into three classes: idiom, collocation, and free combination. Roller et al. (2013) provides 244 German compounds with compositionality scores assigned from 1 to 7 as an average from 30 validations. PARSEME project (Savary et al., 2015) is devoted to the multilingual annotation of multiword expressions (MWE) of arbitrary length and syntactical structure. By design, PARSEME is more suited for MWE extraction tasks rather than compositionality evaluation. This dataet includes annotated verbal MWEs for several Slavic languages Jana et al. (2019) explored the use of hyperbolic embeddings for noun compositionality detection comparing it to the Euclidian embeddings.

Most of the experiments on noun compositionality were conducted for the English language and to the best of our knowledge, to date, there are no datasets for compositionality detection task for any Slavic language structurally similar to (Reddy et al., 2011) and (Farahmand et al., 2015).

Agreement Metric	Value
Pearson's correlation	0.541
Cronbach's alpha	0.700

Table 1: Annotation agreement metrics for our dataset.

3 Noun Compound Dataset

3.1 Data Collection

The compound phrases are collected from the Russian Universal Dependency (UD) treebanks (Nivre et al., 2016) according to part of speech patterns, such as adjectives (ADJ) + noun (NOUN) or noun + noun, based on gold-standard UD annotations, which guarantees that not only no preprocessing but also no POS tagging and no disambiguation is required. We use all Russian treebanks, available in the UD project. They consist of texts from the following genre: news, nonfiction, fiction. To extract nominal compounds, we loop over all nouns and select only those, which has noun or adjective dependant (i.e., are "head" of another noun or adjective). We filter out non-frequent compounds, and from the list of frequent compounds, we randomly select 1,000 compounds to be annotated. Note, that this procedure is coarse and does not rely on more precise compound definition such as the exact type of the dependency between the head and dependant tokens.

Each compound is lowercased and lemmatized. Stress characters are omitted. The head noun is provided in the nominal case and in singular number (if it exists), and the dependant adjectives are put in grammatical agreement with the head noun in case and gender, while dependant nouns remain unchanged.

Type of Compound	Compound Samples
Compositional (1)	*aviatsiannaya bomba* [aircraft bomb], *gimn strany* [national anthem], *gornolyzhnyi kurort* [ski resort], *dno okeana* [ocean bed], *federalnyi zakon* [federal law]
Non-compositional (0)	*goryachaya tochka* [trouble spot], *zheleznyi zanaves* [iron curtain], *kamennyi vek* [the Stone Age], *tsar gory* [king of the hill], *novaya volna* [new wave]
Ambiguous (2)	*novyi god* [New year celebration or new year], *krupnaya set'* [big net or big network], *ogromnaya massa* [big mass of or big amount of], *pozitsiya kompanii* [company place or company position], *drevnyaya professiya* [ancient profession or prostitution]

Table 2: Examples of non-compositional (0), compositional (1) and ambiguous (2) compounds.

3.2 Annotation Setup and Agreement

Each compound phrase in the selected list is annotated by two experts according to the following schema: (0) the phrase is non-compositional; (1) the phrase is compositional; (2) the phrase is ambiguous, which means that exact compositionality of the phrase is dependant on the corresponding context. After that, annotators' answers are reviewed by a moderator. Out of 1,000 randomly selected compounds, moderator samples 220 and resolves the ambiguity left from the first two annotators. We calculate the agreement metrics of the first two annotators on the dataset of 1,000 compounds. Annotators achieved a substantial agreement. We note that the typical problematic cases that are hard to annotate are compounds, which meaning tends to be compositional in a metaphorical way, e.g., *"otkrytoe more"* [open sea] and compounds, that contain polysemic words: *"hod dela"* [justicement or the course of business].

3.3 Dataset Description

The resulting dataset consists of 220 compound phrases with several full sentence contexts, collected from source texts. The number of contexts is not fixed. So far the contexts are not annotated. A few examples are provided in Table 2. Table 3 presents the cross-tabulation of compound pattern and compound compositionality. Each compound is provided with a sentence context. The number of contexts is not fixed as we extract all contexts that contain the compound from the UD treebanks. The contexts so far are not used in the experiments. However, one of the possible directions for the future work would be compound disambiguation, based on the contexts. Examples of the compound contexts are presented in Figure 1.

4 Experiments

We evaluate various methods for detection of compositionality presented in the previous work. For experiments, we train a distributional semantic model (DSM) that includes embeddings not just for single words but also for compounds. We achieve this by replacing in the training corpora all occurrences of compounds from the proposed resource with single tokens composed of their parts. We use two experimental setups in our work.

First, the **unsupervised setup** follows the method and evaluation pipeline presented in (Cordeiro et al., 2016). In this setting, we rely solely on a similarity between a compound embedding and an embedding composed from its parts using an additive function. The value of the similarity should correlate with annotators' judgments in the proposed resource.

