Developing a Russian Database of Regular Semantic Relations Based on Word Embeddings

Ekaterina Enikeeva, Andrey Popov
Saint Petersburg State University
E-mail: protoev@yandex.ru, hedgeonline@gmail.com

Abstract

Recent computational semantic models yield high-quality results with regard to semantic relations extraction tasks, and thus may be applied as a baseline for semantic lexicon construction. Moreover, the stochastic information about lexical compatibility is useful for reducing ambiguity and detecting anomalies during syntactic parsing. We prove that this approach is reasonable and describe a Russian semantic lexical database, acquired in an unsupervised manner and employed as a semantic component of a syntactic parser and a fact extraction system.

Keywords: distributional semantics, word vector representations, semantic lexicon, Meaning ↔ Text model

1 Introduction

Semantic components have usually been seen as crucial in natural language processing and understanding systems. A semantic lexicon may be constructed as a lexical database, such as WordNet[1], which maps concepts into lexemes, groups lexemes into synonymic sets, describes a number of relations between them and provides short definitions and usage examples. A more complicated variation, thus applicable in a wider range of NLP tasks, is an extensive dictionary of a language recording all possible information about lexical units. Such descriptions are rather laborious, however, since they require a comprehensive analysis of each lexeme. Therefore, there are only a few known completed examples of this approach. As regards Russian language resources, we should mention Explanatory Combinatorial Dictionary (Mel'čuk, Zholkovsky 1984) based on the Meaning ↔ Text linguistic model (Mel'čuk 1974/1999). For each described lexeme, an entry comprises its subcategorization frame, lexical co-occurrence data and a set of examples supplied with morphological description.

As noted above, this kind of lexical resource requires lots of manual work, including corpus studies and a thorough lexicographic description. However, recent studies have introduced a variety of computational semantic models that may transform this process into at least a semi-supervised one. These models are based on a distributional hypothesis (cf. the review in Sahlgren (2008)), which claims that a meaning of a word may be derived from its context; machine learning approaches are then applied to produce a computationally effective representation of words (word embeddings) that incorporates contextual information (Mikolov et al. 2013a). A "semantic space" built in such a way appears to reflect regular relations between lexical units; one can perform simple vector operations to infer examples of paradigmatic relations; a widely-cited example being the one of *king – man + woman = queen*.

Semantic word embeddings have numerous advantages: they are learnt from raw corpora and require virtually no external linguistic resources. Recently developed models are easily interpretable and may be applied to a number of tasks, including paradigmatic relations extraction, predicting semantic

[1] https://wordnet.princeton.edu/

analogy and describing selectional restrictions (cf. SemEval task on semantic comparison for words and texts, RUSSE contest, etc.).

We propose an integrated system for building a semantic lexicon for Russian from raw text corpora using few linguistic resources. Our lexical database is designed to reinforce the syntactic parser and fact extraction system working with the Russian language. We aim at overcoming the syntactic ambiguity issue and reducing the number of possible parse trees for a sentence (Popov & Enikeeva 2017), so in this paper we focus mainly on the description of syntagmatic relations. Instead of just listing selectional restrictions for each lexeme, the lexicon includes collocational probability even for those combinations that do not occur in training corpus.

The paper is structured as follows: firstly, we describe the related work, including lexicographical applications of word embeddings. Then we briefly outline our computational approach to describing syntagmatic relations within the lexicon, and provide an overview of corpora and other data we use. In the Evaluation section we report the system quality in terms of widely-used simple metrics (precision of collocates ranking), and discuss the necessity of more elaborated annotation. Finally, an automated lexical description for several lexemes is presented.

2 Related Work

Compositional distributional semantics is successfully applied to a number of semantics-related tasks in NLP. As far as evidence for Russian is concerned, a number of semantic vector models were evaluated during the RUSSE workshop (Panchenko et al. 2015). However, they focused on paradigmatic relations between lexical units: the annotated data includes human judgements on semantic relatedness (synonyms, hypernyms and hyponyms) and free association (collected during a large-scale psychological associative experiment). This remains a gold standard annotation and is used by more recently developed tools for Russian distributional modeling: RusVectōrēs (Kutuzov & Kuzmenko 2017), AdaGram (Bartunov et al. 2015) and RDT (Panchenko et al. 2016).

