Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

WORD SENSE INDUCTION FOR RUSSIAN: DEEP STUDY AND COMPARISON WITH DICTIONARIES[1]

Lopukhin K. A. (kostia.lopuhin@gmail.com), Scrapinghub

Iomdin B. L. (iomdin@ruslang.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, National Research University Higher School of Economics

Lopukhina A. A. (alopukhina@hse.ru), National Research University Higher School of Economics, V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences

The assumption that senses are mutually disjoint and have clear boundaries has been drawn into doubt by several linguists and psychologists. The problem of word sense granularity is widely discussed both in lexicographic and in NLP studies. We aim to study word senses in the wild—in raw corpora—by performing word sense induction (WSI). WSI is the task of automatically inducing the different senses of a given word in the form of an unsupervised learning task with senses represented as clusters of token instances. In this paper, we compared four WSI techniques: Adaptive Skip-gram (AdaGram), Latent Dirichlet Allocation (LDA), clustering of contexts and clustering of synonyms. We quantitatively and qualitatively evaluated them and performed a deep study of the AdaGram method comparing AdaGram clusters for 126 words (nouns, adjectives, and verbs) and their senses in published dictionaries. We found out that AdaGram is quite good at distinguishing homonyms and metaphoric meanings. It ignores disappearing and obsolete senses, but induces new and domain-specific senses which are sometimes absent in dictionaries. However it works better for nouns than for verbs, ignoring the structural differences (e.g. causative meanings or different government patterns). The Adagram database is available online: http://adagram.ll-cl.org/.

Key words: semantics, polysemy, text corpora, word sense induction, semantic vectors

[1] This research was supported by RSF (project No.16-18-02054: Semantic, statistic and psycholinguistic analysis of lexical polysemy as a component of Russian linguistic worldview). The authors would also like to thank students of the Higher School of Economics and Yandex School of Data Analysis for their help in annotating dictionary senses.

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ЗНАЧЕНИЙ СЛОВ ДЛЯ РУССКОГО ЯЗЫКА: ДЕТАЛЬНОЕ ИССЛЕДОВАНИЕ И СРАВНЕНИЕ СО СЛОВАРЯМИ

Лопухин К. А. (kostia.lopuhin@gmail.com), Scrapinghub

Иомдин Б. Л. (iomdin@ruslang.ru), Институт русского языка имени В. В. Виноградова РАН, Национальный исследовательский университет «Высшая школа экономики»

Лопухина А. А. (alopukhina@hse.ru), Национальный исследовательский университет «Высшая школа экономики», Институт русского языка имени В. В. Виноградова РАН

1. Introduction

Several linguists and psychologists have drawn into doubt the assumption that word senses are mutually disjoint and that there are clear boundaries between them. Many psycholinguistic studies have found evidence for processing differences between distinct meanings (homonyms) and related senses (polysemes) (Frazier and Rayner, 1990; Rodd et al., 2002; Beretta et al., 2005; Klepousniotou and Baum, 2007; MacGregor et al., 2015), which shows that related senses are not associated with processing penalties. Moreover, polysemy processing seems to depend on sense overlap—high-overlap metonymic senses are processed easier than moderate- and low-overlap, metaphoric senses (Klepousniotou 2002; Klepousniotou et al., 2008, 2012; MacGregor et al., 2015). Eye movement evidence of how people process polysemous words with metonymic senses suggests that instead of accessing a specific sense, language users initially activate a word's meaning that is semantically underspecified (Frisson 2009, 2015).

The problem of sense distinction has also been discussed by lexicographers. Some of them are skeptical about the view of word meanings as sets of discrete and mutually exclusive senses (Cruse 1995; Kilgarriff 1997; Hanks 2000). Kilgarriff (1997) claims that sense distinction is worthwhile only with respect to a task at hand, while Hanks (2000) calls into question the phenomenon of word senses, showing how different components of the meaning potential of the word are activated in different contexts. Furthermore, word senses descriptions in dictionaries depend on the consistency of lexicographers and their theoretical basis. Sense divisions may be influenced by personal preferences, as lexicographers traditionally distinguish 'lumpers' and 'splitters' among colleagues: those who tend to break up senses further and those who go for large, homonymic senses (Wilks, 1998: 276). One possible solution to the problem of sense distinction was proposed by Erk and colleagues (2013). They found that untrained annotators prefer to disambiguate words in a context in a non-binary manner. People are often inconsistent with disjoint sense partitions and are more comfortable with a graded scale. Thus, the authors proposed to describe word meanings in the

form of graded judgments in the disambiguation task (see also (McCarthy et al., 2016) about the notion of 'partitionability').