Second, the **supervised setup** considers compositionality detection as a binary classification task. We train various supervised machine learning methods on vector representations of a compound and its parts to predict compositionality class. In this setup, we train an additional DSM that does not have any modifications (it does not contain embeddings for compounds). In this setup, embeddings of compound parts are obtained from this unmodified supplementary model.

We train DSMs using fastText (Bojanowski et al., 2016) and word2vec (Mikolov et al., 2013) models with CBOW architecture implemented in *gensim* package (Rehurek and Sojka, 2011). Russian Wikipedia dump is used as a training corpus (as of 02.05.2019, it consists of 1,542,621 articles), with Universal Dependencies raw texts as an enrichment, which helps to deal with cases of missing compounds. Both Wikipedia articles and compounds are lemmatized using MyStem

Под воздействием этого поля ядра	**атомов водорода**	в теле исследуемого , каждый со своим слабым магнитным полем , ориентируются определенным образом относительно сильного поля магнита .
Прозрачная жидкость , в которой на два	**атома водорода**	приходится один атом кислорода , может быть водой , а может быть и смесью жидких водорода и кислорода
Нам удалось сложить кучку из восьми атомов - двух атомов углерода и шести	**атомов водорода**	, изображенную на рисунке .
С чего начинать : сдвинуть два атома углерода или приставить	**атом водорода**	к атому углерода ?
Китайский	**Новый год**	и другие праздники , отмечаемые тайскими китайцами , отличаются в обоих случаях , так как они рассчитываются по китайскому календарю .
Перед самым	**Новым годом**	отключили поселок Никольское .
Речь , конечно же , идет об очередной заморозке до	**нового года**	цен на бензин .
В нашем рейтинге лучших подарков мужчине под	**Новый год**	пневматическая винтовка с ночным прицелом твердо заняла первое место .
Нынешнее заседание Госсовета - первое в	**новом году**	и последнее , на котором Владимир Путин выступит как президент страны .
Но это же был единственный русский фильм на	**Новый год**	, у него были все шансы на успех " .
А у нас политик	**второго эшелона**	ниже этого эшелона не опустится " , - говорит эксперт .
Несмотря на озабоченность Минобрнауки бесконтрольным размножением экономистов и недоверие солидных работодателей к дипломам вузов	**второго эшелона**	, молодой экономист сегодня вряд ли останется на обочине жизни .
Пока потребители	**второго эшелонов**	дожидаются сезона распродаж или приобретают подержанные вещи , лидеры консюмеризма переходят к следующей фазе потребления .
Опускаясь по стратификационной лестнице , они опережают по статусу тех , кто находится во	**втором эшелоне**	, то есть в предшествующей фазе потребительской гонки .

Figure 1: Compound contexts in KWIC format. The compounds and their compositionality classes are: *atom vodoroda* [hydrogen atom] (1), *novyi god* [New year celebration or new year] (2), *vtoroyi eshelon* [second tier] (0).

	Adjective-Noun	**Noun-Noun**	**Total**
Non-compositional (0)	23	10	33
Compositional (1)	71	96	167
Ambiguous (2)	9	11	20
Total	103	117	220

Table 3: The number of compositional and non-compositional compounds in our dataset.

(Segalovich, 2003). Minimal frequency count of 2 is used. We performed experiments on several sets of hyperparameters (dimensionality and amount of training epochs). We found that dimensionality of 300 and five epochs give good or the best results across all considered settings, therefore, we report results only for this set of hyperparameters.

To simplify the task, in experimental evaluation, we do not consider contextual information of compounds. It means that no ambiguity is under consideration and only phrases with compositionality classes of 1 and 0 are qualified for evaluation, which leaves 200 compounds. For three of them, models lack an embedding, which leaves 197 phrases for experiments: 164 are compositional, and 33 are non-compositional according to annotators (approximately 0.83 to 0.17 ratio).

4.1 Unsupervised Setup

For unsupervised setup, we calculate a metric from (Cordeiro et al., 2016) that measures similarity of an embedding of a compound as a whole and an additive embedding composed of its parts. Consider w_1, w_2 are words of a given compound and a function $v(\cdot)$ yielding vector representation of a word/compound. Then the similarity metric is equals to: $cos(v(w_1w_2), v(w_1 + w_2))$, where $v(w_1 + w_2)$ is the normalized sum:

$$v(w_1 + w_2) = \frac{v(w_1)}{\|v(w_1)\|} + \frac{v(w_2)}{\|v(w_2)\|}.$$

In addition to cosine, we use similarity measures based on distance metrics between embeddings: Chebyshev distance (L_∞-norm), Manhattan distance (L_1-norm), and Euclidean distance (L_2-norm). When using these distances, instead

Supervised Model	Spearman's ρ	Precision	Recall	F-measure
Linear Support Vector classifier (LSVC)	**0.47**	0.37	**0.78**	0.48
Multi-layer Perceptron (MLP)	**0.46**	0.32	**0.82**	0.44
Desicion Tree (DT)	0.18	0.31	0.36	0.31
Naïve Bayes (NB)	**0.43**	**0.55**	0.52	**0.52**

Table 4: Performance of the classifiers in the supervised setup (classifier metrics presented for class 0).