Selectional preference extraction is not so obviously captured by distributional models, and the research in this direction is quite sparse. Jauhar and Hovy (2017) develop a minimally supervised frame lexicon induction method based on a predictive embedding model and an Indian Buffet Process posterior regularizer. Interestingly, they show that the model yields some regularities in frame realizations in addition to hand-crafted data. In Rodríguez-Fernández et al. (2016) a distributional baseline metric is introduced: collocates are evaluated against the difference between an example headword and collocate added to the test headword. The main method proposed in the same paper is based on a linear transformation between a headword and collocate space. The approach is tested on manually classified samples drawn from Macmillan Collocations Dictionary (Rundell 2010). A neural network architecture for selectional preference modeling (also based on distributional hypothesis) is described in Van der Cruys (2014).

Experiments described in Bukia et al. (2016) and Kutuzov et al. (2017) lay the foundations of selectional preferences extraction from Russian corpora, and propose alternative solutions to the problem, but there is much to be done in this field. Bukia et al. (2016) compares two distributional approaches to selectional preference modeling. The first implies semantic similarity calculation based on cosine distance, while the second relies on Mikolov's (2013b) assumption about linguistic regularities captured by distributed word vector models. The clustering of attributive collocations in Kutuzov et al. (2017) is also worth mentioning: in this case the authors employ two-step clustering to group collocations with body parts names into semantic classes. Our study follows the line of previous research in the field generalizing the results obtained on specific test sets.

3 Computational Model

3.1 Semantic Relation as a Linear Transformation

As mentioned above, Mikolov and colleagues (Mikolov et al. 2013b) show that regular linguistic relations between two word spaces may be described as a linear transformation on them. The syntagmatic relations may be classified and described in numerous ways (for example, as a lexical function or just semantic collocation class), but in this section we will refer to the relation to be modeled in general.

Our task is to predict possible values of a particular relation for a given target word (base) using training exemplars of this relation. Following Rodríguez-Fernández et al. (2016), we define a base space B and a collocate space C produced by a word embedding model. Let T be a set of collocations t_i comprising base – collocate pairs (b_{t_i}, c_{t_i}) that represent a given relation L.Argument matrix B_T=[b_{t_1},...,b_{t_n}] and collocate matrix C_T=[c_{t_1},...,c_{t_n}] are made up of corresponding word vectors. Then, given the examples of a particular relation (e.g., a lexical function MAGN: *тяжёлая болезнь* ‘serious illness’, *сильный акцент* ‘heavy accent’, etc.), we should find a transformation which converts a base vector to a collocate vector, for instance, predicts a collocate *бурный* ‘wild’ (MAGN value) for a base *аплодисменты* ‘applause’.

A linear transformation matrix $\Psi \in R^{BxC}$ learnt from training set T satisfies the following:

$$A_T\Psi_T = C_T.$$

Therefore, Ψ can be approximated using a singular value decomposition to minimize the sum:

$$\sum_{i=1}^{|T|} ||\Psi_T a_{t_i} - c_{t_i}||^2. \text{ F}$$

Thus, we obtain a transformation matrix for a given relation. Applying it (multiplying it by the base embedding) we obtain a ranked list of potential collocates for given target word and relation.

Rodríguez-Fernández et al. (2016) prove their assumption that base and collocate embeddings should be trained on different corpora. In their work base vectors are obtained from a small corpus containing primarily literal usage (Wikipedia), while collocate vectors are trained on a large corpus full of various figurative meanings. The performance of this model was evaluated on Russian lexical functions data in Enikeeva & Mitrofanova (2017). The authors conduct experiments on the 10 most frequent lexical functions from the Russian National Corpus, fitting linear transformation for each LF and applying it to rank potential collocates. The final scores after heuristic filtering are quite promising (reaching 0.9 in precision).

3.2 Collocation Clustering

We have already noticed that the amount of properly classified syntagmatic relations examples is usually low, hence we need a technique to induce the semantic classes automatically. Following Kutuzov et al. (2017), we can apply clustering algorithms to collocations from corpus represented by the corresponding embeddings. The K-means algorithm is a simple clustering technique producing a known number of classes, and has been successfully applied to various NLP tasks (cf. Berry, Kogan (2010)). This approach may be helpful to fit an existing classification, but in natural language this is not usually the case. Consider, for example, attributive adjective-noun combinations that reflect various relations between an object (noun) and its description (for a detailed study of adjectives categorization see Heyvaert (2010)): the attributive relation in ‘*a sad book*’ is not the same as in ‘*a sad girl*’. Thus, an Affinity Propagation algorithm is used here to infer unknown number of groups. We perform clustering on stacked base and collocate embeddings and the results of such a simple approach will be discussed in the Evaluation section.

