Automated Word Sense Frequency Estimation for Russian Nouns

Anastasiya Lopukhina
Vinogradov Russian Language Institute,
Moscow, Russia

Konstantin Lopukhin
Scrapinghub,
Moscow, Russia

Grigory Nosyrev
Yandex,
Moscow, Russia

Abstract

According to Zipf's observation there is a strong correlation between word frequency and polysemy, and yet word sense frequency distribution is a neglected area in computational linguistics. Furthermore, the study of sense frequency has theoretical interest and practical applications for lexicography and word sense disambiguation. Though WordNet and SemCor contain some information about sense frequency in English, it is not enough for either practical or research purposes. For Russian, even this information is lacking. To fill this gap, we develop and test an automated system based on semantic vectors that deals with the problem of sense frequency for Russian nouns. The model is first trained unsupervised on large corpora and then supplied with contexts and collocations from the Active Dictionary of Modern Russian. Dictionary examples are used either for supervised post-training, or for automatic labeling of clusters that are learnt unsupervised. This allows us to reach a frequency estimation error of 11-15% on different corpora without any additional labeled data. Word sense frequency distributions for 440 nouns are available online.

1 Introduction

Ambiguity is one of the fundamental properties of the lexical system of language. Lexical ambiguity manifests itself in homonymy and polysemy that are usually hard to discern (for Russian, this was proved in (Kachurin 2014)); in this paper we do not separate them. G. K. Zipf (1945) showed that ambiguity and frequency are related: there is a strong correlation between the number of different senses of words and their frequencies, that is, words that are higher in frequency list have more senses than low-frequency words. Thesauri and dictionaries confirm that most common words have several senses (Kilgarriff and Palmer 2000; Apresjan et al. 2014). For example, the most frequent Russian words like *v* 'in' and *byt'* 'to be' (according to the frequency dictionary for Russian (Lyashevskaya and Sharoff 2009)) have 28 and 16 senses respectively (Apresjan et al. 2014). Although the information about word frequency for many languages is widely presented in frequency dictionaries, sense frequencies and their distributions is a neglected area in linguistics.

Several papers discuss the problem of the most frequent (or predominant) sense, its detection and application to automated sense-disambiguation tasks (WSD). In sense-disambiguation evaluations the the first sense is presented as an important baseline (Agirre et al. 2007), which is difficult to overcome for many WSD systems (Navigli 2009). Typically, the most frequent sense is estimated with respect to thesauri or SemCor (Miller et al. 1993) a 220,000 word corpus tagged with WordNet (Fellbaum 1998) senses. In (Mohammad and Hirst 2006) the Macquarie Thesaurus' category information served as a basis for the predominant sense identification. Loukachevitch and Chetviorkin (2015) used the Thesaurus of Russian language (RuThes-lite) to determine the most frequent sense of ambiguous nouns, verbs and adjectives with the help of monosemous multiword expressions that are related to those words. Their results are comparable to the state-of-the-art in this field the highest accuracy rate reaches 57.4% (Loukachevitch and Chetviorkin 2015). McCarty et al. (2007) proposed a technique for ranking word senses on the basis of comparison of a given word with distributionally similar words. The predominant sense of a word was discovered automatically from raw text. This method produces more accurate predominant sense information than SemCor, especially for nouns with low coverage in that corpus. The authors reported that the most frequent sense was identified correctly for 56.3% of SemCor nouns and 45.6% verbs using the proposed technique (McCarthy et al. 2007).

Although the most frequent sense detection

gives some information about the overall sense distribution within a word, this question is rarely put in focus. Still word sense distributions are usually skewed and it is important to study them. In (Jin et al. 2009) a method to estimate the entropy of sense distribution is presented. Lau et al (2014) proposed a topic modelling-based method for estimating word sense distributions, based on Hierarchical Dirichlet Processes and on word sense induction, probabilistically mapping automatically learned topics to senses in a sense inventory. The method developed in this paper can be applied to tasks of detecting word senses which are unattested in a corpus, and identifying novel senses which are found in a corpus but not captured in a word sense inventory (Lau et al. 2014). The research of sense distribution within a word may contribute to theoretical studies of polysemy, lexicography and language learning.