In the field of word sense disambiguation (WSD), the question of sense granularity is one of the key issues as the performance of WSD algorithms crucially depends on the sense inventory used for disambiguation. Although dictionaries and thesauri are the first option that comes to mind, they differ in the words that they cover, and also in the word senses that they distinguish. It was shown that sense distinction in most dictionaries is often too fine-grained for most NLP applications (see (Navigli, 2009) for a survey). This problem especially holds for the WordNet thesaurus (Fellbaum, 1998) and WordNet-like lexical databases—these resources are criticized for their excessive granularity that is not really needed for NLP tasks (Navigli, 2006; Snow et al., 2007) and for the loss of domain-specific senses (Pantel and Lin, 2002). Thus automated word sense induction (WSI) techniques may help establish an adequate level of granularity for NLP application and serve as empirically grounded suggestions for lexicographers.

Word sense induction is a task of automatically identifying the senses of words in raw corpora, without the need for handcrafted resources (dictionaries, thesauri) or manually annotated data. Generally WSI takes the form of an unsupervised learning task with senses represented as clusters of token instances (Navigli, 2009; Navigli, 2012; Nasiruddin, 2013). WSI results are often used as an input to WSD systems (Van de Cruys and Apidianaki, 2011; Navigli and Vannella, 2013) which allows to achieve state-of-the-art results in unsupervised WSD (Panchenko et al., 2016). Another NPL issue that benefits from word sense induction is web search clustering (Kutuzov, 2014). Di Marco and Navigli (2013) proposed a novel approach to web search result clustering based on word sense induction which outperformed both web clustering and search engines. In the fields of linguistics and lexicography WSI was successfully applied to the task of novel sense detection, i.e. identifying words which have taken on new senses over time (Lau et al., 2012, 2014). WSI also provides data for the study of diachronic variation in word senses (Bamman and Crane, 2011).

In this paper, we present an extensive comparison of four word sense induction techniques of different types. We chose two Bayesian approaches—a Latent Dirichlet allocation topic model (LDA) and a vector based Adaptive Skip-gram (AdaGram) model, one feature-based approach that represents each instance as a context vector, then utilizes a clustering algorithm, and an approach that performs clustering of word2vec neighbours. We quantitatively and qualitatively evaluated these techniques and performed a deep study of the AdaGram method comparing AdaGram clusters for 126 words (nouns, adjectives, and verbs) and their senses in published dictionaries. We studied sense overlap and types of senses that can be distinguished distributionally and by means of lexicographic theories. The research was done for Russian, and this is the first extensive study of the WSI methods for the Russian language.

2. Methods

A substantial number of different approaches to WSI has been proposed so far. They can be subdivided into local algorithms that discover senses separately for each word and global algorithms that allow to determine senses by comparing them to the

senses of other words (Navigli, 2009; Van de Cruys and Apidianaki, 2011; Nasiruddin, 2013). In this study we compare four algorithms, two local and two global: Latent Dirichlet allocation that uses topic modeling, context clustering, word2vec neighbours clustering, and AdaGram.

Latent Dirichlet allocation (LDA) posits that each context is a mixture of a small number of topics (senses) and that each word's occurrence is attributable to one of the context's senses. Traditionally, Latent Dirichlet allocation is used for topic modeling in documents: each word in the document is assumed to originate from some topic, and the document can be represented as a mixture of topics. In case of the word sense induction, LDA is applied to contexts of one word, where documents correspond to target word contexts, and topics correspond to target word senses. The number of topics for LDA must be fixed in advance, but there are non-parametric variations like hierarchical Dirichlet process (HDP) that allow variable numbers of topics per document. One drawback of LDA here is that the word contexts are much smaller (just 10–20 words) than the documents that LDA is usually applied to (at least 100–1000 words). Each sense is represented as words that have most weight in the topic. LDA was trained on contexts extracted from the ruWac corpus with 6 topics for each word. No sampling of contexts was performed, most words had at least 10 thousand contexts. The words in each context were additionally filtered: only those with a weight greater than 1.0 were left. This is the same weighting as was used in the context clustering method and is described below and in (Lopukhin and Lopukhina, 2016).

