.

Corpus-based Check-up for Thesaurus

Natalia Loukachevitch
Research Computing Center
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
louk_nat@mail.ru

Abstract

In this paper we discuss the usefulness of applying a checking procedure to existing thesauri. The procedure is based on the analysis of discrepancies of corpus-based and thesaurus-based word similarities. We applied the procedure to more than 30 thousand words of the Russian wordnet and found some serious errors in word sense description, including inaccurate relationships and missing senses of ambiguous words.

1 Introduction

Large thesauri such as Princeton WordNet (Fellbaum, 1998) and wordnets created for other languages (Bond and Foster, 2013) are important instruments for natural language processing. Developing and maintaining such resources is a very expensive and time-consuming procedure. At the same time, contemporary computational systems, which can translate texts with almost human quality (Castilho et al., 2017), cannot automatically create such thesauri from scratch providing a structure somehow similar to resources created by professionals (Camacho-Collados, 2017; Camacho-Collados et al., 2018).

But if such a thesaurus exists, the developers should have approaches to maintain and improve it. In previous works, various methods on lexical enrichment of thesauri have been studied (Snow et al., 2006; Navigli and Ponzetto, 2012). But another issue was not practically discussed: how to find mistakes in existing thesaurus descriptions: incorrect relations or missed significant senses of ambiguous words, which were not included accidentally or appeared recently.

In fact, it is much more difficult to reveal missed and novel senses or wrong relations, if compared to detect novel words (Fremann and Lapata, 2016; Lau et al., 2014). So it is known that such missed senses are often found during semantic annotation of a corpus and this is an additional problem for such annotation (Snyder and Palmer, 2004; Bond and Wang, 2014).

In this paper, we consider an approach that uses embedding models to reveal problems in a thesaurus. Previously, distributional and embedding methods were evaluated in comparison with manual data (Baroni and Lenci, 2011; Panchenko et al., 2015). But we can use them in the opposite way: to utilize embedding-based similarities and try to detect some problems in a thesaurus.

We study such similarities for more than 30 thousand words presented in Russian wordnet RuWordNet (Loukachevitch et al., 2018)[1]. RuWordNet was created on the basis of another Russian thesaurus RuThes in 2016, which was developed as a tool for natural language processing during more than 20 years (Loukachevitch and Dobrov, 2002). Currently, the published version of RuWordNet includes 110 thousand Russian words and expressions.

2 Related Work

Word sense induction approaches (Agirre and Soroa, 2007; Navigli, 2009; Lau et al., 2014; Panchenko et al., 2018) try to induce senses of ambiguous words from their contexts in a large corpus. Sometimes such approaches can find new senses not described in any lexical resources. But the results of these methods are rarely intended to

[1] http://ruwordnet.ru/en/

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5773–5779
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

improve the sense representation in a specific semantic resource.

Lau et al. (2014) study the task of finding unattested senses in a dictionary is studied. At first, they apply the method of word sense induction based on LDA topic modeling. Each extracted sense is represented as top-N words in the constructed topics. To compute the similarity between a sense and a topic, the words in the definition are converted into the probability distribution. Then two probability distributions (gloss-based and topic-based) are compared using the Jensen-Shannon divergence. It was found that the proposed novelty measure could identify target lemmas with high- and medium-frequency novel senses. But the authors evaluated their method using word sense definitions in the Macmillan dictionary[2] and did not check the quality of relations presented in a thesaurus.

A series of works was devoted to studies of semantic changes in word senses (Gulordava and Baroni, 2011; Mitra et al., 2015; Frermann and Lapata, 2016), Gulordava and Baroni, 2011) study semantic change of words using Google n-gram corpus. They compared frequencies and distributional models based on word bigrams in 60s and 90s. They found that significant growth in frequency often reveals the appearance of a novel sense. Also it was found that sometimes the senses of words do not change but the context of their use changed significantly.

