Sense-Annotated Corpus for Russian

Alexander Kirillovich
Higher School of Economics
Kazan Federal University
Moscow & Kazan, Russia
alik.kirillovich@gmail.com

Natalia Loukachevitch
Lomonosov Moscow State University
Institute for System Programming of RAS
Moscow, Russia
louk_nat@mail.ru

Maksim Kulaev
Higher School of Economics
Moscow, Russia
kulaevma@yandex.ru

Angelina Bolshina
Moscow State University
Moscow, Russia
angelina_ku@mail.ru

Dmitry Ilvovsky
Higher School of Economics
Moscow, Russia
dilv_ru@yahoo.com

Abstract

We present a sense-annotated corpus for Russian. The resource was obtained my manually annotating texts from the OpenCorpora corpus, an open corpus for the Russian language, by senses of Russian wordnet RuWordNet. The annotation was used as a test collection for comparing unsupervised (Personalized Pagerank) and pseudo-labeling methods for Russian word sense disambiguation.

Keywords: corpus linguistics, word sense disambiguation, wordnet, Russian

1 Introduction

The task of automatic word sense disambiguation is the central task of automatic semantic analysis of texts and consists in choosing the correct word sense in the context of its use. The best results in this task have been achieved through the use of machine learning methods, which are based on preliminary manual annotation of a text corpus by lexical senses.

Most existing text collections for word sense disambiguation are annotated using sense inventory of WordNet-like resources (Miller et al., 1990; Petrolito and Bond, 2014; Pasini et al., 2021). In this paper we consider a new corpus annotated word senses for Russian, which uses the word sense inventory of Russian wordnet - RuWordNet (Loukachevitch et al., 2016). We also test some baseline methods using the created corpus such as the most frequent sense (MFS), unsupervised personalized pagerank method (Agirre and Soroa, 2009; Agirre et al., 2018), and pseudolabeling based on so-called monosemous relative approach (Martinez et al., 2008; Bolshina and Loukachevitch, 2020a).

2 Related work

2.1 WSD methods

The best results for automatic methods for word sense disambiguation are achieved by supervised methods (Bevilacqua et al., 2021; Pasini et al., 2021). The training of such methods requires manual sense annotation of a large text corpus, which is a laborious work. Large semantically annotated corpora are available mostly for English (Pasini et al., 2021).

There can be two main approaches to reduce data labeling costs. The first approach is based on automatic annotation of data using some additional resources, so-called automatic pseudolabeling. Pseudo-labeling methods can be based on different techniques of annotation such as parallel text collections (Taghipour and Ng, 2015), monosemous related words (so called monosemous relatives) (Martinez et al., 2008) and others. Such automatically annotated data are then used for training supervised methods.

The second group of methods are unsupervised methods, which do not require any labelled dataset for disambiguation. Such methods usually use manual dictionaries or thesauri (such as wordnets), their inventories of senses and corresponding information (word sense definitions, relations between words and senses) to disambiguate words (Navigli and Lapata, 2009; Moro et al., 2014; Agirre and Soroa, 2009). They are the most useful ones in case of dealing with low-resource data or modelling of some link-based dependencies.

The main assumption for unsupervised WSD is that semantically-related senses are presented in similar contexts. In this case a method of disambiguation should include a semantic similarity

metric. In graph-based techniques an analogue of such metric may be a link between entities in a graph. Therefore, it is possible to calculate semantic similarity based on the length of the shortest path between nodes.

One of the most known unsupervised method applied for word sense disambiguation is PageRank method (Agirre and Soroa, 2009; Duque et al., 2018), which was initially proposed for calculating authoritative Internet pages and based on page links (Page et al., 1999). In word sense disambiguation, PageRank is applied to graph-based semantic resources such as WordNet.

2.2 Word Sense Disambiguation in Russian

For Russian, in (Loukachevitch and Chuiko, 2007) the authors studied the all-word disambiguation task on the basis of the RuThes thesaurus (Loukachevitch et al., 2018) - resource for natural language processing of Russian texts. They experimented with various parameters (types of the thesaurus paths, window size, etc). The work (Kobritsov et al., 2005) describes developed disambiguation filters to provide semantic annotation for the Russian National Corpus. The semantic annotation was based on the taxonomy of lexical and semantic facets. In (Mitrofanova and Lyashevskaya, 2009) statistical word sense disambiguation methods for several Russian nouns were described. Alexeyevsky and Temchenko (Alexeyevsky and Temchenko, 2016) tested a number of algorithms based on parsing of monolingual dictionaries.