3.3 Lexicon Structure

Each lexical entry includes the following information about the semantic compatibility of a headword:

- its regular syntagmatic relations;
- paradigmatic relations: as mentioned above, we provide lists of collocates corresponding to a particular relation instead of strict selectional restrictions;
- idioms;
- peculiarities of word meanings.

Semantic relations are represented as a web of lexical units connected by marked links of several types (synonymy, hyponymy, etc.) and probability counts are assigned to each link. This structure may also be useful to define inherited semantic relations. Consider several nouns belonging to a particular semantic class. “A lion”, “a cat”, “a squirrel” are instances of “an animal”, and some properties of the hypernym may be inherited by a specimen of the class: e.g., the probability of combining “an animal” with a particular set of motion verbs (“to run”, “to jump”) is rather high, and this information may be transferred to the next level (“a lion” also tends to appear with the verbs mentioned above). As opposed to the paradigmatic relations weighting described above, syntagmatic relations extraction is a straightforward application of word embeddings. Our model produces synonyms, hypernyms and hyponym lists by means of cosine similarity applied to word vectors; the results are filtered by heuristics. The semantic network may be enhanced by hierarchical clustering of word representations, taking into account syntactic annotation and representing word context as its syntactic neighbors.

4 Data Sources

To the best of our knowledge, the only source of selectional restrictions information for Russian of considerable size is the Framebank project.[2] The main focus of this project is verbal subcategorization frames, so the annotation is verb-oriented. Moreover, the examples of the frame realization may include syntactic constructions and even clauses, which is not of primary interest in our case.

Another option is to use lexical functions (LF) formalism developed within the Meaning↔Text theory (Mel’čuk 1998). At present the inventory of LFs comprises 116 varieties of standard and nonstandard LFs (Apresjan et al. 2007). Russian collocations revealing LF relations are thoroughly described in the *Explanatory Combinatorial Dictionary of Modern Russian* (Zholkovsky & Melchuk 1984). The machine-readable resources containing LF markup for Russian language are quite limited. In our experiments we use SynTagRus Treebank[3] and a verbal combinatory dictionary of Russian abstract nouns.[4]

SynTagRus (Boguslavsky 2014) is a treebank subset of the Russian National Corpus. It consists of more than 28,000 sentences annotated with a dependency parse tree as well as some semantic information including a list of LFs in Meaning↔Text notation and word sense disambiguation. The verbal combinatory dictionary uses its own markup scheme based on LF inventory. About 10,000 collocations are classified in terms of ‘regular abstract meanings’, such as necessity, existence, and action, with additional labels such as phase (start, finish) or semantic class (cognition, perception, etc.).

[2] https://github.com/olesar/framebank

[3] http://ruscorpora.ru/en/corpora-structure.html

[4] http://dict.ruslang.ru/abstr_noun.php

Collocation description within the framework of Meaning↔Text theory rests on the idea that collocations are expected to reveal both the syntagmatic unity and lexical correlation of its parts. Consequently, the majority of LF examples are not free word combinations and reflect a bound usage. Our task of syntactic parser refinement implies that free word combinations should be captured on equal terms with idioms; and we even notice that real texts (especially colloquial speech) show much less restricted usage: word combinations that seem to be abnormal without their real context are in fact acceptable in texts. Finally, some LF attested in SynTagRus annotation are quite rare, so we select only the 20 most frequent types (ignoring special tags marking the action phase: Incep, Fin etc.)

Table 1: Training data sources and size.

Source	Type	Number of classes	Size
SynTagRus treebank	LF examples	20	4958
Verbal combinatory dictionary	LF examples	5 (+ 6 syntactic types)	9729
SynTagRus treebank	syntactically linked collocations	20	75142

