# Comparison of Genres in Word Sense Disambiguation using Automatically Generated Text Collections

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

**Natalia Loukachevitch**
Lomonosov Moscow State University, Moscow, Russia,
Kazan Federal University, Kazan, Russia
louk_nat@mail.ru

## Abstract

The best approaches in Word Sense Disambiguation (WSD) are supervised and rely on large amounts of hand-labelled data, which is not always available and costly to create. In our work we describe an approach that is used to create an automatically labelled collection based on the monosemous relatives (related unambiguous entries) for Russian. The main contribution of our work is that we extracted monosemous relatives that can be located at relatively long distances from a target ambiguous word and ranked them according to the similarity measure to the target sense. We evaluated word sense disambiguation models based on a nearest neighbour classification on BERT and ELMo embeddings and two text collections. Our work relies on the Russian wordnet RuWordNet.

**Keywords**: Word sense disambiguation, Russian dataset, Monosemous relatives.

## 1. Introduction

Word sense disambiguation (WSD) is one of the major challenges of computational semantics and it addresses the issue of lexical ambiguity. The aim of a WSD system is to identify the correct sense of a polysemous word in a context. This task has a wide range of potential applications including information retrieval, machine translation, and a knowledge graph construction. The training of well-performing supervised WSD algorithms involves a vast number of sense-labelled samples for each polysemous word in a language. There exist several hand-crafted sense-annotated datasets for English (Miller et al., 1993; Taghipour and Ng, 2015). However, this requirement is currently beyond reach in many languages and Russian is among them.

In this paper we present a knowledge-driven method based on the concept of monosemous relatives for the automatic generation of a training collection. We exploit a set of unambiguous words (or phrases) related to particular senses of a polysemous word. However, as it was noted in (Martinez et al., 2006), some senses of target words do not have monosemous relatives, and the noise can be introduced by some distant relatives. In our research we tried to address these issues.

In this work we proposed an extended and modified algorithm of training data generation based on monosemous relatives approach. The main contribution of this study is that we have expanded a set of monosemous relatives under consideration: in comparison with earlier approaches now they can be situated at greater distance from a target ambiguous word in a graph. Moreover, we have introduced a numerical estimation of a similarity between a monosemous relative and a particular sense of a target word which is further used in the development of the training collection. In order to evaluate the created training collections, we used contextualized word representations – ELMo (Peters et al., 2018)

and BERT (Devlin et al., 2019). We investigated the application of our algorithm to the training and test collections of different genres and their impact on the resulting performance of the WSD system[1].

The paper is organized as follows. In section two we review the related work. Section three is devoted to the data description. The fourth section describes the method applied to automatically generate and annotate training collections. The procedure of creating the collections is explained in the fifth section. In the sixth section we describe a supervised word sense disambiguation algorithm trained on our collected material and demonstrate the results obtained by four different models. In this section we also present a comparative analysis of the models trained on different kinds of train collections. Concluding remarks are provided in the seventh section.

## 2. Related Work

To overcome the limitations, that are caused by the lack of annotated data, several methods of generating and harvesting large train sets have been developed. There exist many techniques based on different kinds of replacements, which do not require human resources for tagging. The most popular method is that of monosemous relatives (Leacock et al., 1998). Usually WordNet (Miller, 1995) is used as a source for such relatives. WordNet is a lexical-semantic resource for the English language that contains a description of nouns, verbs, adjectives, and adverbs in the form of semantic graphs. All words in those networks are grouped into sets of synonyms that are called synsets.