The theoretical model of word sense frequency distribution was proposed in (Kilgarriff 2004). This model implies that word senses are distributed like words, by Zipfian or power-law distribution (Zipf 1935); more common words have a more dominant first sense, and, especially for the most common words, highly uneven distributions are expected (Kilgarriff 2004). Kilgarriff checked the model with data available from English SemCor and found a positive correlation between word frequency and first sense dominance and that the first sense is more dominant for words with smaller number of senses. It is interesting to validate the proposed model cross-linguistically and study the entropy of sense distributions for Russian nouns.

The lack of word sense frequency information is a problem in language learning and teaching, as dictionaries and learning resources do not provide this type of information. For example, Beck at al. (2013) state that any word that has different meanings appears only once in the list of words to be learned and whether the word *bank* means 'financial institution', 'edge of a river', or 'angle of an airplane' is not taken into account; its associated frequency represents all the different meanings. There is no way to get the frequency of the word *bank* meaning a 'financial institution'. The same problem exists for Russian. For example, the Russian word *vekha* can be described as having two distinctly different senses: 'boundary-mark' and 'a milestone in smb's life', according to the Active Dictionary of Modern Russian (Apresjan et al. 2014). Native speakers would probably agree that the first sense is quite special and rare as compared to the second. So the information about word sense frequency could help students learn the most relevant sense of the word first.

In this paper we present a method for determining noun sense frequency distributions automatically from raw text, the evaluation of this method, its comparison to state-of-the-art systems and a discussion on its applications. The technique we propose is based on semantic context vectors and uses contexts and collocations from the Active Dictionary of Modern Russian (Apresjan et al. 2014). We conduct our research for Russian nouns and this is the first study on sense frequency distributions for Russian.

The article is organized as follows. In Section 2 we give an overview of the three constituents of the word sense frequency estimation task. Section 3 describes our WSD method and compares two ways of clustering context representations. Section 4 is devoted to the evaluation of our method's performance in different corpora and experimental setups. In Section 5 we discuss results we obtained and their applications. We finish with a conclusion in Section 6.

2 Word sense frequency estimation task

Word sense frequencies and their distributions are not easy to estimate. We approach the task of estimating sense frequencies by performing word sense disambiguation on contexts sampled from corpora, and then calculating relative sense frequencies in the sample. For the proper estimation we need a word sense inventory, a source of word contexts (a corpus), and a sense disambiguation technique.

2.1 Word sense inventory

One of the most popular resources used for natural language processing for English is WordNet thesaurus (Fellbaum 1998) with its broad coverage and easy accessibility. Several WordNet-like projects were launched for Russian (Azarowa 2008; Braslavski et al. 2013), but at present there is no large enough and qualitative Russian wordnet (Loukachevitch and Chetviorkin 2015, 22). However WordNet-like resources were criticized for their fine-grained sense distinction that is not really needed for NLP tasks (Navigli 2006; Snow

et al. 2007) and does not reflect the way people represent word meaning (Ide and Wilks 2007; Brown 2008). Thus for our research we choose a reliable resource with a strong theoretical basis in sense distinction that reflects contemporary language Active Dictionary of Modern Russian (AD), an ongoing project of group of researchers from the Russian Language Institute (Apresjan et al. 2014). Word senses in the AD are considered distinct if they have different semantic and syntactic properties, collocational restrictions, synonyms and antonyms. Our current research is focused on nouns because they normally have more distinct senses (compared to other parts of speech), as most of them refer to objects existing in the real world (Iomdin et al. 2014).