In order to train the model on more diverse examples, we extracted word pairs connected by a syntactic link from the SynTagRus treebank. These collocations are classified generally by corresponding syntactic relations, so they were further clustered into more specific semantic classes. The resulting data sources and their sizes are presented in Table 1. Several sources of distributional word representations are freely available for Russian. We have chosen RusVectōrēs[5] (Kutuzov & Kuzmenko 2017) as a primary source: this provide 300-dimensional vectors pre-trained by continuous skip-gram architecture using the word2vec toolkit (Mikolov et al. 2013a) on corpora tagged with Universal Dependencies[6] morphological tags. Word embeddings learnt from Russian Wikipedia were used for modeling headword sense, while the vectors learnt from the Russian National Corpus were applied to possible collocates in order to capture figurative and metaphorical usage. The vocabulary size is 19,5071 and 384,746 lexemes, respectively. Each vector in this model corresponds to a combination of lemma and its part-of-speech tag. More precise results are expected with the AdaGram model (Bartunov et al. 2015), which provides multiple vectors for a word to reflect polysemy; though by now 'one vector per word' models seem to capture semantic ambiguity. The linear transformation model and clustering were trained by means of scikit-learn toolkit[7] (v0.19.1).

5 Evaluation

5.1 Collocate Weighting

Firstly, we describe the collocate ranking process and the evaluation of the results. The training data is heterogenous, and therefore we do not merge it and train the whole model to predict an uninterpretable 'generalized' probability, but process and evaluate different datasets separately. Given a list of

[5] http://rusvectores.org/en/

[6] http://universaldependencies.org

[7] http://scikit-learn.org/stable/

examples of a specific relation, a corresponding linear transformation is learnt from it and the result is applied to a set of collocation hypotheses to be weighted. We define several morphosyntactic types of collocations, such as '*attributive adjective + noun*' (1) or '*verb + noun as a direct object*' (2) and the training sets are divided in the same manner (for instance, the examples of lexical functions Magn, AntiMagn, Bon, and AntiBon are seen as corresponding to collocation type (1)).

The lexical function example lists are used as is with the morphosyntactic annotation from SynTagRus. The collocates from corpus are annotated by the pymorphy2[8] morphological analyzer and clustered by means of the K-means algorithm as a simple baseline. This appears to perform quite well on the task of collocation clustering: for example, the attributive noun+adjective collocations are clustered into 20 groups, which can be easily recognized as containing collocations of a particular semantic class: quantitative description (*длинная дорога* 'long road'), relation to a particular object (*химический опыт* 'chemical experiment'), characteristic feature description (*верный друг* 'true friend'), etc. Some groups belong to the same semantic class, and we merged these manually: for example, the '*relation to a particular object*' collocations were found in three automatically clustered groups.

The performance of ranking collocates for a particular headword is then evaluated using precision and mean reciprocal rank (MRR) on this list of top N collocates, as annotated by experts:

$$precision = \frac{tp}{tp + fp}$$

where tp is the number of correct collocates on the retrieved list, fp is the number of false collocates on the list;

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where Q is the top-N list and $rank_i$ is a rank of the first correct collocate. Gold standard annotation is performed by two experts, and the final test set includes only the cases where the annotators agree (about 85%).

Table 2 shows the top-20 evaluation results on several collocation types for 10 headwords using five-fold cross-validation.

A brief comment on LF notation should be made. We used the following lexical functions corresponding to attributive construction:

- *Magn* means 'very', 'to a (very) high degree', 'intense(ly)': Magn(*проблема* 'problem') = *серьезная* 'serious';
- *AntiMagn* – vice versa.

The following lexical functions are usually represented by verb + direct object in Russian:

- *Oper1* introduces a support verb meaning 'do', 'perform' something, expressed by noun: Oper1(*поддержка* 'support') = *оказывать* '(to) lend';
- Func1 means that its argument belongs to the subject of corresponding verb: Func1(*власть* 'authority') = *принадлежать* 'belong';
- In the verbal combinatory dictionary the collocations are annotated with base classes (action, state, etc.) and relation between headword and its argument (subject, direct object, etc.). Here we use only actions with a direct object tag.

[8] https://github.com/kmike/pymorphy2

Table 2: Top-20 evaluation results on several collocation types for 10 headwords using five-fold cross-validation.