Monosemous relatives are those words or collocations that are related to the target ambiguous word through some connection in WordNet, but they have only one sense, i.e. belong only to one synset. Usually, synonyms are selected as relatives but in some works hypernyms and hyponyms are chosen (Przybyła, 2017). Some researchers replace the target word with named entities (Mihalcea and Moldovan, 2000), some researchers substitute it with meronyms and holonyms (Seo et al., 2004). In the article (Yuret, 2007) a special algorithm was created in order to select the best replacement out of all words contained within synsets of the target word and neighbouring synsets. The algorithm described in (Mihalcea, 2002) to construct an annotated training set is a combination of different approaches: monosemous relatives, glosses and bootstrapping. Monosemous relatives can be also used in other tasks, for example, for finding the most frequent word senses in Russian (Loukachevitch and Chetviorkin, 2015). Other methods of automatic generation of training collections for WSD exploit parallel corpora (Taghipour and Ng, 2015), Wikipedia and Wiktionary (Henrich et al., 2012), topic signatures (Agirre and De Lacalle, 2004). (Pasini and Navigli, 2017) created large training corpora exploiting a graph-based method that took an unannotated corpus and a semantic network as an input.

Various supervised methods including kNN, Naive Bayes, SVM, neural networks were applied to word sense disambiguation (Navigli, 2009). Recent studies have shown the effectiveness of contextualized word representations for the WSD task (Wiedemann et al., 2019; Kutuzov and Kuzmenko, 2019). The most widely used deep contextualized embeddings are ELMo and BERT.

In ELMo (Embeddings from language models) (Peters et al., 2018) context vectors are computed in an unsupervised way by two layers of bidirectional LSTM, that take character embeddings from convolutional layer as an input. Character-based token representations help to tackle the problems with out-of-vocabulary words and rich morphology. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) has a different type of architecture, namely a multi-layer bidirectional Transformer encoder. During the pre-training procedure, the model is "jointly conditioning on both left and right context in all layers" (Devlin et al., 2019: 1). Since these contextualized word embeddings imply capturing polysemy better than any other representations and, thus, we employ them in our investigation.

## 3. Data

In our research as an underlying semantic network, we exploit Russian wordnet RuWordNet (Loukachevitch et al., 2016). It is a semantic network for Russian that has a WordNet-like structure. In total it contains 111.5 thousand of words and word combinations for the Russian language. RuWordNet was used to extract semantic relations (e.g. synonymy, hyponymy, etc.) between a target sense of a polysemous word and all the words (or phrases) connected to it, including those linked via

[1] The source code of our algorithm is publicly available at: https://github.com/loenmac/russian_wsd_data

distant paths. The sense inventory was also taken from this resource. RuWordNet contains 29297 synsets for nouns. There are 63014 monosemous and 5892 polysemous nouns in RuWordNet. Table 1 presents a summary of the number of senses per noun:

| **Number of senses of a polysemous noun** | **Number of nouns in RuWordNet** |
|---|---|
| 2 senses | 4271 |
| 3 senses | 997 |
| 4 senses | 399 |
| 5 senses | 149 |
| > 5 senses | 76 |
| **Total number of senses** | 14 357 |

Table 1: Quantitative characteristics of polysemous nouns in RuWordNet

We utilized two corpora in the research. A news corpus consists of news articles harvested from various news sources. The texts have been cleaned from HTML-elements or any markup. Another corpus is Proza.ru, a segment of Taiga corpus (Shavrina and Shapovalova, 2017), which is compiled of works of prose fiction. We exploit these two corpora because we want to investigate whether the genre of the training corpus has an impact on the performance on the test dataset.

For evaluation of our algorithm of training data generation, we used three distinct RUSSE'18 datasets for Russian (Panchenko et al., 2018) that were created for the shared task on word sense induction for the Russian language. The first dataset is compiled from the contexts of the Russian National Corpus[2]. The second dataset consists of the contexts from Wikipedia articles. And the last dataset is based on the Active Dictionary of the Russian Language (Apresyan et al., 2017) and contains contexts taken from the examples and illustration sections from this dictionary. All the polysemous words are nouns.