2.2 Corpus

The corpus is a source of contexts, and its choice may influence sense frequency, because word sense distributions and predominant senses too vary from corpus to corpus. Many NLP studies were conducted on the domain neutral British National Corpus or on domain specific corpora, for example sports and finance texts (McCarthy et al. 2004; Lau et al. 2014) and the Yandex news database (Loukachevitch and Chetviorkin 2015). For the purposes of the current study we use the contexts from two domain neutral corpora: Russian National Corpus (RNC, ruscorpora.ru, 230 million tokens in the main corpus), a resource made by a consortium of linguists and developers that is the best academic corpus for Russian, and RuTenTen11 web-based corpus, the largest Russian corpus consisting of 18 billion tokens integrated into the Sketch Engine system (Kilgarriff et al. 2004). We sample 1000 random contexts for each word in both corpora and estimate sense frequency on these data. Sample sizes yield a statistical error below 3.1%. Web corpora are known for having more recent data and providing relevant and comparable linguistic evidence for language modeling (Kilgarriff and Grefenstette 2003; Piperski et al. 2013) and lexicographic purposes (Ferraresi et al. 2010). Using neural language models Kutuzov and Kuzmenko (2015) compared RNC and web-based corpora and found that the two corpora agree with each other in most cases. We believe that word sense frequency distributions in two different corpora might be another metric for the comparison of these types of corpora.

2.3 Word sense disambiguation technique

Automated sense frequency estimation requires performing automated word sense disambiguation (WSD) on a sample from a corpus, preferably using only existing knowledge sources. The problem of WSD has a long story in computational linguistics. Supervised methods for WSD were extensively studied, especially during SemEval evaluation series, and reach accuracy of 85-90% given hundreds of labeled examples (Navigli 2009; Pradhan et al 2007), but obtaining enough labeled examples for a large number of words is very processor-intensive. There are also effective semi-supervised WSD methods that use dictionaries, computer thesauri (WordNet, FrameNet) and formal ontologies as sources of information about words meanings (Agirre and Edmonds 2007; Navigli 2009). But the most promising are fully unsupervised, or sense induction (WSI) methods that solve the knowledge acquisition bottleneck by discovering senses from unlabeled corpora. Such sense discovery can be performed either by building vector representations of contexts and applying conventional clustering methods (Schutze 1998), or by learning multiple vector embeddings for each word (Huang 2012; Neelakantan 2014; Bartunov et al. 2015). We base our approach on distributed vector representations of words that are learnt from large corpora unsupervised, building vector representation of contexts from them. Our method uses collocations and examples from a dictionary, and does not require any additional labeled data.

3 Method

As we want to apply the method to a large number of words, we can only use existing linguistic resources, such as dictionaries with a limited number of examples for each sense, and raw corpora. Based on this constraint, we build context representation unsupervised, using distributed vector representations with weighting. Such representation already capture the main semantic properties of contexts in a very compact way, and do not require annotated data. After that, we present two different approaches: the first is clustering context representations and then building a mapping between these clusters and dictionary senses (cluster-map approach), the second is performing supervised classification using dictionary examples (sense-vec approach). In this section we de-

scribe the method in more detail.[1]

3.1 Context representation

Distributed vector representations is a way of representing words as low-dimensional dense real-valued vectors. A particularly efficient way of obtaining such vectors was proposed in (Mikolov et al. 2013a), and is known as word2vec family of methods. They are trained on word contexts from large corpora, and are based on a distributional hypothesis: words that occur in similar contexts tend to have similar meanings (Harris 1954). It was shown in (Mikolov et al. 2013b; c) that word2vec vectors capture semantic and syntactic relations between words. Since then they have been used as features in various NLP tasks such as document classification, machine translation, etc. We built a skip-gram word2vec model with negative sampling, window size 5 and vector dimension 1024, using a corpus consisting of about 2 billion words (combining RuWac, lib.ru and Russian Wikipedia) with lemmatization, which is important for Russian because of its rich morphology.

In order to build a context representation, we need to combine word vectors into a single context vector. The most straightforward way is to take the sum of word vectors:

$$\vec{c} = \sum_{i=1}^{n} \vec{w}_i = (\vec{w}_1 + \ldots + \vec{w}_n) \qquad (1)$$

It is possible to interpret this way of building context representation by analyzing similarity between two contexts. A common measure of semantic similarity is cosine similarity (cosine of the angle between two vectors), which is just a dot product if vectors are normalized. It is then easy to interpret the similarity of two contexts (omitting normalization constant):

$$sim(c^1, c^2) \sim c^1 \cdot c^2 = \\ = \sum_{i}^{n} \vec{w}_i^1 \cdot \sum_{j}^{n} \vec{w}_j^2 = \sum_{i,j}^{n,m} (\vec{w}_i^1 \cdot \vec{w}_j^2) \qquad (2)$$

This means that the similarity between contexts is a normalized sum of similarities between all pairs of words from these contexts. In most cases only several summands will be significant, as similarity between random words is usually close to zero.