Collocation type	Training data	Precision	MRR
attributive adjective + noun	LF examples (Magn, AntiMagn)	0.84	0.9
attributive adjective + noun	corpus collocation (clustered as 'characteristic feature')	0.89	0.92
attributive adjective + noun	corpus collocations (clustered as 'relation to object')	0.9	0.94
verb + direct object (noun)	LF examples (Oper1, Func1)	0.64	0.73
verb + direct object (noun)	LF examples from verbal combinatory dictionary	0.6	0.66
verb + subject (noun)	LF examples (Func0)	0.42	0.89

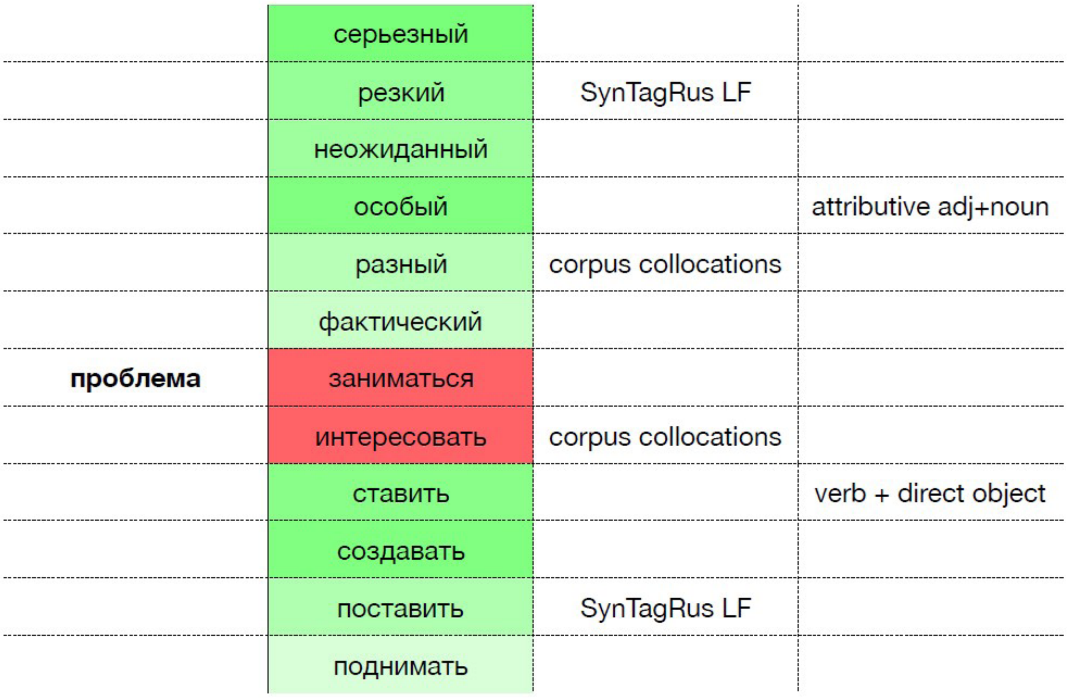
A specific relation between verb and subject is represented by the Func0 lexical function: thus means that an event described by a headword takes place: Func0(*снег* 'snow') = *идёт* 'falls'.

5.2 Lexical Description

To the best of our knowledge, there are no semantic lexicon examples for Russian to compare with the results. Instead, we compile several entries automatically and assess them manually. For each syntactic type we learn from multiple data sources and then assign each collocate hypothesis the probability from the model in which it is scored the highest. The examples for frequent Russian lexemes are presented in Figures 1 – 3.

The rows in figures represent collocates grouped by the source of training data and possible syntactic construction. The erroneous predictions of the distributional model are marked in red color, green shades correspond to collocational probability: the more intense the color, the higher the score.

We include examples of errors in entries to discuss the matter in detail. Figure 1 represents collocational examples from the 'syntagmatic' part of the entry for a lexeme *проблема* 'problem/issue'. It shows various degrees of collocational strength for attributives with the words *серьезный*, *резкий*, *неожиданный* 'serious, sharp, unexpected' being more probable collocates and the words *разный*, *фактический* 'different, actual' being less probable ones. As for verb + direct object relation, we list acceptable verb examples: *ставить*, *создавать*, *поставить*, *поднимать* 'raise (an issue), cause, raise, raise' along with the erroneously predicted with a high probability collocates *заниматься*, *интересовать* 'deal with, be interested in'. In fact, these verbs are usually



	серьезный		
	резкий	SynTagRus LF	
	неожиданный		
	особый		attributive adj+noun
	разный	corpus collocations	
	фактический		
проблема	заниматься		
	интересовать	corpus collocations	
	ставить		verb + direct object
	создавать		
	поставить	SynTagRus LF	
	поднимать		

Figure 1: Part of lexicon entry for a noun проблема ‘problem’.