| **Explanation** | **Number of words** | **Example** |
|---|---|---|
| A word has only one sense in RuWordNet | 34 | The word двойник (*dvojnik*, "doppelganger") has only one sense in RuWordNet whereas in RUSSE'18 it has 4. |
| A word is missing in the RuWordNet vocabulary | 9 | The word гипербола (*giperbola*, "hyperbole"). |
| The senses from RuWordNet and RUSSE'18 dataset have only one sense in common | 4 | The word мандарин (*mandarin*) has two senses in RUSSE'18: its sense "tangerine" is included in the thesaurus, whereas its sense "mandarin, bureaucrat" is absent. |
| Controversial cases of sense mapping | 29 | The word демократ (*democrat*, "democrat") has 2 senses: "supporter of democracy" and "a member of the Democratic Party". But there's another one in RUSSE'18: "a person of a democratic way of life, views". |
| Not enough examples for senses in the corpora | 2 | Words карьер (*kar`er*, "quarry/a very fast gallop") and шах (*shax*, "shah/check"). |
| Words with morphological homonymy | 1 | The word суда (*suda*, "court (Gen, Sg)/ship (Nom, Pl)") can have two distinct lemmas. |

Table 2: Cases when a word from RUSSE'18 dataset was not included in the final test dataset

[2] http://www.ruscorpora.ru/new/index.html

From the RUSSE dataset we excluded some polysemous words, and in Table 2 we overview the common reasons why it was done. The final list of the target ambiguous words contains 30 words in total, each having two different senses. All the texts with the target ambiguous nouns in this dataset have sense annotation. We will call the resulting test dataset RUSSE-RuWordNet because it is a projection of RUSSE'18 sense inventory on the RuWordNet data.

We also created a small training dataset, that consists of the word sense definitions and examples of uses from Ozhegov dictionary (Ozhegov, 2014) for every target polysemous word. This training data is utilized as a baseline for the WSD task. In this set each sense of an ambiguous word has one definition and between 1 and 3 usage examples.

Table 3 demonstrates quantitative characteristics of all of the above-mentioned corpora:

| | **Taiga-Proza.ru** | **News Corpus** | **RUSSE-RuWordNet** | **Dictionary Corpus (Baseline)** |
|---|---|---|---|---|
| **Number of sentences** | 32,8 million | 24,2 million | 2 103 | 144 |
| **Number of lemmas** | 246,8 million | 288,1 million | 39 311 | 657 |
| **Number of unique lemmas** | 2,1 million | 1,4 million | 12 110 | 475 |

Table 3: Quantitative characteristics of the corpora and datasets used in the experiments

## 4. Candidate Selection and Ranking Algorithm

The central idea of our method is based on the assumption that a training collection can be built not only with the direct relations like synonymy, hypernymy and hyponymy but also with far more distant words, such as co-hyponyms. For example, most contexts for the word крона (*krona*) in the sense "krona, currency" match the contexts of the other words denoting currency like английский фунт (*anglijskij funt*, "pound sterling") as they have common hypernym валюта (*valyuta*, "currency").

The principal features of our approach are as follows:

1. We take into consideration not only the closest relatives to a target word sense, as it was done in previous works, but also more distant relatives.
2. We utilize similarity scores between a candidate monosemous relative and synsets close to a sense of a target polysemous word in order to evaluate how well this candidate can represent the sense of an ambiguous word.
3. We introduce the notion of ***a nest*** that is used to assess the potential of a candidate's usage contexts for displaying target sense of a polysemous word. In order to measure the relevance and suitability of a monosemous candidate, we exploit a thesaurus set of words similar to a target sense. The group of synonyms to a target sense and all the words from directly related synsets within 2 steps from a target word comprise ***the nest*** for a target sense.
4. We check similarity scores to the nest for both closest and further located monosemous relatives because a word described as monosemous in the thesaurus can actually have polysemous usage in a corpus. For example, Russian word ириска (*iriska*, "toffee") can also denote a nickname of Everton Football Club (The Toffees) (Loukachevitch, 2019). Thus, all candidate monosemous relatives should be further checked on the source corpus.
5. We propose two distinct methods of compiling a training collection based on the monosemous relatives rating.

A target word sense is a sense of a polysemous word that we want to disambiguate. Candidate monosemous relatives are unambiguous words (or phrases), that can be located in up to four-step relation paths to a polysemous word and include co-hyponyms, two-step (or more) hyponyms and hypernyms. We consider the words (or phrases), that have more than 50 occurrences in the corpus.