[1] Source code and the sense frequency database are available on http://sensefreq.ruslang.ru

This way of context representation gives equal weight to all words in the context, which can be problematic if some words are just randomly shared between two contexts. For the task of word sense disambiguation, there are words that are much more indicative of the current sense, and we often can determine the sense just by seeing a single such word in the context. For example, for a polysemous word *gorshok* ('clay pot' / 'flowerpot' / 'potty') such indicative words include *rasteniye* ('a plant'; for 'flowerpot') and *priuchat'* ('to train'; for 'potty'). We would like to give more weight to such words when building a context representation. Such words are more likely to be seen in the context of a disambiguated word than on their own, so we can take the logarithm of the ratio of the relative frequency in context to the relative frequency without context, and use this weight for the context representation:

$$q_i = \ln \frac{P(w_i|c)}{P(w_i)}, \quad \vec{c} = \sum_{i=1}^{n} q_i \vec{w}_i \qquad (3)$$

A good context representation technique must produce similar vectors for contexts where the target word is used in the same sense, and dissimilar vectors for contexts with different senses. We can test this property using t-SNE embedding of annotated contexts for a single word. t-SNE is a method for projecting high-dimensional vectors into a two-dimensional space where vectors that are close in high-dimensional space are still close in a two-dimensional space (Fig. 1). Each point is a context colored according to its sense (human annotated). A better separation of senses means that we will need fewer annotated examples for automatic disambiguation.

3.2 Disambiguation method

Having built context representations, we now need a method for determining word sense by these representations. We evaluated two different approaches: cluster-map and sense-vec. In cluster-map we cluster a large number of contexts, and then build a mapping between these clusters and dictionary senses. This approach is appealing because it is much less reliant on annotated data: each cluster corresponds to a sense discovered by the clustering algorithm, and only mapping between these senses and dictionary senses requires annotated data. This approach could also be used to discover senses unattested in the dictionary.

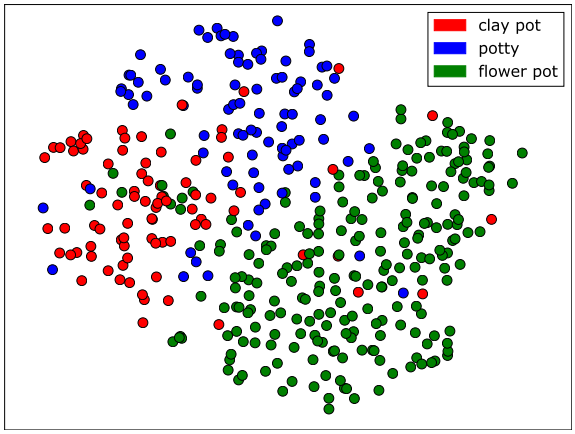

Figure 1: t-SNE embedding of context vectors

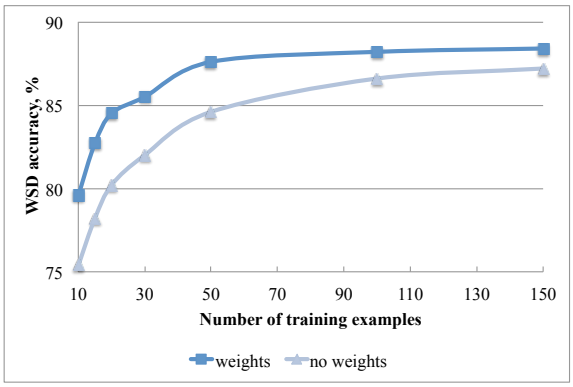

Figure 2: Effect of weighting: disambiguation accuracy with and without weighting as a function of the number of training examples

However, cluster-map is complicated, because a good clustering technique is required. In our experiments, spherical k-means clustering (a variation of k-means clustering that uses cosine similarity as a distance measure) gave best results.