In (Mitra et al., 2015), the authors study the detection of word sense changes by analyzing digitized books archives. They constructed networks based on a distributional thesaurus over eight different time windows, clustered these networks and compared these clusters to identify the emergence of novel senses. The performance of the method has been evaluated manually as well as by comparison with WordNet and a list of slang words. But Mitra et al. (2015) did not check if WordNet misses some senses.

3 Comparison of Distributional and Thesaurus Similarities

To compare distributional and thesaurus similarities for Russian according to RuWordNet, we used a collection of 1 million news articles as a reference collection. The collection was lemmatized. For our study, we took thesaurus words with frequency more than 100 in the corpus. We obtained 32,596 words (nouns, adjectives, and verbs). For each of these words, all words located in the three-step relation paths (including synonyms, hyponyms, hypernyms, co-hyponyms, indirect hyponyms and hypernyms, cross-categorial synonyms, and some others) were considered as related words according to the thesaurus. For ambiguous words, all sense-related paths were considered and collected together. In such a way, for each word, we collected the thesaurus-based "bag" of similar words (TBag).

Then we calculated embeddings according to word2vec model with the context window of 3 words, planning to study paradigmatic relations (synonyms, hypernyms, hyponyms, co-hyponyms). Using this model, we extracted the twenty most similar words w_i to the initial word w_0. Each w_i should also be from the thesaurus. In such a way, we obtained the distributional (word2vec) "bag" of similar words for w_0 (DBag) with their calculated word2vec similarities to w_0.

Now we can calculate the intersection between TBag and DBag and sum up the word2vec similarities in the intersection. Figure 1 shows the distribution of words according to the similarity score of the TBag-DBag intersection. The axis X denotes the total similarity in the TBag-DBag intersection: it can achieve more than 17 for some words, denoting high correspondence between corpus-based and thesaurus-based similarities.

Relative adjectives corresponding to geographical names have the highest similarity values in the TBag-DBag intersection, for example, *samarskii* (related to Samara city), *vologodskii* (related to Vologda city), etc. Also nouns denoting cities, citizens, nationalities, nations have very high similarity values in the TBag-DBag intersection.

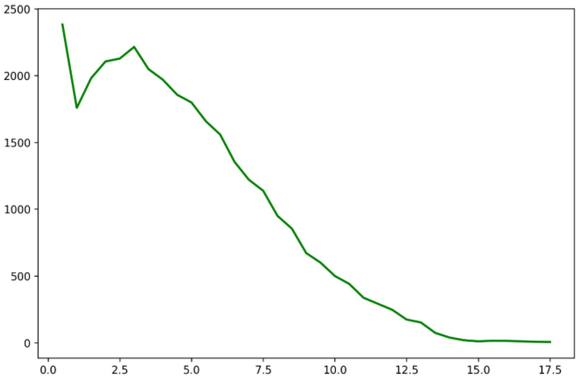

Figure 1. Distribution of thesaurus words according to the total similarity in the TBag-Dbag intersection

[2] https://www.macmillandictionary.com/

.

Among verbs, verbs of thinking, movement (*drive – fly*), informing (*say – inform – warn*), value changing (*decrease – increase*), belonging to large semantic fields, have the highest similarity values (more than 13).

At the same time, the rise of the curve in the low similarity values reveals the segment of problematic words.

4 Analyzing Discrepancies between Distributional and Thesaurus Similarities

We are interested in cases when the TBag-DBag intersection is absent or contains only 1 word with small word2vec similarity (less than the threshold (0.5)). We consider such a difference in the similarity bags as a problem, which should be explained. We obtained 2343 such problematic "words". Table 1 shows the distribution of these words according to the part of speech.

It can be seen that verbs have a very low share in this group of words. It can be explained that in Russian, most verbs have two aspect forms (Perfective and Imperfective) and also frequently have sense-related reflexive verbs. All these verb variants (perfective, imperfective, reflexive) are presented as different entries in RuWordNet. Therefore, in most cases altogether they should easily overcome the established threshold of discrepancies. In the same time, if some verbs are found in the list of problematic words, they have real problems of their description in the thesaurus.