In the **word2vec neighbours** method, we took word vectors closest to the target word and clustered them using spherical k-means, and then merged close clusters. This method is based on two assumptions. The first one is that the word2vec vector of the polysemous word will capture the properties of all senses that are encountered in the corpus frequently enough. The second assumption is that each sense of the polysemous word has at least one monosemous word with a similar meaning that occurs in similar contexts and thus has a similar embedding. Both of these assumptions have their weaknesses. The first assumption does not hold for rare senses. If a word in one of its senses is used in a small number of contexts, the word vector will not capture its meaning. The second assumption causes even more trouble, as many senses will not have any reasonable synonyms that are used in similar contexts often enough. Still, this method is very efficient, easy to implement and produces reasonable results for many words. Senses are represented as words closest to the center of each cluster. The clustering method and sense merging are described in more detail in the description of the context clustering method below.

Context clustering represents contexts as dense vectors, taking a weighted average of word2vec vectors of individual words. Context vectors are clustered using spherical k-means, and then close clusters are merged. In more detail, in this study each context was represented as a weighted average of word2vec embeddings of 10 words before and after the target word. Weights were equal to the pointwise mutual information of contexts words (Lopukhin and Lopukhina, 2016) and allowed to give more weight to words that are more important for disambiguation. This method of context representation proved to be efficient for word sense disambiguation (Lopukhin and Lopukhina, 2016; Lopukhina et al., 2016). The spherical k-means method was used

for clustering of the context representations. Spherical k-means is similar to regular k-means clustering, but uses cosine distance instead of euclidian distance, which is a preferable measure of closeness for representations based on word2vec embeddings (Mikolov et al., 2013). The k-means clustering requires fixing the number of clusters in advance. But the number of senses is clearly different for different words, and k-means clusters often converge to very close points. To overcome both of these problems, clusters whose centers were closer than a certain threshold were merged. Senses are represented as most informative context words for a given sense.

AdaGram is a non-parametric Bayesian extension of the Skip-gram method. It automatically learns the required number of representations for all words at desired semantic resolution (Bartunov et al., 2015). It is able to learn the vector embedding for each sense of the word, where the number of senses is adapted depending on the number and diversity of contexts for each word. AdaGram has an efficient online learning algorithm that learns sense vectors for all words simultaneously. In practice, training is p times slower than for word2vec Skip-gram algorithm, where p is the maximum number of senses for a word (hyperparameter set in advance, typically 10–20). The model was evaluated on word sense induction tasks of SemEval-2007 and 2010 (Bartunov et al., 2015: 8–9) and achieved results superior to other extensions of word2vec to multiple senses. Besides p, the most important hyperparameter of AdaGram is α that controls granularity of produced senses. Other hyperparameters, such as vector dimension and window size, have the same meaning as in word2vec Skip-gram method. AdaGram can perform word sense disambiguation using induced senses and represents senses with nearest neighbors. We extended the sense representation with context words that give most information about a particular sense and typical sense contexts, and developed a Python library that allows loading AdaGram models and performing disambiguation. AdaGram model was built for about 190,000 most frequent words. Mean number of senses across all words is just 1.4, but more frequent words have more senses: 5.1 for the first 1,000 words and 3.6 for the first 10,000. The model is available online: http://adagram.ll-cl.org/about.

All models were trained on a 2 billion token corpus combining the ruWac Internet corpus (Sharoff, 2006), a Russian online library lib.ru and the Russian Wikipedia. All words were lowercased and lemmatized, no stop-word removal was performed. The word2vec Skip-gram model for word2vec neighbours and context clustering was trained with vector dimension 1024, window 5 and minimal token frequency 100 (forming a vocabulary of about 190,000 words). The AdaGram model was trained with maximum number of senses $p = 10$, sense granularity $\alpha = 0.1$, vector dimension 300, window 5 and minimal frequency 100. AdaGram has lower vector dimensionality, but this is compensated by the fact that multiple vectors are learnt for most words.

3. Evaluation

WSI evaluation is particularly arduous because there is no easy way of comparing and ranking different representations of senses. In fact, all the proposed measures in the literature tend to favour specific cluster shapes and consistency of the

senses produced as output. Here we apply two clustering measures—V-measure and adjusted Rand Index. Moreover, we qualitatively evaluated the obtained clusters and compared them with sense distinction in dictionaries.

3.1. Quantitative evaluation

For the quantitative evaluation of different WSI methods we compared induced senses with dictionary senses for 8 polysemous nouns and 10 polysemous verbs. For each word, 100–500 contexts were sampled from RuTenTen11 (Kilgarriff et al., 2004) and RNC corpora (http://ruscorpora.ru/en/) and labeled with dictionary senses from the Active Dictionary of Russian (Apresjan, 2014) by a human annotator. The methods assigned each context to one of the induced senses. Thus we obtained two different clusterings of contexts for each word: one by a human annotator and one by a WSI method, and used two different clustering similarity measures to compare them. We did not do a quantitative evaluation of the word2vec neighbours method as it lacks a natural disambiguation approach: senses are induced directly from word-2vec embeddings without using contexts; only a qualitative evaluation (below) was performed.