A fragment of the nest for the word *такса* (*taksa*, "dachshund") is given below:

(1) охотничий пёс (*oxontichij pyos*, "hunting dog"), пёсик (*pyosik*, "doggie"), четвероногий друг (*chetveronogij drug*, "four-legged friend"), собака (*sobaka*, "dog"), терьер (*ter`er*, "terrier") … etc.

The choice of the distance constant for the nest was motivated by the fact that the senses of the relatives located at the 2-step relation path are close to the target sense of the polysemous word and, thus, these relatives are more reliable and do not require sophisticated additional verification. As for the distance used to extract candidate monosemous relatives, we decided to stick to the maximum distance of 4, because usually the words located at 5 or more steps from the target sense are too generic. For example, the monosemous candidate for the word такса (*taksa*, "dachshund") located at 4-step path is животное (*zhivotnoe*, "animal") and the candidate at the 5-step path is биологический организм (*biologicheskij organizm*, "biological organism"). We can see that the second word is more general and can be used in a wide variety of contexts, and many of them may not at all be related to animals and dogs in particular. Another similar example is гвоздика (*gvozdika*, "clove"): its 4-step relative is продовольственные продукты (*prodovolstvenny`je producty* "food products") and 5-step relative is вещество (*veshhestvo* "substance").

Our method of extracting monosemous relatives is based on comparison of distributional and thesaurus similarities. Embedding models are utilized to select the most appropriate monosemous relatives whose contexts serve as a good representation of a target word sense. We used the word2vec models to extract 100 most similar words to each monosemous word from the candidates list. In that way, we collected the words that represent a distributional set of close words with the respective cosine similarities measures. Our selection and ranking method, thus, consists of the following steps:

1. We extract all the candidate monosemous relatives within 4 steps from a target polysemous word sense $\boldsymbol{s_j}$.
2. We compile the nest $\boldsymbol{ns_j}$ which consists of synonyms to a target sense and all the words from the synsets within 2 steps from a target word $\boldsymbol{s_j}$. The nest $\boldsymbol{ns_j}$ consists of $\boldsymbol{N_k}$ synsets.
3. For each candidate monosemous relative $\boldsymbol{r_j}$, we find 100 most similar words according to the word2vec model trained on a reference corpus.
4. We intersect these top-100 words with the words included in the nest $\boldsymbol{ns_j}$ of the target sense $\boldsymbol{s_j}$.
5. For each word in the intersection, we take its cosine similarity weight calculated with the word2vec model and assign it to the synset it belongs to. The final weight of the synset in the nest $\boldsymbol{ns_j}$ is determined by the maximum weight among the words $\boldsymbol{w^j_{k_1}, \dots, w^j_{k_i}}$ representing this synset in the intersection.
6. The total score of the monosemous candidate $\boldsymbol{r_j}$ is the sum of the weights of all synsets from the nest $\boldsymbol{ns_j}$. In such a way more scores are assigned to those candidates, that resemble a greater number of synsets from the nest close the target sense of the ambiguous target word. Thus, the final weight of the candidate can be defined as follows:

$$Weight_{r_j} = \sum_{k=1}^{N_k} \max\left[cos\left(r_j, w^j_{k_1}\right), \dots, cos\left(r_j, w^j_{k_i}\right)\right]$$

The following fragment of list of monosemous relatives with similarity scores (given in brackets) was obtained for the noun гвоздика (*gvozdika*, "clove"):

(2) мускатный орех (*muskatny`j orex*, "nutmeg") (6), имбирь (*imbir`*, "ginger") (6.4), корица (*korica*, "cinnamon") (6.5), кардамон (*kardamon*, "cardamom") (6.8), чёрный перец (*cherny`j perec*, "black pepper") (7.5)… etc.

We have also found some examples where a monosemous word is connected to a sense of a target word but got zero similarity weight. For example, the word марля (*marlya*, "gauze") is a cohyponym to the word байка in the sense (*bajka*, "thick flannelette") but was not included in the monosemous relatives list because its distributional set of close words did not have any intersection with the nest.

