Determining the Most Frequent Senses Using Russian Linguistic Ontology RuThes

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

Ilia Chetviorkin
Lomonosov Moscow State University
Moscow, Russia
Ilia2010@yandex.ru

Abstract

The paper describes a supervised approach for the detection of the most frequent senses of words on the basis of RuThes thesaurus, which is a large linguistic ontology for Russian. Due to the large number of monosemous multiword expressions and the set of RuThes relations it is possible to calculate several context features for ambiguous words and to study their contribution to a supervised model for detecting frequent senses.

1 Introduction

The most frequent sense (MFS) is a useful heuristic in lexical sense disambiguation when the training or context data are insufficient. In sense-disambiguation (WSD) evaluations the first sense labeling is presented as an important baseline (Agirre et al., 2007), which is difficult to overcome for many WSD systems (Navigli, 2009).

Usually MFS is calculated on the basis of a large sense-tagged corpus such as SemCor, which is labeled with WordNet senses (Landes et al., 1998). However, the creation of such corpora is a very labor-consuming task. Besides Princeton WordNet (Fellbaum, 1998), only for several other national wordnets such corpora are labeled (Perolito and Bond, 2014). In addition, the MFS of a given word may vary according to the domain, therefore the automatic processing of documents in a specific domain can require re-estimation of MFS on the domain-specific text collection. The distributions of lexical senses also can depend on time (Mitra et al., 2014).

Therefore automatic calculation of the most frequent sense is studied in several works (Mohammad and Hirst, 2006; McCarthy et al., 2004; McCarthy et al., 2007). One of the prominent approaches in this task is to use distributional vectors to compare contexts of an ambiguous word with sense-related words (Koeling et al., 2005; McCarthy et al., 2007). In such experiments mainly WordNet-like resources are studied. In (Mohammad, Hirst, 2006), the Macquarie Thesaurus serves as a basis for the predominant sense identification.

In this paper we present our experiments demonstrating how unambiguous multiword expressions can help to reveal the most frequent sense if they are allowed to be included in a thesaurus. The experiments are based on newly-published Thesaurus of Russian language RuThes-lite, which has been developed since 1994 and was applied in a number of tasks of natural language processing and information retrieval (Loukachevitch and Dobrov, 2014).

This paper is organized as follows. Section 2 compares our study with related works. In Section 3, we describe the main principles of RuThes-lite linguistic ontology construction. Section 4 is devoted to the manual analysis of the distribution of word senses described in RuThes, which is performed on the basis of Russian news flow provided by Yandex news service. Section 5 describes the experiments on supervised prediction of the most frequent sense of an ambiguous word.

2 Related Work

It was found in various studies that the most frequent sense is a strong baseline for many NLP tasks. For instance, only 5 systems of the 26 submitted to the Senseval-3 English all words task outperformed the reported 62.5% MFS baseline (Snyder and Palmer, 2004).

However, it is very difficult to create sense-labeled corpora to determine MFS, therefore techniques for automatic MFS revealing were proposed. McCarthy et al. (2004, 2007) describe

Natalia Loukachevitch and Ilia Chetviorkin 2015. Determining the most frequent senses using Russian linguistic ontology RuThes. *Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015.* NEALT Proceedings Series 27 / Linköping Electronic Conference Proceedings 112: 21–27.

an automatic technique for ranking word senses on the basis of comparison of a given word with distributionally similar words. The distributional similarity is calculated using syntactic (or linear) contexts and the automatic thesaurus construction method described in (Lin, 1998). WordNet similarity measures are used to compare the word senses and distributional neighbors. McCarthy et al. (2007) report that 56.3% of noun SemCor MFS (random baseline – 32%), 45.6% verb MFS (random baseline – 27.1%) were correctly identified with the proposed technique.

In (Koeling et al., 2005) the problem of domain specific sense distributions is studied. They form samples of ambiguous words having a sense in one of two domains: SPORTS and FINANCE. To obtain the distribution of senses for chosen words, the random sentences mentioning the target words in domain-specific text collections are extracted and annotated.