Table 1. V-measure for the word sense induction task

	Nouns	**Verbs**	**Average**
LDA	0.16	0.10	0.13
Context clustering	**0.39**	**0.22**	**0.31**
AdaGram	0.33	0.18	0.26

Table 2. Adjusted Rand Index for the word sense induction task

	Nouns	**Verbs**	**Average**
LDA	0.12	0.02	0.07
Context clustering	**0.34**	**0.14**	**0.24**
AdaGram	0.25	**0.13**	0.18

V-measure is a harmonic average of homogeneity and completeness of clusters. It was used in the SemEval-2010 Word Sense Induction & Disambiguation competition (Manandhar et al., 2010), but was criticized for favoring clusterings with a large number of clusters. This is less of a problem for our evaluation as we cap the maximum number of clusters for all methods at 10. Still, it is important to use an evaluation metric that corresponds to human intuition of having a reasonable number of clearly distinct senses, so we additionally used adjusted Rand Index (ARI) (Hubert and Arabie, 1985). It does not have the abovementioned issue and was used by Bartunov and colleagues (2015) in the AdaGram evaluation.

The quantitative comparison shows that context clustering and AdaGram are clearly better than LDA for nouns and especially for verbs. Context clustering performs better than AdaGram in this test for both ARI and V-measure, especially for nouns,

but this comparison is not entirely fair: hyperparameters of context representation for context clustering were specifically tuned during WSD evaluation in (Lopukhin and Lopukhina, 2016; Lopukhina et al., 2016) that were performed on a similar set of words, while AdaGram hyperparameters were left at their default values. Close senses were not merged for AdaGram, this could also improve V-measure and especially ARI.

3.2. Qualitative evaluation

We also performed a qualitative evaluation of these methods on 15 nouns: 7 polysemous, having 3–9 senses in the Active Dictionary of Russian, and 8 nouns that have just one sense in the dictionary, but at least 5 of them have new and slang meanings (e.g. *bomba* 'crib', 'sexually attractive woman' and *bajan* 'old joke'). All induced senses were divided into three groups by a human annotator. The first group represented quality senses: senses that have an intuitively clear meaning, even if they are more or less fine-grained than the dictionary senses, or are completely absent from the dictionary. The second group represented duplicate senses that did not have sufficient distinctions from other similar senses. The third group represented senses that were hard to interpret: either a mixture of several clearly distinct senses, or just uninterpretable sense descriptions. Therefore, an ideal WSI method would produce a large number of quality senses and minimal number of duplicate or hard to interpret senses. The average number of senses in each group for all studied methods is presented in Table 3.

Table 3. Average number of quality, duplicate and unclear senses for the four WSI method

	Quality senses	Duplicate senses	Hard to interpret
Word2vec neighbours	2.4	1.1	0.5
Context clustering	2.8	1.0	0.9
LDA	1.8	2.1	1.3
AdaGram	3.6	3.7	2.5

The two best methods according to this metric are AdaGram and context clustering. AdaGram produces the largest number of quality senses, while also having more duplicates and hard to interpret senses. Context clustering has fewer duplicates and hard to interpret senses, while still giving a high number of quality senses.

While AdaGram and context clustering use conceptually similar context representation (bag of word vectors), AdaGram has one computational advantage over the context clustering method: it learns sense vectors for all words simultaneously, while context clustering requires extracting contexts and clustering for each word separately. On one hand, this makes it much easier to change the algorithm and its hyperparameters, but on the other hand, AdaGram is able to produce sense vectors for all words in the corpus much faster. This is why we chose AdaGram for a deeper qualitative evaluation on more words.

3.3. AdaGram qualitative evaluation

First, we compared the average recall. We prepared a dataset of 51 nouns, 40 verbs and 35 adjectives with different ambiguity types—homonyms, words with metaphoric and metonymic senses, terms and frequent highly polysemous words (according to the Frequency Dictionary of Russian (Lyashevskaya and Sharoff, 2009)). For all these words we compared senses that are distinguished in four dictionaries (the Russian Language Dictionary (Evgenyeva, 1981–1984), the Explanatory Dictionary of Russian (Shvedova, 2007), the Large Explanatory Dictionary of Russian (Kuznetsov, 2014) and the Active Dictionary of Russian (Apresjan, 2014)) with clusters induced by AdaGram. A cluster was considered a hit if it represented only one dictionary sense: mixed or broader clusters were rejected. This part of the evaluation was performed by many annotators without overlap, so inter-annotator agreement is unknown. Overall, the average recall for nouns is higher than for adjectives and verbs and is lower in comparison with the Active Dictionary of Russian than with other dictionaries.