In (Bolshina and Loukachevitch, 2020a) the authors study an approach to automatic semantic annotation of a text corpus based on so called "monosemous relatives" technique, which exploits monosemous related words. The proposed approach involves not only monosemous synonyms, hyponyms or hypernyms as usual, but also "far" relatives located up to four relations from the initial sense according to Russian wordnet RuWordNet (Loukachevitch et al., 2016). Gathered related words are then filtered according to corpus-based vector similarity to synsets corresponding senses of the target word. In such a way, the approach allows adapting to specific genre-specific or domain collections (Bolshina and Loukachevitch, 2020b).

In (Panchenko et al., 2018) the authors describe the results of the first shared task on word sense induction (WSI) for the Russian language. The participants were asked to group contexts of a given word in accordance with its senses that were not provided beforehand. For the task, new evaluation datasets based on sense inventories with different sense granularity were created. The contexts in the datasets were sampled from texts of Wikipedia, the academic corpus of Russian, and an explanatory dictionary of Russian. In the Russian SuperGLUE benchmark (Shavrina et al., 2020) the datasets from RUSSE-2018 were transformed into the Word-in-Context task, which is a binary classification task: given two sentences containing the same polysemous word, the task is to determine, whether the word is used in the same sense in both sentences, or not.

Thus we see that some research has been done for word sense disambiguation in Russian. But by this time there is no text corpus annotated with word senses. The above-mention annotation in the Russian National Corpus is based on general semantic categories, not specific word senses.

3 Sense-annotated collection

For creating a sense-annotated collection, we use texts collected in the OpenCorpora project [1]. The OpenCorpora corpus gathered Russian texts and develop several layers of annotation for the open use of these data by researchers (CC BY-SA license) (Bocharov et al., 2011). Currently, the Opencorpora corpus has a subcorpus with morphological annotation annotated by crowdsourcing. The morphological corpus was used for developing one of the most known Russian morphological analyzers PyMorphy2 (Korobov, 2015). But the OpenCorpora does not contain texts with word sense annotation.

3.1 RuWordNet

For word sense annotation, we use sense inventory of Russian lexical-semantics resource RuWordNet[2] (Loukachevitch et al., 2016; Nikishina et al., 2022). RuWordNet is a resource similar to WordNet (Miller et al., 1990). It was semi-automatically created from other Russian resource - RuThes thesaurus (Loukachevitch et al., 2018). As other WordNet-like resources, RuWordNet consists of synsets, connected with semantic relations. Current RuWordNet version includes more than 133 thousand Russian words and expressions of three parts

[1] http://opencorpora.org/
[2] ruwordnet.ru

Entity type	Count
Synset	59,905
Lexical entry	133,468
Word	71,365
Multiword expression	62,103
Sense	154,111
Synset relation	254,007
hypernym / hyponym	74,736
instance hypernym / hyponym	5,803
part holonym / meronym	3,450
antonym	922
entailment	1,033
cause	568
domain topic	38,608
POS synonym	44,898
Link to inter-lingual index	23,162
Definition	20,054

Table 1: RuWordNet statistics.

of speech: nouns, verbs and adjectives. RuWordNet contains more than 15 thousand ambiguous Russian words presented in more than 20 thousand synsets. Tables 1 presents detailed RuWordNet statistics.

3.2 Manual sense annotation

For sense annotation, texts of average length were selected from the OpenCorpora corpus, beginning from texts containing several sentences. The texts were subdivided into sentences, lemmatized, matched with RuWordNet lexical entries, and transformed into the text format covering maximal information, useful for selecting an appropriate word sense in context. The created format presents the following items in structure:

- sentence,
- list of words in a column,
- each word is associated with a lemma and a part of speech,
- list of senses for each word found in RuWordNet,
- each sense is provided with the synset name, synonyms and hypernyms, presenting several levels up along the RuWordNet hierarchy.