As a result of this procedure, all monosemous relatives are sorted by the weight they obtained. The higher-rated monosemous relatives are supposed to be better candidates to represent the sense of the target word and, consequently, their contexts of use are best suited as the training examples in the WSD task. The candidate ranking algorithm identifies which monosemous relatives are most similar to the target ambiguous word's sense. Once we have detected the monosemous candidates, we can extract from the corpus the contexts in which they occur. Then in these texts we substitute the monosemous relatives with the target ambiguous word and add the texts with the respective sense labels to a training collection.

In order to verify the applicability of our method to the RuWordNet material, we found candidate monosemous relatives for the ambiguous words in the thesaurus using our algorithm but without word2vec filter. Only two words out of 5895 do not have monosemous relatives within four-step relation path in RuWordNet graph. The quantitative characteristics of the candidate monosemous relatives are presented in Table 4. As it was mentioned in (Taghipour and Ng, 2015: 339), 500 samples per sense is enough for training data. Table 5 demonstrates how many target senses have at least 500 samples of their monosemous relatives in a reference corpus. We also take into consideration the case when word2vec filter was applied to the candidate monosemous relatives. These tables show that by applying our approach to the RuWordNet data we would be able to find monosemous relatives to almost all the polysemous words in the thesaurus and create a training collection for a WSD system.

| **Distance to a candidate monosemous relative** | **Number of target senses, that have at least one relative at this distance** |
|---|---|
| 0 (synset) | 9 818 |
| 1 | 13 095 |
| 2 | 14 129 |
| 3 | 14 021 |
| 4 | 13 768 |

Table 4: Quantitative characteristics of candidate monosemous relatives for RuWordNet target senses

| | **Number of target senses when word2vec filter was not applied** | **Number of target senses when word2vec filter was applied** |
|---|---|---|
| **Taiga-Proza.ru** | 13 738 | 12 797 |
| **News Corpus** | 14 017 | 13 099 |

Table 5: Target senses with more than 500 occurrences of monosemous relatives in the corpora.

## 5. Generating Training Data using Monosemous Relatives

For comparison, we decided to create two separate training collections compiled from the news and Proza.ru corpora, and we also exploited two distinct approaches to a collection generation. According to the first method, we compiled the collection only with a monosemous relative from the top of the candidate rating. We wanted to obtain 1000 examples for each of the target words, but sometimes it was not possible to extract so many contexts with one particular candidate. That is why in some cases we also took examples with words next on the candidates' list. For simplicity, we call this collection Corpus-1000 because we obtained exactly 1000 examples for each sense.

The second approach enables to harvest more representative collection with regard to the variety of contexts. The training examples for the target ambiguous words were collected with the help of all respective unambiguous relatives with non-zero weight. The number of extracted contexts per a monosemous candidate is in direct proportion to its weight. We name this collection a balanced one because the selection of training examples was not restricted to the contexts which have only one particular monosemous relative.

In Table 6 we present the quantitative characteristics of the two collections, such as the relations connecting the target senses and their monosemous relatives, distances between them, and a proportion of monosemous relatives expressed as a phrase.

| Feature | Proportion of occurrences in the news collection | Proportion of occurrences in Proza.ru collection |
|---|---|---|
| **Distance to a target sense** | | |
| 0 (synset) | 2% | 4% |
| 1 | 13% | 9% |
| 2 | 38% | 37% |
| 3 | 31% | 34% |
| 4 | 16% | 16% |
| **Relation between a target sense and a monosemous relative** | | |
| Synonyms | 2% | 4% |
| Hyponyms | 13% | 8% |
| Hypernyms | 11% | 9% |
| Cohyponyms | 28% | 28% |
| Cohyponyms situated at three-step path | 24% | 28% |
| Cohyponyms situated at four-step path | 19% | 22% |
| Other | 3% | 1% |
| **Word combinations** | 48% | 29% |

Table 6: Quantitative characteristics of monosemous relatives

Two word2vec embedding models that we used in our experiments were trained separately on the news and Proza.ru corpora with the window size of 3. As a preprocessing step, we split the corpora into separate sentences, tokenized them, removed all the stop words, and lemmatized the words with pymorphy2 tool (Korobov, 2015). The words obtained from the word2vec model were filtered out – we removed the ones not included in the thesaurus.