Table 4. Average number of senses discovered by AdaGram in comparison to dictionaries (recall)

	Apresjan, 2014	**Kuznetsov, 2014**	**Evgenyeva, 1981–1984**	**Shvedova, 2007**	**Average**
adjectives	0.44	0.72	0.68	0.66	0.62
nouns	0.50	0.70	0.72	0.74	0.69
verbs	0.35	0.61	0.68	0.71	0.61
Average	0.43	0.68	0.70	0.71	0.64

In order to compare the sets of senses induced by AdaGram and described by lexicographers, we performed a following experiment. 98 polysemic words (30 nouns, 38 adjectives, 30 verbs) were chosen from the Active Dictionary of Russian. Then we performed a manual evaluation of the AdaGram clusters (an example of the model's output is presented in the Appendix). The Active Dictionary of Russian was chosen because it uses a series of linguistic criteria to systematically distinguish between senses of a given word (called lexemes).

In many cases, AdaGram distinguishes less senses than the dictionary does. As it appears, AdaGram usually does not induce obsolete, obsolescent, vernacular, special etc. senses, e.g. *bort* 'a front lap' (of a jacket or coat: *bort pidžaka <sjurtuka>*), *balovat'* 'to horse around' (*Smotri ne baluj!*), *vstupit'* 'to come in' (*vstupit' na pomost* 'to mount a dais'), *žaba* as in *grudnaja žaba* 'cardiac angina'. For some words, most of the senses are quite rare and therefore ignored by AdaGram, e.g. all senses of the verb *axnut'* except for the first and direct one ('to gasp'): 'to go off' (*v nebe axnulo* 'boom went the sky), 'to hit smb' (*axnut' po skule*), 'to hit smth' (*axnut' kulakom po stolu* 'to thump a table'), 'to drop smth with a loud noise' (*axnut' printer ob pol* 'to flop the printer down'), 'to empty a glass' (*axnut' stakan vodki* 'to gulp down a glass of vodka'). This might be explained by the simple fact that these senses might not occur in the corpus at all, or occur very rarely.

More interestingly, AdaGram does not distinguish senses which differ in argument structure rather than in semantic components or domain, e.g. causative meanings: *gasit'* 'to extinguish' (*gasit' svet* 'to switch off the lights') and 'to be the cause of extinguishment' (*Dožd' gasit koster* 'The rain puts out the fire'); *brit'* 'to shave' (*On ne breet podborodok* 'He does not shave his chin') and 'to get shaved (by a barber)' (*On breet borodu v barberšope* 'He gets shaved in a barbershop'). Lexicalized grammatical forms of adjectives are not considered by AdaGram as specific senses, e.g. *bližajšij* ('the closest', a superlative form of *blizkij* 'close', but also 'near': *v blizhajšie dni* 'in the next few days') or *vysšij* ('the highest', a superlative form of *vysokij* 'high', but also 'higher': *vyssee obrazovanie*' 'higher education').

On the other hand, in some cases AdaGram offers more senses than the dictionary. First of all, these are proper names, e.g. *Blok* (a surname, literally 'a block, a pulley'), *Avangard* (a hockey team, literally 'advance guard'), *Vidnoe* (a town, literally 'smth visible'), *Groznyj* (the Russian tzar and the capital of Chechnya, literally 'menacing'), etc., which are normally excluded from explanatory dictionaries (at least in the Russian lexicographic tradition). More often, AdaGram distinguishes between groups of contexts referring to different domains. For example, it divides into two clusters the following sets of collocates of the word *babočka* 'a butterfly': (1) *motylek* 'a moth', *strekoza* 'a dragonfly', *porxat'* 'to flutter', *krylyško* 'a winglet', *roit'sja* 'to swarm', (2) *gusenica* 'a caterpillar', *kajnozojskij* 'Cainozoic', *nasekomoe* 'an insect', *češuekrylyj* 'lepidopterous', *dvukrulyj* 'dipterous'. Obviously, these are not two different senses of the word *babočka*, but rather two types of texts (fiction vs. non-fiction) where it apparently occurs in distinctly different contexts. Similarly, AdaGram postulates two meanings for the word *oružie* 'weapon' with contexts corresponding to wars vs. computer games, *brak* (civil vs. religious marriage), *anglijskij* 'English' (history books vs. sports). For the noun *graf,* apart from the mathematical sense ('a graph'), not listed in the Active Dictionary of Russian, AdaGram gives as many as four types of contexts corresponding to counts or earls in Russia, France, Britain and Western Europe in general, which the dictionary considers belonging to the same sense.