In sense-vec we build a sense vector by averaging context vectors of available annotated examples for each sense (in this study examples were taken from the dictionary). Each context is then assigned to a sense vector closest in context vector space. More sophisticated training methods are often used for WSD, but they require more labeled examples to avoid overfitting. In our case, the number of examples is often just 5-10 for each sense, so a method with a lower number of parameters was chosen. We present the comparison of sense-vec and cluster-map in the next section.

For building a mapping between clusters and dictionary senses for cluster-map, and for building dictionary sense vectors for sense-vec, we used collocations and examples from Active Dictionary. More precisely, for each sense we extracted all examples (short and common usages), illustrations (longer, full-sentence examples from the Russian National Corpus), collocations, synonyms and analogues. Each example, illustration, etc. was treated as a separate context of a word used in a particular sense. The obtained vectors contributed to the average sense vector, which was used in a supervised classification in sense-vec. In a cluster-map approach we used these sense vectors to build a mapping between clusters and senses by assigning a cluster to the closest sense vector.

4 Evaluation

In this section we evaluate disambiguation accuracy and frequency estimation accuracy of sense-vec and cluster-map approaches, and study the effect of weighting when building context representations. Most experiments in this section use the same set of 20 words. All words have at least 200 hand-labeled contexts, with 100 contexts sampled from RuTenTen11 and 100 from the RNC. Seven of these words have additional 400 hand-labeled contexts from RuTenTen11, so the total number of hand-labeled contexts for all words is 6800.

In order to evaluate the effect of weighting on the quality of context representations, we compared performance of sense-vec with and without weighting. The sense-vec approach is convenient here because it is simpler than the cluster-map one, and allows us to measure how accuracy changes with the number of training examples. In this experiment we use 7 words that have 500 labeled contexts[2] from RuTenTen11. We vary the number of training examples from 10 to 150, leaving the rest for testing, and taking the average accuracy over several random train/test splits. The results are presented in Fig. 2.

We see that adding weighting reduces the number of training examples required to reach similar asymptotic accuracy of 87-88%. The most frequent sense (MFS) baseline on these words is 64% (MFS is a baseline method commonly used in WSD evaluation: it always outputs the most frequent sense of the word, regardless of context). The setup of this experiment is similar to

[2]Only 7 words had enough labeled contexts required for this experiment

SemEval-2007 Task 17 (lexical sample), where the best systems achieved a disambiguation accuracy of 89% on English nouns with an MFS baseline of 70% and hundreds of labeled contexts.

Comparison of sense-vec and cluster-map approaches is presented in Table 1. In this experiment we use contexts and collocations from the Active Dictionary (AD) for training. Disambiguation accuracy is evaluated on contexts sampled from RuTenTen11 and the RNC for 20 words (average accuracy is shown). We see that sense-vec performance is much better than cluster-map on both corpora.

Method	RuTenTen11	RNC
MFS baseline	66.8	69.9
Sense-vec	**76.7**	**74.8**
Cluster-map	72.7	71.0

Table 1: Comparison of sense-vec and cluster-map methods: disambiguation accuracy averaged on 20 nouns, higher is better.

In Table 2 we present the comparison of two different kinds of training data for sense-vec: AD examples (like in Table 1), and RuTenTen11 contexts (like in Fig. 2). Disambiguation accuracy is evaluated on the same 7 nouns used in Fig 2, since only these words have enough labeled contexts to use both for training and evaluation. We see that accuracy on AD is lower, but still it is 15% higher than the baseline. Two factors could explain better accuracy when using labeled RuTenTen11 contexts for training: greater number of long contexts (AD usually has only a couple of full sentences for each sense, the rest are collocations and short examples), and the fact that contexts used for training and testing come from the same corpus (RuTenTen11). Besides, using AD allows us to obtain comparable WSD accuracy without any additional human-annotated data.

Sense-vec training data	WSD Acc., %
AD (Active Dictionary)	79.2
20 contexts (RuTenTen11)	84.5
100 contexts (RuTenTen11)	**88.2**

Table 2: Comparison of sense-vec and cluster-map methods: disambiguation accuracy averaged on 20 nouns, higher is better.