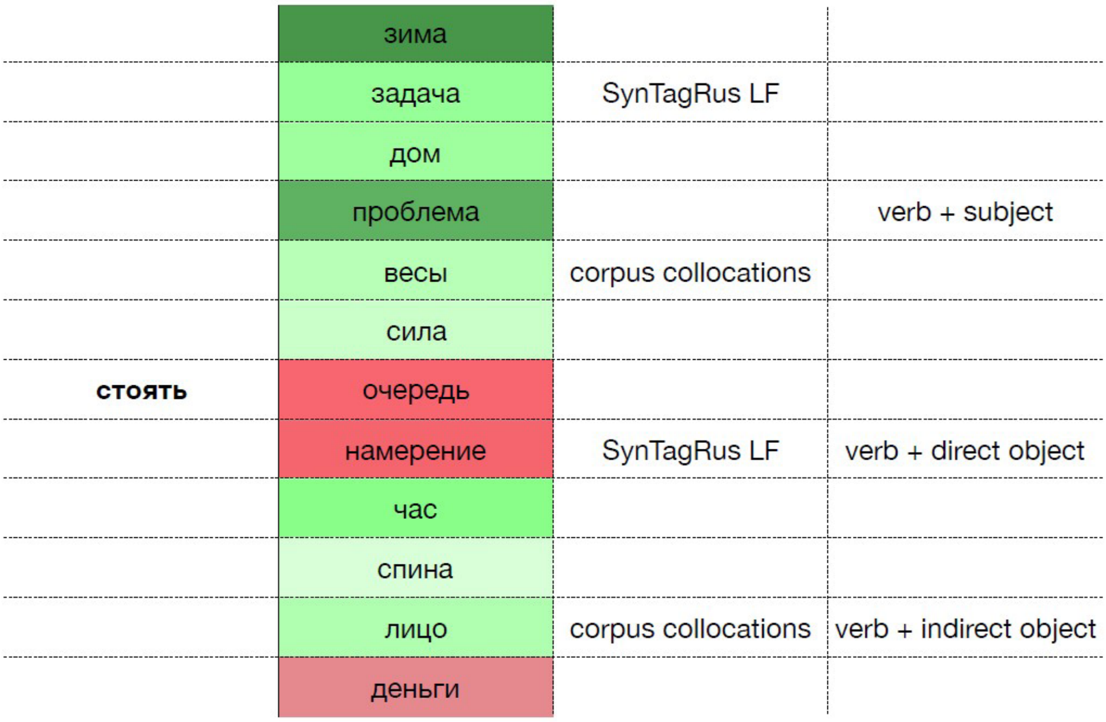
collocated with the word *проблема* in Russian, but also appear in other types of construction, as in the following examples[9] (1-2):

> Мировой опыт показывает, что строители вообще не должны *заниматься проблемами* (Verb + Noun instrumental) подключения к ресурсам. ‘The global experience shows, that the builders should not deal with resource connection problems.’ (1)
> Но его *интересовали проблемы* (Verb + Noun nominative), не имевшие отношения к науке вообще. ‘He was interested in issues unrelated to science.’ (2)

Figure 2 briefly shows several fillers of an argument structure of a verb *стоять* ‘stand’. We would like to emphasize the ability of the model to describe ambiguous examples: consider the upper part of the figure representing collocates with a subject role. In Russian, one of the frequent figurative usages of the verb *стоять* ‘stand’ is related to time periods, especially seasons: *Стояла зима* ‘It was winter’; a similar one is attested in collocation with abstract nouns such as *проблема* ‘issue’: *стоящая перед нами проблема* ‘the issue we are facing’. On the other hand, the literal collocations with nouns denoting physical objects are also scored quite high: *дом* ‘house’ is colored with a bright shade in Figure 2.