In many cases, AdaGram offers several clusters of contexts which do not overlap with the dictionary senses. For the adjective *vozdušnyj*, it offers two groups of contexts: (1) *vozdušnyj potok* 'air flow'*, vozdušnyj fil'tr* 'air filter'*, vozdušnyj nasos* 'air pump'*, vozdušnyj poršen'* 'air piston', (2) *vozdušnyj šarik* 'party balloon', *vozdušnyj poceluj* 'air kiss', *vozdušnaja figurka* 'a feathery figurine', *vozdušnoe plat'e* 'a vapory dress', which the dictionary subdivides into five different senses: 'consisting of air', 'happening in the air', 'using air', 'using the energy of the air', 'lightweight'. Party balloons, kisses and dresses are more likely to be referred to in fiction, while air filters, pumps and pistols are more charachteristic for technical prose, and clearly these genres have quite different classes of contexts.

Finally, there are some special senses found by AdaGram but not listed in the dictionary; apart from the aforementioned *graf* 'graph', consider *agent* 'chemical agent', *vint* 'hard disk' (apparently a shortening from *vinčester* < Winchester), *gorjačij* 'used for communication' (*gorjačaja linija, gorjačij telefon* 'hotline').

Our analysis shows that less distant senses are less likely to be distinguished by AdaGram, according to the following hierarchy: homonyms > senses belonging

to different subgroups > senses beloging to the same subgroup > exploitations of the same sense (all these entities are systematically distinguished and marked in the Active Dictionary or Russian). It is also worth noting that metaphors are much more recognizable by AdaGram than metonymic shifts, which might also correspond to the way they are treated by native speakers. Although AdaGram distinguishes less senses than the Active Dictionary of Russian, the feedback we received from our annotators shows that they are often more satisfied with smaller sets of senses found by the former than with the fine-grained distinctions provided by the latter.

4. Conclusion

In this paper we have explored the question of word sense induction for Russian. We applied four methods with different underlying approaches—a Latent Dirichlet allocation topic model (LDA) a vector based AdaGram model, a feature-based context clustering method and an approach that performs clustering of word2vec neighbours. Quantitative evaluation performed for nouns and verbs showed that context clustering and AdaGram are better than LDA for nouns and much better for verbs. The overall qualitative evaluation of the interpretability of the obtained clusters revealed that the two best methods are AdaGram and context clustering. They produce the largest number of quality senses while word2vec neighbours and LDA performance is less powerful. This result can be explained by the fact that LDA has access to less information in this setup: it works only on contexts of each word individually, while the other methods have access to the whole corpus, either directly for AdaGram or indirectly via word2vec embeddings for other methods. Context clustering works better than word2vec neighbours because it uses the contexts too, while word2vec neighbours requires existence of monosemous neighbours.

In a deeper study of AdaGram and its comparison with dictionaries, we found that the method performs consistently well for different parts of speech and induces overall 63% of senses that are distinguished by dictionaries. Moreover, AdaGram allows to get new and domain-specific senses that may not be included in domain-neutral lexical resources. The major limitation of the method is its inability to take syntactic information into account. The problem of sense discrimination by context is most evident for verbs. Although, AdaGram may not allow to solve the problem of the excessive granularity of lexical resources for NLP tasks (as it produces quite fine-grained clusters-senses), its clustering seems more corresponding to human intuition. And similarly to the conclusions from psycholinguistic experiments with ambiguous words, AdaGram distinguishes homonyms and metaphors better than more closely related senses.

The AdaGram database for Russian is available online (http://adagram.ll-cl.org/). The AdaGram method can be applied as a tool for lexicographers dealing with Russian—for sense induction from the large corpus and for novel sense detection. One of the AdaGram's possible application is the regular polysemy patterns detection which was discussed in (Lopukhina and Lopukhin, 2016). Besides, this instrument may help find patterns in big corpora that a human can not access.

One possible development of this study may be making context representations "aware of" the word order and the syntactic relations. This might allow distinguishing

senses that are currently lumped together. This goal can be achieved either by corpus preprocessing (e.g. applying a syntactic parser), or by using richer context representations (e.g. moving from the bag of words to recurrent neural networks). Another possible development may be adjusting the methods to produce more distinct senses, and improving the sense presentations to make them clearer for the users.