It is valuable to understand not only the average disambiguation accuracy, but also how it varies from word to word. We have observed that some words and senses are harder to disambiguate both for the automatic method and, usually, for human annotators. One such word is *veshalka* ('coat-hanger' / 'rack' / other senses), where 'coat-hanger' and 'rack' are hard to distinguish because the context words are very similar for both senses, although they denote clearly distinct objects. Another example is *block* ('block') with 9 dictionary senses with abstract meanings. The distribution of disambiguation accuracy on 20 nouns is presented as a histogram in Fig 3.

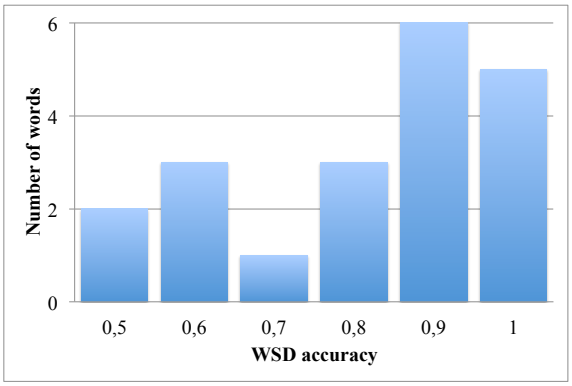

Figure 3: Histogram of disambiguation accuracy on 20 nouns

We see in Fig. 3 that accuracy can vary significantly for different words. As we would like to apply our method to a large list of nouns without any annotated data, it would be useful to have a way of estimating accuracy automatically. It turns out that it is indeed possible to estimate disambiguation accuracy for the sense-vec approach. For any given context (represented as a vector) we assign it to the closest sense vector. The closer a context vector is to a sense vector, the more accurate is the sense prediction. So a good estimate of accuracy is the ratio of predictions where the closeness between the sense and the context is within some threshold. Using this technique, the Pearson correlation between estimated and actual disambiguation accuracy for 20 nouns is 0.80, which has p-value 2.4E-05.

Our ultimate target is the estimation of word sense frequency, so it is important to study the frequency estimation error. We use two different measures. The first is maximum frequency error: the maximum absolute difference between predicted and actual sense frequency for every sense of an ambiguous word. An advantage of this met-

ric is that it is easily interpretable. The other metric is Jensen-Shannon divergence, which is a measure of how different two distributions are. Results are presented in Table 3. We see that again sense-vec is able to achieve higher accuracy, and that the frequency estimation accuracy is higher than the disambiguation accuracy.

Method	RuTenTen11	RNC
Frequency error, %		
Sense-vec	**11.4**	**15.4**
Cluster-map	17.4	20.3
JS divergence		
Sense-vec	**0.027**	**0.043**
Cluster-map	0.050	0.070

Table 3: Comparison of sense-vec and cluster-map methods on AD: frequency estimation error and JS divergence, averaged over 20 nouns, lower is better.

The main results of our evaluation are:

- The sense-vec method is able to achieve a disambiguation accuracy of 85-88% trained on 20-100 contexts respectively, which is comparable to SemEval-2007 results.
- Trained on AD, sense-vec achieves a frequency estimation error of 11-15% on different corpora
- Accuracy varies significantly for different words, and we are able to estimate it without labeled data.

5 Discussion

We applied our method to all homonymous and polysemous nouns from the first issue of the Active Dictionary of Modern Russian and obtained word sense frequencies for 440 Russian nouns. These data may have several applications: to lexicography and language learning, to the theoretical study of polysemy, to different NLP-tasks. Some of the applications are discussed below.

5.1 Corpora comparison and statistics

We compared sense frequencies for 440 nouns on contexts sampled from the academic Russian National Corpus (RNC) and web-based RuTenTen11 and found out that 82% of nouns have the same most frequent sense in both corpora. The difference can be explained by the content of the corpora.