Lau et al. (2014) propose to use topic models for identification of the predominant sense. They train a single topic model per target lemma. To compute the similarity between a sense and a topic, glosses are converted into a multinomial distribution over words, and then the Jensen – Shannon divergence between the multinomial distribution of the gloss and the topic is calculated.

Mohammad and Hirst (2006) describe an approach for acquiring predominant senses from corpora on the basis of the category information in the Macquarie Thesaurus.

A separate direction in WSD research is automatic extraction of contexts for ambiguous words based on so called "monosemous relatives" (Leacock et al., 1998; Agirre and Lacalle, 2004; Mihalcea 2002) that are related words having only a unique sense. It was supposed that extracted sentences mentioning monosemous relatives are useful for lexical disambiguation. These approaches at first determine monosemous related words for a given ambiguous word, then extract contexts where the relatives were mentioned, and use these contexts as automatically annotated data to train WSD classifiers.

In our case we use monosemous relatives in another way: to determine the most frequent senses of ambiguous words. We conduct our research for Russian and this is the first study on MFS prediction for Russian.

3 RuThes Linguistic Ontology

One of the popular resources used for natural language processing and information-retrieval applications is WordNet thesaurus (Fellbaum, 1998). Several WordNet-like projects were also initiated for Russian (Balkova et al., 2008; Azarowa, 2008; Braslavski et al. 2013). However, at present there is no large enough and qualitative Russian wordnet. But another large resource for natural language processing – RuThes thesaurus, having some other principles of its construction, has been created and published. The first publicly available version of RuThes (RuThes-lite) contains 96,800 unique words and expressions and is available from http://www.labinform.ru/ruthes/index.htm .

RuThes Thesaurus of Russian language is a linguistic ontology for natural language processing, i.e. an ontology, where the majority of concepts are introduced on the basis of actual language expressions. RuThes is a hierarchical network of concepts. Each concept has a name, relations with other concepts, a set of language expressions (words, phrases, terms) whose senses correspond to the concept, so called ontological synonyms.

Ontological synonyms of a concept can comprise words belonging to different parts of speech (*stabilization, stabilize, stabilized*); language expressions relating to different linguistic styles, genres; idioms and even free multiword expressions (for example, synonymous to single words).

So a row of ontological synonyms can include quite a large number of words and phrases. For instance, the concept *ДУШЕВНОЕ СТРАДАНИЕ* (*wound in the soul*) has more than 20 text entries (several English translations may be as follows: *wound, emotional wound, pain in the soul* etc.).

Besides, in RuThes introduction of concepts based on multiword expressions is not restricted and even encouraged if this concept adds some new information to knowledge described in RuThes. For example, a concept such as *ЗАСНУТЬ ЗА РУЛЕМ* (*falling asleep at the wheel)* is introduced because it denotes a specific important situation in road traffic, has an "interesting" text entry *заснуть во время движения* (*falling asleep while driving*). Also, this concept has an "interesting" relation to concept *ДОРОЖНО-ТРАНСПОРТНОЕ ПРОИСШЕСТВИЕ* (*road accident*) (Loukachevitch and Dobrov, 2014). The word

"interesting" means here that the synonym and the relation do not follow from the component structure of phrase *заснуть за рулем*.

Thus, RuThes principles of construction give the possibility to introduce more multiword expressions in comparison with WordNet-like resources.

An ambiguous word is assigned to several concepts – this is the same approach as in WordNet. For example, the Russian word *картина (picture)* has 6 senses in RuThes and attributed to 6 concepts.