Appendix

An example of the AdaGram model's output for the word *goršok* (# 0 'flower pot', #2 'potty' and 'clay pot', #1 'clay pot' and 'potty', #4 'clay pot', #3 'given name').

горшок

Word ipm: 16.89, occurrences: 34188.

#0 0.33	#2 0.29	#1 0.26	#4 0.10	#3 0.03
Contexts: ...	Contexts: ...	Contexts: ...	Contexts: ...	Contexts: ...
Neighbours: *цветочный, цветок, растение, грунт, клумба*	**Neighbours:** *приучать, бог, отучать, ребенок, ходить*	**Neighbours:** *глиняный, ночной, щи, звон, котел*	**Neighbours:** *переработка, деформация, межкомнатный, театрализовать, эллинский*	**Neighbours:** *адмирал, флот, крейсер, полузащитник, вмф*
Similar senses:	**Similar senses:**	**Similar senses:**	**Similar senses:**	**Similar senses:**
вазон 0.77	*памперс* 0.52	*миска* 0.75	*сосуд* 0.60	*горшков* 0.67
кадка 0.66	*садик* 0.49	*кастрюля* 0.74	*пифос* 0.58	*головко* 0.64
ваза 0.66	*ребенок* 0.49	*котел* 0.71	*амфора* 0.55	*кузнецов* 0.61
кашпо 0.65	*ребеночек* 0.49	*глиняный* 0.70	*кувшин* 0.54	*касатонов* 0.60
клумба 0.63	*доча* 0.49	*плошка* 0.69	*лепной* 0.52	*макаров* 0.58

References

1. *Apresjan Ju. D.* (ed.). (2014). Active Dictionary of Russian. Vol. 1 (A–B), Vol. 2 (V–G) [Aktivnyj slovar' russkogo jazyka. Tom 1 (A-B), tom 2 (V–G)]. Jazyki slavjanskih kul'tur, Moscow.
2. *Bamman, David and Gregory Crane.* (2011). Measuring historical word sense variation. Proceedings of the 2011 Joint International Conference on Digital Libraries (JCDL 2011), pp. 1–10, Ottawa, Canada.
3. *Bartunov, Sergey, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov.* (2015). Breaking sticks and ambiguities with adaptive skip-gram. Accessed February 15, 2017. https://arxiv.org/abs/1502.07257.
4. *Beretta, Alan, Robert Fiorentino, and David Poeppel.* (2005). The effects of homonymy and polysemy on lexical access: An MEG study. In: Cognitive Brain Research 24.1: 57–65.
5. *Cruse, D. A.* (1995). Polysemy and related phenomena from a cognitive linguistic viewpoint. In Philip Saint-Dizier and Evelyne Viegas, editors, Computational Lexical Semantics. Cambridge University Press, pages 33–49.

6. *Di Marco, Antonio and Roberto Navigli.* (2013). Clustering and diversifying Web search results with graph-based word sense induction. In: Computational Linguistics, 39(3):709–754.
7. *Erk, Katrin, Diana McCarthy, and Nick Gaylord.* (2013). Measuring word meaning in context. In: Computational Linguistics, 39(3):511–554.
8. *Evgenyeva A. P.* (ed.). (1981–1984), Russian Language Dictionary. Russian language, Moscow.
9. *Fellbaum, Christiane, editor.* (1998). WordNet, An Electronic Lexical Database. The MIT Press, Cambridge, MA.
10. *Frazier, Lyn and Keith Rayner.* (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. In: Journal of Memory and Language, 29:181–200.
11. *Frisson, Steven.* (2009). Semantic underspecification in language processing. In: Language and Linguistic Compass, 3, 111–127.
12. *Frisson, Steven.* (2015). About bound and scary books: The processing of book polysemies. In: Lingua, 157, 17–35.
13. *Hanks, Patrick.* (2000). Do word meanings exist? Computers and the Humanities. In: Senseval Special Issue, 34(1–2):205–215.
14. *Hubert, L. and Arabie, P.* (1985). Comparing partitions. Journal of classification, 2(1):193–218.
15. *Kilgarriff, Adam.* (1997). 'I don't believe in word senses'. In: Computers and the Humanities, 31(2):91–113.
16. *Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell.* (2004). The Sketch Engine. In Euralex 2004. Proceedings, 105–116. Lorient, France.
17. *Klepousniotou, Ekaterini.* (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. In: Brain and Language, 81:205–223.
18. *Klepousniotou, Ekaterini, & Baum, S. R.* (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. In: Journal of Neurolinguistics, 20(1), 1–24.
19. *Klepousniotou, Ekaterini, Debra Titone, and Caroline Romero.* (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. In: Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6):1,534–1,543.
20. *Klepousniotou, Ekaterini, Pike, G. B., Steinhauer, K., & Gracco, V.* (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. In: Brain and Language, 123(1), 11–21.
21. *Kutuzov, Andrey.* (2014). Semantic clustering of Russian web search results: possibilities and problems. In: Russian Summer School in Information Retrieval. Springer International Publishing.
22. *Kuznetsov S. A.* (ed.). (1998), Large Explanatory Dictionary of Russian. Norint, St. Petersburg.
23. *Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin.* (2012). Word sense induction for novel sense detection. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 591–601. Association for Computational Linguistics.