For example, for the word *batareya* with a more special and archaic sense 'several large guns used together' is the most frequent in RNC, while a more neutral and intuitively frequent sense 'an electric battery' is the most frequent in RuTenTen11. The web-based corpus reflects domain specific most frequent senses of the words *vyderzhka* ('exposure' vs. 'stamina/self-control' first in RNC) and *bas* ('low-pitched musical sounds' vs. 'bass' first in RNC); and very colloquial most frequent senses of the words *blin* ('Damn!' vs. 'kind of pancake' first in RNC) and *bred* ('gibberish' vs. 'delirium' first in RNC). The discrepancy between the corpora also appears in the domain of political and social phenomena: for example, the word *bulleten'* in RNC has the most frequent sense 'bulletin', while in RuTenTen11 'voting paper'; for the word *gramota* the most frequent sense in RNC is 'charter' and in RuTenTen11 'diploma'. In general, our preliminary observations are similar to what was found in (Kutuzov and Kuzmenko 2015).

Both corpora show very similar overall statistics for sense frequency distribution: the average frequency of the first sense is 63% (which is lower than 70% for 30 English nouns used in Semeval-2007 lexical sample task 17), and 25% for the second sense. The ratio of words with one dominant sense (with frequency more than 80%) is only 21%. The average number of senses with frequency above 10% is 2.0, whereas the average total number of senses is 3.3.

5.2 Theoretical observations

Sense frequency data for a large number of Russian nouns are an interesting dataset for testing theories of sense frequency distributions. The model and the experimental evaluation proposed in (Kilgarriff 2004) suggest a positive correlation between word frequency and first sense dominance. Closely following Kilgariff's experiment setup, we checked his assumptions on our data and found a very similar tendency.

We think that obtaining sense frequency data for a larger number of words might help to determine the correlation between sense distributions and frequency. A separate study of sense distributions within words of one semantic type (e.g. 'plants' or 'nations') could also give interesting, more accurate and well interpretable results.

5.3 Lexicographic applications

The ordering of senses in dictionaries in Russian lexicographic tradition generally follows etymological principles the first sense of a polysemous word usually is the original, non-figurative meaning which does not always correspond to the most common sense in contemporary language. Thus accurate word sense frequency data can be useful for lexicographers and may help reconsider sense ordering and the depth of description in dictionaries. We compared the first sense in the Active Dictionary (AD) with the most frequent sense obtained by our method on two corpora (RNC and RuTenTen11). The ratio of polysemous words where the first dictionary sense is the most frequent is 71% in RNC and 68% in RuTenTen11. So in more than two thirds of cases etymologically the first sense coincides with the most frequent sense in texts.

The following examples show the difference between the first sense in the AD and the most frequent sense in the RNC. The word *garderob* in AD is described as having three senses: 'a large piece of furniture where you can hang your clothes' accounts for 23% of all occurrences of the word in RNC; 'a room with coat-hangers in a public place where visitors can leave their clothes' 31%; and 'the collection of clothes that someone has' 46%. The third sense that represents metonymical shift is now the most frequent in texts.

The word *volneniye* shows the same tendency. The first sense of volneniye in AD is 'heavy sea', cognate to the word *volna* 'wave'. This sense is presented in 8% of examples in RNC, while the third sense 'agitation or deep emotion' that represents metaphoric shift reaches 83% in RNC. The information about the most frequent sense of a polysemous word may be important for language learners and thus should be reflected in dictionaries.

6 Conclusions

In this paper we developed and tested an automated system for sense frequency estimation based on semantic context vectors, supplied with contexts and collocations from the Active Dictionary (Apresjan et al. 2014). The system achieved an average disambiguation accuracy of 75-77% and an average frequency estimation error of 11-15%, using only Active Dictionary contexts and collocations. As a result, we obtained sense frequencies for 440 ambiguous nouns from the first issue of the AD. The database is available on http://sensefreq.ruslang.ru and contains frequency distributions using contexts sampled from the academic Russian National Corpus and the web-based corpus RuTenTen11.

The system we developed is able to achieve a disambiguation accuracy of 85-88% for Russian nouns, trained on 20-100 hand-labeled contexts from RuTenTen11. This result is comparable to the state-of-the-art SemEval-2007 results for English nouns (top result of 89% on hundreds of labeled contexts).