Part of speech	**Number**
Nouns	1240
Adjectives	877
Verbs	226
Total	2343

Table 1. Distribution of parts of speech among problematic words

To classify the causes of discrepancies, we ordered the list of problematic words in decreasing similarity of their first most similar word from the thesaurus, that is in the beginning words with the most discrepancies are gathered (further, ProblemList). Table 2 shows the share of found problems in the first 100 words of this list.

In the subsections, we consider specific reasons, which can explain discrepancies between thesaurus and corpus-based similarities.

4.1 Morphological Ambiguity and Misprints

The most evident source of the discrepancies is morphological ambiguity when two different words w_1 and w_2 have the same wordform and words from DBag of w_1 in fact are semantically related to w_2 (usually w_2 has larger frequency). For example, in Russian there are two words *bank* (financial organization) and *banka* (a kind of container). All similar words from Dbag to *banka* are from the financial domain: *gosbank* (state bank), *sberbank* (saving bank), *bankir* (banker), etc. The analyzed list of problematic words includes about 90 such words. 32 of such words are located in the top of ProblemList.

The technical reasons of some discrepancies are frequent misprints. For example, frequent Russian word *zayavit* (to proclaim) is often erroneously written as *zavit* (to curl). Therefore the DBag of word *zavit* includes many words similar to *zayavit* such as *soobshchit'* (*to inform*), or *otmetit* (*to remark*). Another example is a pair words *statistka (showgirl*) and *statistika (statistics)*. In the top-100 of ProblemList, two such words were found. Such cases can be easily excluded from further analysis.

4.2 Named Entities and Multiword Expressions

The natural reason of discrepancies are named entities, whose names coincide with ordinary words, they are not described in the thesaurus, and are frequent in the corpus under analysis. For example, *mistral* is described in RuWordNet as a specific wind, but in the current corpus French helicopter carrier Mistral is actively discussed.

Frequent examples of such named entities are names of football, hockey and other teams popular in Russia coinciding with ordinary Russian words or geographical names (*Zenith, Dynamo*, etc.). Some teams can have nicknames, which are written with lowercase letters in Russian and cannot be revealed as named entities. For example, Russian word *iriska* means a kind of candy. In the same time, it is nickname of Everton Football Club (*The Toffees*).

Some discrepancies can be based on frequent multiword expressions, which can be present or absent in the thesaurus. A component w_1 of multiword expression w_2 can be distributionally similar to other words frequently met with w_2 or it can be similar to words related to the whole phrase w_1 w_2.

For example, word *toplenyi* (*rendered*) occurs in the phrase *toplenoe maslo* (*rendered butter*) 78 times of 112 of its total frequency. Because of this, this word is the most similar to word *mindalnyi* (adjective to almond), which is met in the phrase *mindalnoe maslo* (*almond oil*) 57 of 180 times. But two words *toplenyi* and *mindalnyi* cannot be considered as sense-related words.

Explanation	**Number of words**
Morphological ambiguity	**32**
Misprints	**2**
Unknown names, including	**11**
- Sports teams names	6
- Sports teams nick names	2
Multiword expression	**5**
Incorrect relations	**6**
Lost Senses	**10**

Table 2. Explanations of discrepancies between thesaurus and distributional similarities for Top-100 of ProblemList

4.3 Correcting Thesaurus Relations

In some cases, the idea of distributional similarity is clear, but the revision cannot be made in the thesaurus. We found two types of such cases. First, such epithet as *gigant* (*giant*) in the current corpus is applied mainly to large companies (*IT-giant, cosmetics giant,* etc.). But it can be strange to provide the relations between words *giant* and *company* in a thesaurus. The second case can be seen on the similarity row to word *massazhistka* (*female masseur*), comprising such words as hairdresser, housekeeper, etc. This is a kind of specialists in specific personal services but it seems that an appropriate word or expression does not exist in Russian. So, we do not have any language means to create a more detailed classification of such specialists.