- *ФИЛЬМ* (*moving picture*)
- *ПРОИЗВЕДЕНИЕ ЖИВОПИСИ* (*piece of painting*)
- *КАРТИНА* (*ОПИСАНИЕ*) (*picture as description*)
- *КАРТИНА СПЕКТАКЛЯ* (*scene as a part of a play*)
- *ЗРЕЛИЩЕ (ВИД)* (*sight, view*)
- *КАРТИНА ПОЛОЖЕНИЯ, СОСТОЯНИЯ* (*picture as general circumstances*)

The relations in RuThes are only conceptual, not lexical (in contrast to antonyms or derivational links in wordnets). The main idea behind the RuThes set of conceptual relations is to describe the most essential, reliable relations of concepts, which are relevant to various contexts of concept mentioning. The set of conceptual relations includes the class-subclass relation, the part-whole relation, the external ontological dependence, and the symmetric association (Loukachevitch and Dobrov, 2014).

Thus, RuThes has considerable similarities with WordNet including concepts based on senses of real text units, representation of lexical senses, detailed coverage of word senses. At the same time, the differences include attribution of different parts of speech to the same concepts, formulating names of concepts, attention to multiword expressions, a set of conceptual relations. A more detailed description of RuThes and RuThes-based applications can be found in (Loukachevitch and Dobrov, 2014).

4 Manual Analysis of Sense Distribution

To check the coverage of lexical senses described in RuThes we decided to verify their usage in a text collection. At this moment we do not have the possibility to create a sense-tagged corpus based on RuThes senses. In addition, as it was indicated in (Petrolito and Bond, 2014), in sense-labeling most time and efforts are spent on adding new word senses to a source resource. Another problem of a sense-labeled corpus is that it fixes the described sets of senses, and it is impossible to automatically update them for a new version of a thesaurus.

To verify the coverage of lexical senses described in RuThes, the most important issue is to check that at least frequent senses have been already described. With this aim, it is not necessary to label all senses of a word in a large text collection, it is enough to check out senses in a randomly selected sample of word usages in contemporary texts as it was made in (Koeling, 2005). In addition, from this analysis we obtain manual estimation of MFS.

We decided to check RuThes senses on news texts and articles through Yandex news service[1]. We based our evaluation on a news collection because news reports and articles are one of the most popular documents for natural language processing, such as categorization, clustering, information extraction, sentiment analysis. Besides, the news collection comprises a lot of other text genres as legal regulations or literature pieces. Finally, this collection contains recently appeared senses, which can be absent in any fixed collection such as, for example, Russian national corpus (Grishina and Rakhilina, 2005) and dictionaries.

Yandex.news service collects news from more than 4,000 sources (including main Russian Newspapers), receiving more than 100,000 news articles during a day. The news flow from different sources is automatically clustered into sets of similar news. When searching in the service, retrieval results are also clustered. Usually three sentences from the cluster documents (snippets) are shown to the user.

For a given ambiguous word, linguists analyzed snippets in Yandex news service, which returns the most recent news reports and newspaper articles containing the word. Considering several dozens of different usages of the word in news, the linguists estimated the distribution of senses of the word, which later would allow defining the most frequent sense of the word. In news snippets, repetitions of the

[1] http://news.yandex.ru/

same sentences can be frequently met – such repetitions were dismissed from the analysis. Table 1 presents the results of the analysis for Russian ambiguous words *провести (provesti), картина (kartina),* and *стрелка (strelka).* The sense distributions for these three words have quite different behavior. Word *провести* has a single predominant sense; word *картина* has two main senses with approximately similar frequencies. Word *стрелка* has three enough frequent senses.

Because of insufficient amount of data under consideration, the experts could designate several senses as the most frequent ones if they saw that the difference in the frequencies does not allow them to decide what a sense is more frequent. For example, for word *картина* two main senses were revealed: *ФИЛЬМ (moving picture)* and *ПРОИЗВЕДЕНИЕ ЖИВОПИСИ* (*piece of painting)* (Table 1).