Sense frequency distributions for a large list of nouns provide exciting opportunities for theoretical studies of polysemy and crosslinguistic comparisons. The results may enrich language learning resources and help lexicographers order senses of a word according to frequency if needed. The information on the predominant sense may be important for automated sense-disambiguation tasks on Russian.

The method presented in this paper can be applied to any language with a sufficiently large corpus and a dictionary that provides examples for each sense.

Acknowledgments

The authors thank Daria Shavarina for her careful reading of the draft.

References

Eneko Agirre. Llus marquez, and richard wicentowski, editors. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). Association for Computational Linguistics, Prague, Czech Republic*.

Eneko Agirre and Philip Glenny Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Jury D. Apresjan. 2014. *Active Dictionary of Modern Russian*, volume A-G.

Irina Azarowa. 2008. Russnet as a computer lexicon for russian. *Proceedings of the Intelligent Information systems IIS-2008*, pages 341–350.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2015. Breaking sticks and ambiguities with adaptive skip-gram. *arXiv preprint arXiv:1502.07257*.

Isabel L Beck, Margaret G McKeown, and Linda Kucan. 2013. *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Pavel Braslavski, Dmitrii Ustalov, and Mikhail Mukhin. A spinning wheel for yarn: user interface for a crowdsourced thesaurus. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden*, pages 101–104.

Susan Windisch Brown. Choosing sense distinctions for wsd: Psycholinguistic evidence. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 249–252. Association for Computational Linguistics.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database mit press.

Adriano Ferraresi, Silvia Bernardini, Giovanni Picci, and Marco Baroni. 2010. Web corpora for bilingual lexicography: a pilot study of english/french collocation extraction and translation. *Using Corpora in Contrastive and Translation Studies. Newcastle: Cambridge Scholars Publishing*.

Zellig S Harris. 1954. Distributional structure. *Word*.

Nancy Ide and Yorick Wilks, 2006. *Making sense about sense*, pages 47–73. Springer.

Boris Iomdin, Anastasiya Lopukhina, and Grigory Nosyrev. Towards a word sense frequency dictionary. In *Dialog 2014*, pages 199–212.

Peng Jin, Diana McCarthy, Rob Koeling, and John Carroll. Estimating and exploiting the entropy of sense distributions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 233–236. Association for Computational Linguistics.

Dmitry V. Kachurin. 2014. *The problem of distinguishing homonymy and polysemy for the purpose of compiling explanatory dictionaries*. Thesis.

Adam Kilgarriff. How dominant is the commonest sense of a word? In *Text, Speech and Dialogue*, pages 103–111. Springer.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

A Kilgarriff and J Rosenzweig. 2000. English framework and results computers and the humanities 34 (1-2). *Special Issue on SENSEVAL*.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105:116.

Andrey Kutuzov and Elizaveta Kuzmenko, 2015. *Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian*, pages 47–58. Springer.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of ACL*, pages 259–270.

Natalia Loukachevitch and Ilia Chetviorkin. Determining the most frequent senses using russian linguistic ontology ruthes. In *Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA*, page 22.

Olga N Lyashevskaya and Serge A Sharoff. 2009. Frequency dictionary of modern russian based on the russian national corpus [chastotnyy slovar'sovremennogo russkogo jazyka (na materiale nacional'nogo korpusa russkogo jazyka)]. *Azbukovnik, Moscow*.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics.

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

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Saif Mohammad and Graeme Hirst. Determining word sense dominance using a thesaurus. In *EACL*.

Roberto Navigli. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual*

meeting of the Association for Computational Linguistics, pages 105–112. Association for Computational Linguistics.

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

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Alexander Piperski, Vladimir Belikov, Nikolay Kopylov, Eugene Morozov, Vladimir Selegey, and Serge Sharoff. Big and diverse is beautiful: A large corpus of russian to study linguistic variation. In *Proc. Web as Corpus Workshop (WAC-8)*.

Sameer S Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 87–92. Association for Computational Linguistics.

Hinrich Schtze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y Ng. Learning to merge word senses. In *EMNLP-CoNLL*, volume 2007, pages 1005–1014.

George Kingsley Zipf. 1935. The psycho-biology of language.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.