## 6. Experiments

We conducted several experiments to determine which text collection used as training data for a WSD model gives the best performance on the test dataset. Following (Wiedemann et al., 2019), in our research we used an easily interpretable classification algorithm – non-parametric nearest neighbor classification (kNN) based on the contextualized word embeddings ELMo and BERT.

In our experiments we exploited two distinct ELMo models - the one trained by DeepPavlov on Russian WMT News and the other is RusVectōrēs (Kutuzov and Kuzmenko, 2017) lemmatized ELMo model trained on Taiga Corpus (Shavrina and Shapovalova, 2017). The difference between these two models is that from the first model we extracted a vector for a whole sentence with a target word, whereas from the second model we extracted a single vector for a target ambiguous word. As for BERT, we used two models: BERT-base-multilingual-cased released by Google Research and RuBERT, which was trained on the Russian part of Wikipedia and news data by DeepPavlov (Kuratov and Arkhipov, 2019). To extract BERT contextual representations, we followed the method described by (Devlin et al., 2019) and (Wiedemann et al., 2019) and concatenated "the token representations from the top four hidden layers of the pre-trained Transformer" (Devlin et al., 2019: 9).

The Tables 7 and 8 demonstrate the results obtained by different types of contextualized word embeddings, the training collections, and model parameters.

| Model | ELMo RusVectōrēs (target word) | | ELMo DeepPavlov (whole sentence) | | RuBERT DeepPavlov | | Multilingual BERT | |
|---|---|---|---|---|---|---|---|---|
| k | Proza.ru | News collection | Proza.ru | News collection | Proza.ru | News collection | Proza.ru | News collection |
| 1 | 0.809 | 0.794 | 0.765 | **0.752** | 0.751 | 0.735 | 0.668 | 0.67 |
| 3 | 0.826 | 0.811 | **0.773** | 0.749 | 0.781 | 0.756 | 0.684 | 0.673 |
| 5 | 0.834 | **0.819** | 0.77 | 0.748 | 0.793 | 0.771 | 0.694 | 0.667 |
| 7 | **0.841** | **0.819** | 0.767 | 0.746 | **0.804** | **0.774** | 0.699 | 0.673 |
| 9 | 0.84 | 0.816 | 0.762 | 0.747 | 0.802 | 0.769 | **0.7** | **0.677** |
| **Baseline** | 0.772 | | 0.716 | | 0.667 | | 0.672 | |

Table 7: F1 scores for ELMo- and BERT-based WSD models, Corpus-1000 collections

| Model | ELMo RusVectōrēs (target word) | | ELMo DeepPavlov (whole sentence) | | RuBERT DeepPavlov | | Multilingual BERT | |
|---|---|---|---|---|---|---|---|---|
| k | Proza.ru | News collection | Proza.ru | News collection | Proza.ru | News collection | Proza.ru | News collection |
| 1 | 0.812 | 0.797 | 0.745 | 0.758 | 0.746 | 0.75 | 0.669 | 0.662 |
| 3 | 0.833 | 0.81 | 0.775 | 0.753 | 0.778 | 0.755 | 0.707 | 0.681 |
| 5 | 0.845 | 0.81 | 0.776 | 0.756 | 0.792 | 0.769 | 0.717 | 0.682 |
| 7 | **0.857** | 0.815 | **0.793** | **0.759** | 0.802 | 0.768 | 0.723 | 0.683 |
| 9 | 0.856 | **0.821** | 0.791 | 0.753 | **0.812** | **0.774** | **0.729** | **0.688** |
| **Baseline** | 0.772 | | 0.716 | | 0.667 | | 0.672 | |

Table 8: F1 scores for ELMo- and BERT-based WSD models, balanced collections

As it can clearly be seen, all the systems surpassed the quality level of the baseline solution trained on the dataset of the dictionary definitions and usage examples. This means that we have managed not only to collect training data sufficient to train the WSD model but also to show a good performance on the RUSSE-RuWordNet dataset.