Word	Name of concept corresponding to senses of the word	Number of contexts
Провести (provesti) 9 senses	*ПРОВЕСТИ, ОРГАНИЗОВАТЬ, УСТРОИТЬ* (organize)	19
	ПРОЛОЖИТЬ ЛИНИЮ, ПУТЬ (build road, pipe)	1
Картина (kartina) 6 senses	*ПРОИЗВЕДЕНИЕ ЖИВОПИСИ* (piece of painting)	10
	ФИЛЬМ (moving picture)	10
Стрелка (strelka) 7 senses	*СТРЕЛКА РЕК* (river spit)	8
	СТРЕЛКА ПРИБОРА (pointer of the device)	6
	ЗНАК СТРЕЛКИ (arrow sign)	4
	ЖЕЛЕЗНОДОРОЖ-НАЯ СТРЕЛКА (railroad point)	1
	СТРЕЛКА НА ЧАСАХ (clock hand)	1

Table 1. Sense distribution of several Russian ambiguous words in the news flow (20 different contexts in current news flow were analyzed)

In total, around 3,000 ambiguous words with three or more senses described in RuThes (11,450 senses altogether) were analyzed in such a manner. As a result of such work, about 650 senses (5.7%) were added or corrected. So the coverage of senses in RuThes was enough qualitative and improved after the analysis.

Certainly, the distribution of word senses in news service search results can be quite dependent on the current news flow; in addition, the subjectivity of individual expertise can appear. Therefore for 400 words the secondary labeling was implemented, which allows us to estimate inter-annotator (and inter-time) agreement. 200 words from these words had three senses described in RuThes, other 200 words had four and more described senses.

The table 2 demonstrates that for 88% of the words, experts agreed or partially agreed on MFS for the analyzed words (Kappa=0.83). The partial agreement means in this case that experts agreed on prominent frequency of at least one sense of a word and indicated other different senses as also prominent. For example, for word *картина,* the first expert indicated two main senses (*moving picture* and *piece of painting*) with equal frequencies. The second expert revealed that the *piece of painting* sense is much more frequent than other senses. Therefore we have here partial agreement between experts and suppose that the most frequent sense of a word is the *piece of painting* sense.

Number of words analyzed by two experts	400
Number of words for that experts agreed on MFS	216
Number of words for that experts partially agreed on MFS	125
Number of words for that experts did not agreed on MFS	49

Table 2. The agreement in manual estimation of the most frequent senses for ambiguous words described in RuThes.

5 Supervised Estimation of Most Frequent Sense

The described in the previous section expert annotation of the most frequent senses was performed only for ambiguous words with three or more senses described in RuThes. Besides, RuThes contains about 6,500 words with two senses, which were not analyzed manually. In addition, MFS can vary in different domains; natural language processing of documents in a

specific domain can require re-estimation of MFS on the domain collection.

Therefore we propose a method for supervised estimation of MFS based on several features calculated on the basis of a target text collection. To our knowledge, this is the first attempt to apply a supervised approach to MFS estimation. In addition, in contrast to previous works our method of MFS estimation is essentially based on unambiguous text entries of RuThes, especially on multiword expressions, which were carefully collected from many sources.

The automatic estimation of the most frequent sense was performed on a news collection of two million documents. Computing features for the supervised method we used several context types of a word: *the same sentence context, the neighbor sentence context, full document context*.

From the thesaurus, we utilize several types of conceptual contexts of an ambiguous word *w*:

- one-step context of word *w* attached to concept *C* ($ThesCon_{w1}$) that comprises other words and expressions attached to the same concept C and concepts directly related to C as described in the thesaurus;
- two and three-step contexts of word *w* attached to C ($ThesCon_{w2(3)}$) comprising words and expressions from the concepts located at the distance of maximum 2 (3) relations to the initial concept C (including C); the path between concepts can consist of relations of any types,
- one-step thesaurus context including only unambiguous words and expressions: $UniThesCon_{w1}$.