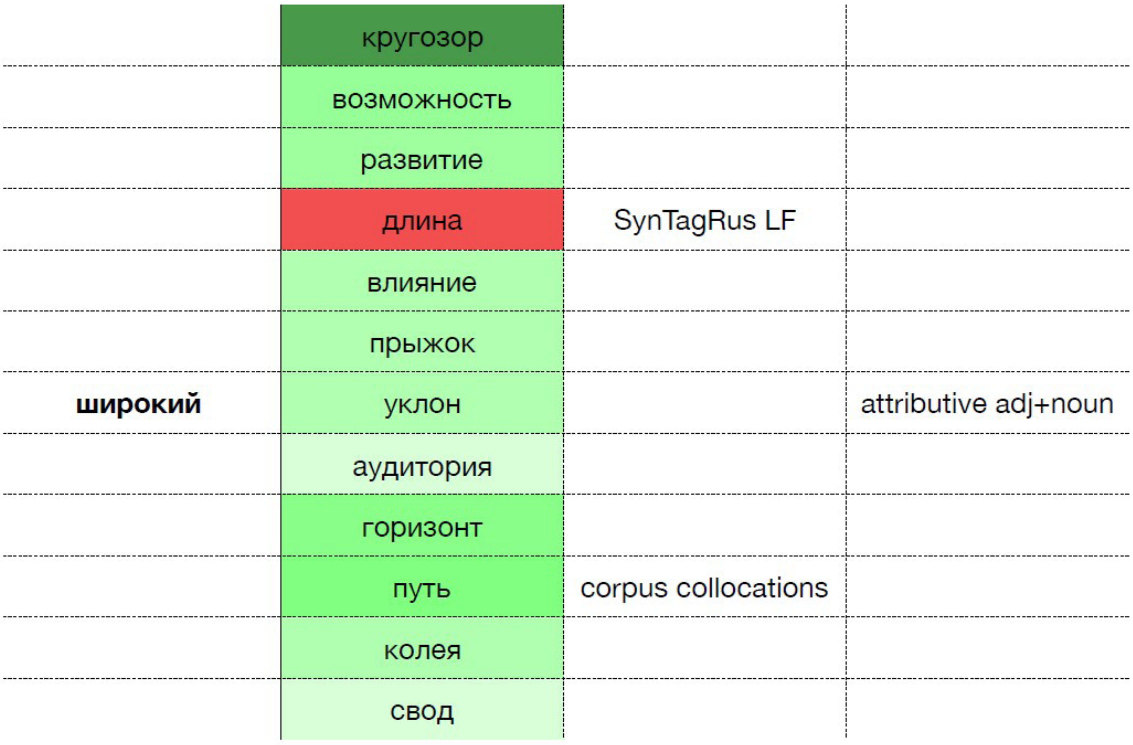
Figure 3 illustrates an adjective широкий ‘wide’ as it reveals the peculiarities of different training sources within one syntactic type. Possible collocates obtained during learning from SynTagRus lexical functions кругозор, возможность, развитие ‘horizon/outlook, possibility, development’ are abstract nouns and typical illustrations of Magn LF. The distributional model also assigns high scores to other nouns with abstract meanings such as влияние ‘influence’ and even to прыжок ‘jump’, which has a figurative abstract meaning. However, the level of generalization is too high, because some unacceptable collocates (bearing abstract meaning) are ranked in top, as it is the case with длина

[9] The examples are taken from Russian National Corpus: http://ruscorpora.ru/en/search-main.html



	зима		
	задача	SynTagRus LF	
	дом		
	проблема		verb + subject
	весы	corpus collocations	
	сила		
стоять	очередь		
	намерение	SynTagRus LF	verb + direct object
	час		
	спина		
	лицо	corpus collocations	verb + indirect object
	деньги		

Figure 2: Part of lexicon entry for a verb *стоять* ‘stand’.



	кругозор		
	возможность		
	развитие		
	длина	SynTagRus LF	
	влияние		
	прыжок		
широкий	уклон		attributive adj+noun
	аудитория		
	горизонт		
	путь	corpus collocations	
	колея		
	свод		

Figure 3: Part of lexicon entry for an adjective *широкий* ‘wide’.

‘length’. On the other hand, learning from corpus collocations yields collocates with more literal meaning: аудитория, горизонт, путь, колея ‘audience, horizon, way, track’ and even less frequent as свод ‘vault’, as well as several abstract nouns such as уклон ‘tendency/direction’.

The described procedure is applied to frequent Russian words belonging to the main parts of speech – 693 nouns, 282 verbs, 221 adjectives and 109 adverbs.

6 Conclusion

In the paper we introduce an approach to automatically constructing a semantic lexicon for the Russian language based on distributional word representations. The lexicon is constructed in order to simplify syntactic disambiguation and the fact extraction process. The issues of evaluation are discussed and several examples of retrieved lexicon entries are presented. We hope that the search interface to the lexicon and some visualization features will be made available online soon.

References

Apresjan Ju.D. (1995) Selected works. Vol. 1. Lexical Semantics: Synonymic Means of Language. [Izbrannyje trudy. T.1. Leksicheskaja semantika: sinonimicheskije sredstva jazyka]. Moscow.