Main statistics of the annotated corpus is presented in Table 2.

Metric	Num
Documents	807
Sentences	6,751
Lemmas	109,893
Annotated lemmas	46,320
Lexical entries	17,126
Annotated lexical entries	10,683
RWN synsets	8,619

Table 2: Description of the collection.

4 Evaluation of WSD methods on the collection

We experimented with two approaches for Russian word sense disambiguation: unsupervised PageRank method and automatic pseudo-labeling based on 'monosemous relatives'.

4.1 Applying PageRank for Russian word sense disambiguation

The assumption is that it is possible to solve WSD task for Russian as well as for English using PageRank. However, a WordNet-like database should be used to correctly repeat all steps. RuWordNet enables us to apply it because its structure is close to the structure of original WordNet.

The main idea of PageRank is to calculate the relative importance of a node (rank) in the graph G. It may be calculated using a number of directed links incoming a considered node. Besides, the strength of the link from i to j depends on the rank of node i: the more important node i is, the more strength its votes will have. Alternatively, PageRank can also be viewed as the result of a random walk process, where the final rank of node i represents the probability of a random walk over the graph ending on node i, at a sufficiently large time.

The calculation of the PageRank vector Pr for N nodes of graph G is equivalent to resolving the following equation:

$$Pr = cM \cdot Pr + (1 - c) \cdot v$$

where M is $N \times N$ transition probability matrix, $M_{ij} = \frac{1}{d_i}$, d_i is the number of outbound links of node i. V is a $N \times 1$ vector whose elements are $\frac{1}{N}$ and c is the so called damping factor, a scalar value between 0 and 1. The first term of the sum represents the above-described voting scheme. The second term correspond to the probability of a surfer

Procedure	**Train**	**Test**
Random	63.9	63.6
Most frequent sense	85.7	71.1
Pseudo-labelling	73.6	74.1
Basic PPR	-	67.4
PPR with a subset of relations	-	71.1
(previous) & not incl. target word	-	73.7
(previous) & hyperparameter optimization (damping_factor=0.95, n_iter=30)	73.7	74.2
(previous) & sliding window optimization (w=5)	74.2	74.3
(previous) & collocations	75.0	75.4

Table 3: Precision of considered methods.

randomly jumping to any node, e.g. without following any paths on the graph. The second term in the equation can be be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible. It allows avoiding deadlocks and loops in the graph, thereby guaranteeing that PageRank calculation converges to a unique stationary distribution (Page et al., 1999).

In the traditional PageRank formulation the vector v assigns equal probabilities to all nodes in the graph in case of random jumps. However, the vector v can be modified to be non-uniform. For example, stronger probabilities can be assigned to certain kinds of nodes - creating so called Personalized PageRank (PPR) method (Haveliwala, 2003).

In (Agirre and Soroa, 2009), the authors applied the PPR algorithm to word sense disambiguation based on WordNet (Miller, 1995) and showed that the results are better than for other graph-based algorithms.

To apply the PPR algorithm, several steps should be performed:

1. Determine types of relations between synsets of WordNet-like resource to be used. Some relations may be weak and may add noise to this graph. It is proposed to save the following relations: part meronym, part holonym, instance hyponym, instance hypernym, hyponym, hypernym.

2. Convert this resource to a graph.

   (a) Each sense corresponds to a node,

   (b) Each selected relation corresponds to an edge.

3. Decide whether a target word will be included in this context graph while solving disambiguation or not. The main benefit of the first variant is that it is more computationally effective. However, it leads to a problem of importance increase of related senses in the context (Agirre and Soroa, 2009). In the second variant, for each target word W_i, initial probability mass is concentrated in the senses of the words surrounding W_i, but not in the senses of the target word itself, so that context words increase its relative importance in the graph (Agirre and Soroa, 2009).

4. Determine a sliding context window, i.e. a number of words before and after a target one to be considered as a context.