From these text and thesaurus contexts we generate the following features for ambiguous word *w* and its senses C_w:

- the overall collection frequency of expressions from $UniThesCon_{w1}$ – here we estimate how often unambiguous relatives of *w* were met in the collection – $Freqdoc_1$ and logarithm of this value *logFreqdoc*, Table 3 depicts frequencies of monosemous relatives of word *картина* in the source collection,
- the frequency of expressions from $UniThesCon_{w1}$ in texts where *w* was mentioned – $FreqdocW_1$,
- the overall frequency and the maximum frequency of words and expressions from $ThesCon_{wi}$ co-occurred with *w* in the same sentences – $FreqSentWmax_i$ and $FreqSentWsum_i$ (i=1, 2, 3),
- the overall frequency and the maximum frequency of words and expressions from $ThesCon_{wi}$ occurred in the neighbor sentences with *w* – $FreqNearWmax_i$ and $FreqNearWsum_i$ ((i=1, 2, 3).

All real-valued features are normalized by dividing them by their maximal value.

Monosemous relatives of word *картина*	sense of *картина*	document frequency
Фильм (*film*)	*moving picture*	45285
мультфильм (*cartoon*)	*moving picture*	4097
документальный фильм (*documentary film*)	*moving picture*	3516
живопись (painting)	*piece of painting*	3200
съемка фильма (shooting a film)	*moving picture*	2445
кинофильм (movie)	*moving picture*	1955
произведение искусства (art work)	*piece of painting*	1850
художественный фильм (fiction movie)	*moving picture*	1391
изобразительное искусство (visual art)	*piece of painting*	1102
режиссер картины (*director of the movie*)	*moving picture*	978
общая картина (general picture)	*general circums-tances*	932

Table 3. Document frequencies of monosemous relatives of word *картина* in the source collection of 2 mln. documents

We conduct our experiments on two sets of ambiguous words with three or more senses. The first set (*Set1*) consists of 330 words of 400 words that were analyzed by two linguists. They agreed with each other on one or two the most frequent senses. We used this set to train machine learning models. We apply the trained model to the second set of ambiguous words – 2532 words (*Set2*), for which only one expert provided MFS. Both sets include words of three parts of speech: nouns, verbs and adjectives.

The Table 4 presents accuracy results of MFS detection for single features. One can see that many single features provide a quite high level of accuracy.

Feature	Accuracy
$Freqdoc_1$	42.4%
$FreqdocW_1$	46.4%
$FreqSentWsum_1$	41.2%
$FreqSentWmax_1$	43.3%
$FreqSentWsum_2$	**48.2%**
$FreqSentWmax_2$	**48.2%**
$FreqSentWsum_3$	47.0%
$FreqSentWmax_3$	47.6%
$FreqNearWsum_1$	43.0%
$FreqNearWmax_1$	44.2%
$FreqNearWsum_2$	39.7%
$FreqNearWmax_2$	46.7%
$FreqNearWsum_3$	38.8%
$FreqNearWmax_3$	43.3%
Supervised algorithm	**50.6**%
Random	23.5%

Table 4. Accuracy of MFS prediction for single features and the supervised algorithm for Set1

To combine the features regression-oriented methods implemented in WEKA machine learning package were utilized. The best quality of classification using labelled data was shown by the ensemble of three classifiers: Logistic Regression, LogitBoost and Random Forest. Every classifier ranged word senses according to probability of this sense to be the most frequent one. We averaged probabilities of MFS generated by these methods. We obtained 50.6% accuracy of MFS prediction, the random baseline for this set is very low – 23.5% (Table 4). Our estimation is based on ten-fold cross validation.

To check the robustness of the obtained supervised model we applied it to the Set2. Table 5 describes the accuracy results for the best single features and the supervised method. The average level of results is higher than on the Set1, because Set2 contains the larger share of 3-sense words.