Metric / Model	fastText	word2vec
cos (norm.)	**0.42**	0.37
L_∞ (avg.)	0.33	0.09
L_1 (avg.)	0.33	0.14
L_2 (avg.)	0.33	0.14

Table 5: Spearman correlation (ρ) of the metric with annotator judgments in the unsupervised setup.

of normalized sum, we use a simple averaging:

$$v(w_1 + w_2) = \frac{1}{2}(w_1 + w_2).$$

We evaluate the performance of these metrics to predict compositionality based on Spearman rank correlation (Spearman's ρ) between them and the compositionality class in the annotated dataset as considered in (Cordeiro et al., 2016).

4.2 Supervised Setup

In supervised setup, we access average performance on 25 stratified randomized splits of the selected dataset into 75% for training and 25% for testing with the following machine learning algorithms: linear support vector machine (LSVC) with $C = 1$ (Platt, 1999); three-layer perceptron (MLP) with α=1, solver='lbfgs', sizes of layers=200/20/20 (Hinton, 1989); decision tree (DT) with maximum depth=10, max features=20 (Breiman, 2017); Naïve Bayes (NB) (Zhang, 2004). For feature representation, we use a concatenation of compound embedding with embeddings of compound parts. We evaluate the Spearman correlation with the annotation class, as well as precision, recall, and F_1-score.

4.3 Results and Discussion

The results of the experimental evaluation for unsupervised setup are presented in Table 5, for supervised setup – in Table 4. Of presented metrics, L_1, L_2, and L_∞ present substantial negative correlation. That can be explained by the nature of embedding vectors. The bigger the distance value, the further compound is from its components in a semantic sense. If the sense of the compound widely differs from corresponding senses of its components, it is deemed as non-compositional. To be comparable with previous papers, we present a positive correlation bringing minus of a distance instead. Taking this into consideration, all metrics perform comparably on the dataset. We can see a not strong, yet stable and substantial correlation between similarity and compositionality class.

Considering the supervised classification task, precision, recall, and F_1 metrics are presented alongside Spearman rank correlation. As non-compositional compounds are in the minority in this dataset, and detecting idiomatic phrases provides more interest practice-wise, we report on zero-class quality metrics to access algorithm performance. LSVC, MLP, and NB present higher ρ than the unsupervised counterpart. LSVC and MLP also give relatively high recall on non-compositional examples. Overall, linear SVC and multi-layer perceptron perform better than the other models across all metrics.

5 Conclusion

We presented the first Russian-language dataset of noun compounds annotated, where each compound follows one of the noun compound patterns (noun+noun or adjective+noun) and is annotated with as non-compositional, compositional or ambiguous compounds. The latter can be either compositional or not, depending on the context. Each compound is provided along with the sentence contexts. The inter-annotator agreement metrics show that annotator judgments on the scores agree well. We investigated the performance of various algorithms from previous work and showed that the achieved evaluation metrics correspond with other state-of-the-art results for English. We hope that our resource will foster the research in the area of compositionality detection for Russian and other Slavic languages.

Acknowledgements

Dmitry Puzyrev and Ekaterina Artemova were supported by the framework of the HSE University Basic Research Program and Russian Academic Excellence Project "5-100".

References

Luis Espinosa Anke and Steven Schockaert. 2018. Seven: Augmenting word embeddings with unsupervised relation vectors. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2653–2665.

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*, pages 89–96.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of CoNLL*, pages 1–7.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Leo Breiman. 2017. *Classification and regression trees*. Routledge.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1986–1997.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.

Antton Gurrutxaga and Iñaki Alegria. 2013. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 116–125.

Geoffrey E. Hinton. 1989. Connectionist learning procedures. *Artif. Intell.*, 40(1-3):185–234.

Abhik Jana, Dmitry Puzyrev, Alexander Panchenko, Pawan Goyal, Chris Biemann, and Animesh Mukherjee. 2019. On the compositionality prediction of noun phrases using poincaré embeddings. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 73–80.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 32–41.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language*

Technologies as a Challenge for Computer Science and Linguistics (LTC 2015).

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (v-n) collocations by integrating features. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 899–906.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1371.

Harry Zhang. 2004. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, volume 2.