24. *Lau, Jey Han, 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, 259–270. Baltimore, Maryland, USA.
25. *Lopukhin, Konstantin and Anastasiya Lopukhina.* (2016). Word Sense Disambiguation for Russian Verbs Using Semantic Vectors and Dictionary Entries. In: Computational Linguistics and Intellectual Technologies. No. 15. P. 393–405.
26. *Lopukhina, Anastasiya and Konstantin Lopukhin.* (2016). Regular polysemy: from sense vectors to sense patterns. In: The 26th International Conference on Computational Linguistics (COLING 2016). P. 19–23.
27. *Lopukhina, Anastasiya, Konstantin Lopukhin, Boris Iomdin, and Grigory Nosyrev.* (2016). The Taming of the Polysemy: Automated Word Sense Frequency Estimation for Lexicographic Purposes, in: Proceedings of the XVII EURALEX International Congress. Lexicography and Linguistic Diversity (6–10 September, 2016). Tbilisi: Ivane Javakhishvili Tbilisi State University.
28. *Lyashevskaya, Olga N., and Serge A. Sharoff.* (2009). Frequency dictionary of modern Russian based on the Russian National Corpus. Moscow.
29. *MacGregor, L. J., Bouwsema, J., & Klepousniotou, E.* (2015). Sustained meaning activation for polysemous but not homonymous words: Evidence from EEG. In: Neuropsychologia, 68, 126–138.
30. *Manandhar, S., Klapaftis, I. P., Dligach, D., and Pradhan, S. S.* (2010). SemEval-2010 task 14: Word sense induction & disambiguation. In: International workshop on semantic evaluation (SemEval), pp. 63–68.
31. *McCarthy, Diana, Marianna Apidianaki, and Katrin Erk.* (2016). Word sense clustering and clusterability. In: Computational Linguistics, Vol. 42, No. 2, Pages: 245–275.
32. *Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey and Dean.* (2013) Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26 (NIPS 2013), 3111–3119.
33. *Nasiruddin, Mohammad.* (2013). A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. arXiv preprint arXiv:1310.1425.
34. *Navigli, Roberto.* (2006). 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, 105–112. Association for Computational Linguistics, USA.
35. *Navigli, Roberto.* (2009). Word sense disambiguation: A survey. ACM Computing Surveys, 41(2):1–69.
36. *Navigli, Roberto.* (2012). A quick tour of word sense disambiguation, induction and related approaches. In: International Conference on Current Trends in Theory and Practice of Computer Science. Springer Berlin Heidelberg.
37. *Navigli, Roberto, and Daniele Vannella.* (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In: Second Joint Conference on Lexical and Computational Semantics (* SEM). Vol. 2.

38. *Panchenko A., Simon J., Riedl M., Biemann C.* (2016). Noun Sense Induction and Disambiguation using Graph-Based Distributional Semantics. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS). Bochum, Germany.
39. *Pantel, Patrick, and Dekang Lin.* (2002). Discovering word senses from text. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM.
40. *Rodd, Jennifer, Gareth Gaskell, and William Marslen-Wilson.* (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. In: Journal of Memory and Language 46.2: 245–266.
41. *Sharoff, S.* (2006). Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini, pp. 63–98.
42. *Shvedova N. Yu.* (2007), Explanatory Dictionary of Russian, Moscow.
43. *Snow, Rion, Sushant Prakash, Dan Jurafsky, and Andrew Y. Ng.* (2007). Learning to merge word senses. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 1005–1014. Prague, Czech Republic.
44. *Van de Cruys, Tim, and Marianna Apidianaki.* (2011). Latent semantic word sense induction and disambiguation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics.
45. *Wilks Y.* (1998). Senses and texts. In: Computational linguistics and Chinese language processing. V. 3. No. 2.