Feature	Accuracy
$FreqSentWsum_1$	53.7%
$FreqSentWsum_2$	**57.4%**
$FreqSentWmax_2$	53.7%
$FreqSentWsum_3$	54.6%
$FreqNearWsum_2$	53.7%
Supervised algorithm trained on Set1	**57.8%**
Random	33.4%

Table 5. Accuracy of MFS prediction for words from Set2 including accuracy of the best single features and accuracy of the supervised algorithm trained on Set1

We can see that simple context features give the accuracy results comparable with those described in (McCarthy et al., 2004; McCarthy et al., 2007), which have similar levels of random baselines (see Section 2). At this moment machine-learning combination of features did not demonstrate the significant growth in accuracy but the machine-learning framework allows adding distributional features utilized in the above-mentioned works.

6 Conclusion

In this paper we describe a supervised approach to detecting the most frequent senses of ambiguous words on the basis of thesaurus of Russian language RuThes. The approach is based on monosemous relatives of ambiguous words, in particular multiword expressions, described in RuThes. To check the proposed approach two linguists manually estimated the most frequent senses for 3,000 ambiguous words described in RuThes with three or more senses.

Our approach demonstrates its quality, which is quite comparable to the state-of-art distributional approaches, but our approach is based on simpler context features.

We found that some simple features (such as frequency of 2-step monosemous relatives of a word in sentences with this word – $FreqSentWsum_2$) provide high level of prediction of the most frequent sense.

We believe that in combination with other distributional features of words proposed in previous works it is possible to achieve better results in future experiments on MFS prediction.

Acknowledgments

This work is partially supported by Russian Foundation for Humanities grant N15-04-12017.

References

Eneko Agirre, and Oier Lopez De Lacalle. 2004. Publicly Available Topic Signatures for all WordNet Nominal Senses. *Proceedings of LREC-2004*.

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski., Eds. 2007. *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval). Association for Computational Linguistics, Prague, Czech Republic.

Irina Azarowa. 2008. RussNet as a Computer Lexicon for Russian. *Proceedings of the Intelligent Information systems IIS-2008*: 341-350.

Valentina Balkova, Andrey Suhonogov, and Sergey Yablonsky. 2008. Some Issues in the Construction of a Russian WordNet Grid. *Proceedings of the Forth International WordNet Conference*, Szeged, Hungary: 44-55.

Pavel Braslavski, Dmitrii Ustalov, and Mikhail Mukhin. 2014. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. *Proceedings of EACL-2014*, Sweden.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Elena Grishina, and Ekaterina Rakhilina. 2005. Russian National Corpus (RNC): an overview and perspectives. *AATSEEL- 2005.*

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. *Proceedings EMNLP-2005*, Vancouver, B.C., Canada: 419-426.

Shari Landes, Claudia Leacock, and Randee Tengi. 1998. Building semantic concordances. *In Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database*. Cambridge (Mass.): The MIT Press.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning Word Sense distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of ACL-2014*, pages 259-270.

Claudia Leacock, George Miller, and Martin Chodorow. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1): 147-165.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,* V(2): 768-774.

Natalia Loukachevitch and Boris Dobrov. 2014. RuThes Linguistic Ontology vs. Russian Wordnets. *In Proceedings of Global WordNet Conference GWC-2014.*

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of ACL-2004*.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4): 553-590.

Rada Mihalcea. 2002. Bootstrapping large sense tagged corpora. In *Proceedings of LREC-2002.*

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. *Proceedings of ACL-2014.*

Saif Mohammad and Graeme Hirst. 2006. Determining word sense dominance using a thesaurus. *Proceedings of EACL-2006*: 121-128.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2), 10.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Mihalcea, R. and Chklowski, T., editors, Proceedings of SENSEVAL-3: Third International Workshop on Evaluating Word Sense Disambiguating Systems*: 41-43.

Tommaso Petrolito and Francis Bond. 2014. A Survey of WordNet Annotated Corpora. *In Proceedings Global WordNet Conference, GWC-2014*: 236-245.