The Proza.ru model achieves better results and outperforms the news model. The qualitative analysis of the classification errors caused by the model trained on the news collection showed that the main cause of mistakes were lexical and structural differences between training and test sets. The examples from the test dataset were from the Russian National Corpus and Wikipedia, whereas the training collections were composed of news articles. On the contrary, Proza.ru collection consists of various works of fiction, so, the training samples have more similar representations to the test ones. We thus conclude that similar genres of train and test collections give higher results in the WSD task.

The algorithm based on the ELMo pre-trained embeddings by RusVectōrēs outperformed all other models achieving 0.857 F1 score. The second-best model in the WSD task is RuBERT by DeepPavlov, followed by ELMo model by DeepPavlov. The lowest F1 score belongs to Multilingual BERT. As for the difference in F1 scores between the Corpus-1000 and the balanced collection, we can observe the performance drop for the Corpus-1000 for all the models, which means that the approach used to generate the balanced collection is better suited for the task. Corpus-1000 does not include all possible monosemous relatives, so the collection lacks contextual diversity, the balanced collection, on the contrary, is more representative with regard to the variety of contexts.

## 7. Conclusion

The issue that we addressed in this article is the lack of sense-annotated training data for supervised WSD systems in Russian. In this paper we have described our algorithm of automatic collection and

annotation of training data for the Russian language. The main contribution of the paper is that we have utilized in the selection algorithm not only close monosemous relatives but also more distant ones. Moreover, we implemented the procedure of ranking monosemous relatives' candidates. Our training collections consist of the texts extracted from the news and Proza.ru corpora. The candidate scores were obtained from two word2vec models trained separately on each corpus.

In order to evaluate the training collections, we applied kNN classifier to the contextualized word embeddings extracted for target polysemous words and measured its performance on the RUSSE-RuWordNet test dataset. We have investigated the capability of different deep contextualized word representations to model polysemy. The best result was obtained with RusVectōrēs ELMo model and amounted to 0.857 F1 score. We have also found out that the training collection harvested from the Proza.ru corpus gave higher F1 scores on the RUSSE-RuWordNet test dataset than the collection from the news corpus.

## Acknowledgements

The work of Loukachevitch N. in the current study concerns formulation of the disambiguation approach for RuWordNet data, calculation of paths between synsets, criteria for selecting contexts; this work is supported by the Russian Science Foundation grant no. 19-71-10056 financed through Kazan Federal University.

## References

Agirre, E., De Lacalle, O. L. (2004). *Publicly Available Topic Signatures for all WordNet Nominal Senses*. In LREC.

Apresyan, V. Yu., Apresyan, Yu. D., Babaeva, E. E., Boguslavsaya, O. Yu., Glovinskaya, M. Ya., Iomdin, B. L., Krylova, T. V., Levontina, I. B., Lopukhina, A. A., Ptentsova, A. V., Sannikov, A. V., Uryson, E. V. (2017). *Active Dictionary of the Russian Language* [*Aktivny`j slovar` russkogo yazyka*]. Publishing House Nestor-Istoria, Moscow, Vol. 3.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. In Proc. of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 4171-4186.

Henrich, V., Hinrichs, E., Vodolazova, T. (2012). *Webcage: A Web-harvested Corpus Annotated With GermaNet Senses*. In Proc. of the 13th Conf. of the European Chapter of the ACL, pp. 387-396.

GOST 7.79-2000: *System of Standards for Information, Library Services and Publishing. Rules for Transliteration of Cyrillic Letters into the Latin Alphabet.*

Korobov, M. (2015). *Morphological Analyzer and Generator for Russian and Ukrainian Languages*. In Analysis of Images, Social Networks and Texts, pp. 320-332.

Kuratov, Y., Arkhipov, M. (2019). *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*. arXiv preprint arXiv:1905.07213.