Bartunov S., Kondrashkin D., Osokin A., Vetrov D. (2015) Breaking Sticks and Ambiguities with Adaptive Skip-gram. https://arxiv.org/abs/1502.07257

Berry M. W., Kogan J. (2010) Text Mining: Applications and Theory. Wiley.

Boguslavsky I. (2014). SynTagRus – a Deeply Annotated Corpus of Russian. In Blumenthal, P., Novakova, I., and Siepmann, D., editors, Les émotions dans le discours-Emotions in Discourse, pages 367–380, Peter Lang, Frankfurt am Main, Germany.

Bojanowski P., Grave E., Joulin A., Mikolov T. (2017) Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics. Vol. 5. Pp. 135–146.

Bukia G. T., Protopopova E. V., Panicheva P. V., Mitrofanova O. A. (2016) Estimating Syntagmatic Association Strength Using Distributional Word Representations. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue'16". Vol. 15. pp. 112-122.

Enikeeva E., Mitrofanova O. (2017) Russian Collocation Extraction based on Word Embeddings. In: Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference «Dialogue 2017». Issue 16. Vol. 1. Moscow. Pp. 52–64.

Gak V.G. (1998) Language transformations [Jazykovyje preobrazovanija.] Moscow.

Heyvart F. An outline for a semantic categorisation of adjectives. In Anne Dykstra & Tanneke Schoonheim (eds.). 2010 Proceedings of the XIV EURALEX International Congress. 6-10 July 2010. Leeuwarden/Ljouwert: Fryske Akademy.

Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer.

Kutuzov A., Kuzmenko E., Pivovarova L. Clustering of Russian Adjective-Noun Constructions Using Word Embeddings. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing.

Jauher S. K., Hovy E. Embedded Semantic Lexicon Induction with Joint Global and Local Optimization. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017).

Mel'čuk I. (1974/1999) Experience in Theories of «Meaning ↔ Text» Linguistic Models [Opyt teoriji lingvisticheskih modelej «Smysl ↔ Tekst»]. Moscow.

Mel'čuk I., Zholkovsky A. (1984) Explanatory Combinatorial Dictionary of Modern Russian [Tolkovo-kombinatornyj slovar russkogo jazyka]. Vienna.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a) Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR.

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. (2013b) Distributed Representations of Words and Phrases and their Compositionality. In: Advances in neural information processing systems. Pp. 3111–3119.

Panchenko A., Loukachevitch N., Ustalov D., Paperno D., Meyer C., Konstantinova N. (2015) RUSSE: The first workshop on Russian semantic similarity. In: Computational Linguistics and Intellectual Technologies, Proceedings of the Annual International Conference "Dialogue". Pp. 89-105.

Panchenko A., Ustalov D., Arefyev N., Paperno D., Konstantinova N., Loukachevitch N. and Biemann C. (2016): Human and Machine Judgements about Russian Semantic Relatedness. In Proceedings of the 5th Conference on Analysis of Images, Social Networks, and Texts (AIST'2016). Communications in Computer and Information Science (CCIS). Springer-Verlag Berlin Heidelberg.

Popov A., Enikeeva E. (2017) Template Search Algorithm for Multiple Syntactic Parses. In: IMS'17, June 21–23, 2017, St. Petersburg, Russia. DOI: http://dx.doi.org/10.1145/12345.67890 (In print)

Rodríguez-Fernández S., Anke L., Carlini R., Wanner L. (2016) Semantics-driven recognition of collocations using word embeddings, Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany.

Rundell, M. (2010) Macmillan Collocations Dictionary, Macmillan.

Sahlgren M. (2008) The Distributional Hypothesis. From context to meaning. In: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), Rivista di Linguistica. Vol. 20. №1. Pp. 33–53.

Acknowledgements

The reported study is supported by the Russian Fund of Basic Research (RFBR) grants 16-06-00529 "Development of a linguistic toolkit for semantic analysis of Russian text corpora by statistical techniques" and 17-29-09159 "Quantitative grammar of Russian prepositional constructions". We are deeply grateful to SynTagRus team, especially to Leonid L. Iomdin and colleagues from IPPI RAS, for providing access to the Russian treebank and lexical functions annotation. We would also like to thank Olga Mitrofanova and anonymous reviewers for their valuable comments.