5. Set PPR hyperparameters – number of iterations and damping factor (probability of random jumps).

Changes in each of these steps lead to different realisations of this method. Then, a resulting algorithm is the following:

1. For each TEXT in COLLECTION:

   (a) For each TARGET_WORD in TEXT:

      i. Take CONTEXT_WORDS using WINDOW.

      ii. Insert CONTEXT_WORDS in a graph – create a directed link from them to their possible senses.

      iii. Declare PPR method and assign initial probability mass to nodes of CONTEXT_WORDS .

iv. Fit PPR on this graph.
v. Take all possible senses of TARGET_WORD and their final probabilities.
vi. Choose a sense with a maximum probability.

It can be seen from Table 1 that RuWordNet contains a large number of multiword expressions (collocations). For each collocation, senses of word components (sense_id) are described. For example, component senses of phrase "отвратительный на вид" (disgusting looking) are described as follows:

- <sense name="отвратительный" id="118920-A-145306" synset_id="118920-A"/>
- <sense name="вид" id="107545-N-134500" synset_id="107545-N"/>

Therefore the PPR algorithm may be modified using collocations from the RuWordNet knowledge base. Collocations can be inserted in a graph, they also may be considered as an additional information for disambiguation. There are two ways of introducing collocations into the algorithm implementation:

1. Take a sense for target word from an expression if it is a component of such expression in the given text.
2. Use tokens of collocations contained in the context to resolve disambiguation of other words.

The first method is simpler because it does not require to consider context while resolving disambiguation.

This method was implemented for both original and personalized ways. Moreover, hyperparameters were optimized and some of previously mentioned improvements were introduced. Results will be presented in the appropriate section.

4.2 Pseudo-labeling method

Automatic pseudo-labeling method is based on the monosemous relative technique. The related monosemous words or expressions can be located on the distance up to 4 RuWordNet relations from the initial sense (Bolshina and Loukachevitch, 2020a). For example, a single-sense co-hyponym can serve as a monosemous relative (2 relations).

We suppose that contexts of monosemous relatives can be appropriate for the target sense and we can use for training disambiguation models. Any monosemous relative in fact can be quite different in context of usage from the target sense, therefore additional check and selection of monosemous relatives are needed. The monosemous relatives of the target words are additionally scored in accordance to the cosine similarity between word2vec vector of the relative and averaged vector of so-called *synset nest*.

The synset nest represents a set of words (or phrases) most closely related to a particular sense of the target word, specifically target word synonyms and all the words from directly related synsets within two steps from the target word (Bolshina and Loukachevitch, 2020a). A fragment of the nest for the Russian word *taksa* ("dachshund") is as follows: *hunting dog, hunting dog, doggie, four-legged friend, dog, dog, terrier, dog, greyhound dog...* (translated from Russian).

The word2vec vectors can be calculated on different text collections, which allows tuning of relative selection on the specific genre of texts (Bolshina and Loukachevitch, 2020b). The pseudolabeling includes the following steps:

- selection of monosemous related words for each sense of ambiguous word in RuWordNet at the distance up to 4 relations from the sense synset,
- scoring monosemous relatives according to word2vec similarity to the synset nests for each word sense calculated on a selected text corpus,
- extraction of monosemous relatives' contexts for training a supervised model training taken in proportion to similarity scores between monosemous relatives and synset nest.

In the current study word2vec training and context extraction was implemented on a Russian news corpus (2 million documents). For each sense, 200 contexts originating from different monosemous relatives were extracted. For context representation, the ELMO model[3] was used. Logistic regression

[3]https://rusvectores.org/ru/models/

model was trained for disambiguation of each ambiguous word on the automatically annotated word sense contexts.

4.3 Results

The approaches described in this article were implemented on the created corpus. Moreover, different settings and hyperparameters were tried. Precision was calculated as a performance measuse of disambiguation methods. It was measured in two different ways: including one-sense words and not. This should be considered because a human annotator might indicate that there is no correct sense for this word (in the context, of course) in our knowledge base.

Some simple methods were considered as baselines. They include: the most frequent sense method and the random method. The sense annotated collection was randomly split on train and test sets (it makes sense only for a limited number of methods) to exclude over-fitting. Final results are presented in Table 3.

It can be seen that the most frequent sense method demonstrates the best performance on the training set and nearly the worst one on the test set. And it is notable that the unsupervised PPR method outperforms the supervised pseudo-labeling approach only when preliminary parameter setting and optimisation were conducted.