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, Cham.

`https://rusvectores.org/ru/`

Kutuzov, A., Kuzmenko, E. (2019). *To Lemmatize or not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation*. In Proc. of the First NLPL Workshop on Deep Learning for Natural Language Processing, pp. 22-28.

Leacock, C., Miller, G. A., Chodorow, M. (1998). *Using Corpus Statistics and WordNet Relations for Sense Identification*. Computational Linguistics, vol. 24(1), pp. 147-165.

Loukachevitch, N. (2019). *Corpus-Based Check-Up for Thesaurus*. In Proc. of the 57th Annual Meeting of the ACL, pp. 5773-5779.

Loukachevitch, N. V., Lashevich, G., Gerasimova, A. A., Ivanov, V. V., Dobrov, B. V. (2016). *Creating Russian WordNet by Conversion*. In Proc. of Conference on Computational linguistics and Intellectual technologies Dialog-2016, pp. 405-415.

Loukachevitch, N., Chetviorkin, I. (2015). *Determining the Most Frequent Senses Using Russian Linguistic Ontology RuThes*. In Proc. of the workshop on Semantic resources and semantic annotation for NLP and the Digital Humanities at NODALIDA 2015, pp. 21-27.

Martinez, D., Agirre, E., Wang, X. (2006). *Word Relatives in Context for Word Sense Disambiguation*. In Proc. of the Australasian Language Technology Workshop 2006, pp. 42-50.

Mihalcea, R. (2002). *Bootstrapping Large Sense Tagged Corpora*. In Proc. of the Third International Conference on Language Resources and Evaluation (LREC-2002), vol. 1999.

Mihalcea, R., Moldovan, D. I. (2000). *An Iterative Approach to Word Sense Disambiguation*. In FLAIRS Conference, pp. 219-223.

Miller, G. (1995). *WordNet: A Lexical Database for English*. In Communications of the ACM, vol.38(11), pp. 39-41.

Miller, G. A., Leacock, C., Tengi, R., Bunker, R. T. (1993). *A Semantic Concordance*. In Proc. of the workshop on Human Language Technology, pp. 303-308.

Navigli, R. (2009). *Word Sense Disambiguation: A survey*. ACM computing surveys (CSUR), vol. 41(2), 10.

Ozhegov, S.I. (2014). *Explanatory Dictionary of the Russian Language* [*Tolkovy`j Slovar` Russkogo Yazyka*]. Edited by Skvortsova S.I., 8, pp. 1376.

Panchenko, A., Lopukhina, A., Ustalov, D., Lopukhin, K., Arefyev, N., Leontyev, A., Loukachevitch, N. (2018). *RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language*. In Computational Linguistics and Intellectual Technologies: Dialogue-2018, pp. 547–564.

Pasini, T., Navigli, R. (2017). *Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages Without Manual Training Data*. In Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 78-88.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). *Deep Contextualized Word Representations*. In Proc. of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 2227–2237.

Przybyła, P. (2017). *How Big is Big Enough? Unsupervised Word Sense Disambiguation Using a Very Large Corpus*. arXiv preprint arXiv:1710.07960.

Seo, H. C., Chung, H., Rim, H. C., Myaeng, S. H., Kim, S. H. (2004). *Unsupervised Word Sense Disambiguation Using WordNet Relatives*. Computer Speech & Language SPEC. ISS., vol. 18, no. 3, pp. 253-273.

Shavrina, T., Shapovalova, O. (2017). *To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser.* In Proc. of "CORPORA2017", international conference, Saint-Petersburg.

Taghipour, K., Ng, H. T. (2015). *One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction*. In Proc. of the 19th Conf. on computational natural language learning, pp. 338-344

Wiedemann, G., Remus, S., Chawla, A., Biemann, C. (2019). *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings*. arXiv preprint arXiv:1909.10430.

Yuret, D. (2007). *KU: Word Sense Disambiguation by Substitution*. In Proc. of the 4th International Workshop on Semantic Evaluations, pp. 207-213.