5 Conclusion

We presented a sense-annotated corpus for Russian. The total size of the corpus is 109,893 lemmas, out of which 46,320 ones are manually annotated by 8,619 RuWordNet synsets.

The obtained corpus was used as a test collection for evaluating two word-sense disambiguation methods: personalized PageRank and pseudo-labelling. The precision of PPR is 75.4% and the precision of pseudo-labelling is 74.1%.

Our future work will be undertaken in two directions: (1) Firstly, we are going to use the corpus not only as test data, but also as a training collection for supervised methods. (2) Secondly, we are going to further develop the corpus itself, including annotating multi-word expressions and publishing the corpus in the Linguistic Linked Open Data cloud.

The corpus has been published on GitHub: https://github.com/LLOD-Ru/OpenCorpora-RuWordNet.

Acknowledgments

This work is supported by the Russian Science Foundation, grant no. 19-71-10056. The work of Natalia Loukachevitch in manual and automatic word sense annotation is supported by a grant for research centers in the field of artificial intelligence (agreement identifier 000000D730321P5Q0002 dated November 2, 2021 No. 70-2021-00142 with ISP RAS).

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. *ACL 2018*, page 29.

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41.

Daniil Alexeyevsky and Anastasiya V Temchenko. 2016. Word sense disambiguation in monolingual dictionaries for building russian wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 9–14.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Victor Bocharov, Svetlana Bichineva, Dmitry Granovsky, Natalia Ostapuk, and Maria Stepanova. 2011. Quality assurance tools in the opencorpora project. *Computational linguistics and intellectual technologies*.

Angelina Bolshina and Natalia Loukachevitch. 2020a. All-words word sense disambiguation for russian using automatically generated text collection. *Cybernetics and Information Technologies*, 20(4):90–107.

Angelina Bolshina and Natalia Loukachevitch. 2020b. Comparison of genres in word sense disambiguation using automatically generated text collections. In *Fourth International Conference Computational Linguistics in Bulgaria*, page 155.

Andres Duque, Mark Stevenson, Juan Martinez-Romo, and Lourdes Araujo. 2018. Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial intelligence in medicine*, 87:9–19.

Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

Bors Kobritsov, Olga Lyashevskaya, and Olga Shemanaeva. 2005. Disambiguation of lexico-semantic ambiguity in news and newspaper-magazine texts: surface filters and statistical evaluation. *Proceedings of the Contest "Internet Mathematics*.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *International conference on analysis of images, social networks and texts*, pages 320–332. Springer.

Natalia Loukachevitch and Daria Chuiko. 2007. Thesaurus-based word sense disambiguation [avtomaticheskoe razreshenie leksicheskoy mnogoznachnosti na baze tezaurusnykh znaniy]. *Proceedings of the Contest "Internet Mathematics*, pages 108–117.

Natalia Loukachevitch, German Lashevich, and Boris V Dobrov. 2018. Comparing two thesaurus representations for russian. In *Proceedings of the 9th Global Wordnet Conference*, pages 34–43.

Natalia V Loukachevitch, German Lashevich, Anastasia A Gerasimova, Vladimir V Ivanov, and Boris V Dobrov. 2016. Creating russian wordnet by conversion. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue*, pages 405–415.

David Martinez, O Lopez de Lacalle, and Eneko Agirre. 2008. On the use of automatically acquired examples for all-nouns word sense disambiguation. *Journal of Artificial Intelligence Research*, 33:79–107.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

George A. Miller, Richard Beckwith, Christiane. Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Olga Mitrofanova and Olga Lyashevskaya. 2009. Disambiguation of taxonomy markers in context: Russian nouns. In *17th Nordic Conference of Computational Linguistics NODALIDA—2009, Odense, Denmark*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Mirella Lapata. 2009. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):678–692.

Irina Nikishina, Mikhail Tikhomirov, Varvara Logacheva, Yuriy Nazarov, Alexander Panchenko, and Natalia Loukachevitch. 2022. Taxonomy enrichment with text and graph vector representations. *Semantic Web*, 13(33):441–475.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Alexander Panchenko, Anastasiya Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. Russe'2018: a shared task on word sense induction for the russian language. *arXiv preprint arXiv:1803.05795*.

Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–245.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 338–344.