Another interesting example of a similarity grouping is the group of “flaws in the appearance”: word *tsellyulit* (*cellulite*)[3] is most similar to words: *morshchina* (*crease of the skin*), *perkhot'* (*dandruff*), *kariyes* (*dental caries*), *oblyseniye* (*balding*), *vesnushki* (*freckles*). It can be noted that a bald head or freckles are not necessary flaws of a specific person, but on average they are considered as flaws. On the other hand, such a phrase as *nedostatki vneshnosti* (*flaws in the appearance*) is quite frequent in Internet pages according to global search engines. Therefore maybe it could be useful to introduce the corresponding synset for correct describing the conceptual system of the modern personality.

But also real problems of thesaurus descriptions were found. They included word relations, which could be presented more accurately (6 cases in Top-100). For example, word *tamada* (*toastmaster*) was linked to a more general word, not to *veduschii* (*master of ceremonies*), and it was revealed from the ProblemList analysis.

4.4 Senses Unattested in Thesaurus

Also significant missed senses including serious errors for verbs were found. As it was mentioned before, in Russian there are groups of related verbs: perfective, imperfective, and reflexive. These verbs usually have a set of related senses, and also can have their own separate senses. In the comparison of discrepancies between TBag and Dbag of verbs, it was found that at least for 25 verbs some of senses were unattested in the current version of the thesaurus, which can be considered as evident mistakes. For example, the imperfective sense of verb *otpravlyatsya* (*depart*) was not presented in the thesaurus.

Several dozens of novel senses, which are the most frequent senses in the current collection, were identified. Most such senses are jargon (sports or journalism) senses, i.e. *derbi* (derby as a game between main regional teams) or *naves* as a type of a pass in football (*high-cross pass*). Also several novel senses that belong to information technologies were detected: *proshivka* (*firmware*), *socset'* (abbreviation from *sotsial'naya set'* – *social network*).

Several colloquial (but well-known) word senses absent in RuWordNet were found. For example, verb *obzech'sya* in the literary sense means ‘burn oneself’. In Dbag the colloquial sense ‘make a mistake’ is clearly seen.

For word *korrektor* (*corrector*), two most frequent unattested senses were revealed. The Dbag of this word looks as a mixture of cosmetics and stationary terms: *guash'* (*gouache*), *kistochka* (*tassel*), *tonal'nyy* (*tonal*), *chernila* (*ink*), *tipografskiy* (*typographic*), etc.

[3] https://en.wikipedia.org/wiki/Cellulite

Word	Absent senses	Type and Domain	Distributional Similarity to	Frequency
otpravlyatsya	Missed imperfective to Perfective *otpravit'sya*	Mistake, General	*otpravit'sya* 0.85	10712
oblachnyy (adjective for *oblako* – cloud)	As in cloud computing, cloud service, etc.	Newly appeared, Computer	*geterogennyy* (heterogenous) 0.5	4662
konyushnya	Formula-1 team	Newly appeared, Sport, Jargon	*gonshchik* (racer) 0.63	3854
derbi (derby)	Derby as a game between main regional teams	Sport, Jargon,	*match* (match as a competition) 0.62	3743
leibl (label)	As a record company	Newly appeared, Journalism, Jargon,	*plastinka* (vinyl disk) 0.56	2147
proshivka (firmware)	As firmware (kind of software)	Newly appeared, Computer,	*updeit* (update), 0.67	1311
korrektor (corrector)	Two senses 1. as correction fluid 2. as a cosmetic preparation (skin corrector)	Newly appeared, 1. Stationary, 2. Cosmetics	*guash'* (gouache) 0.49 *pomada* (lipstick) 0.44	237
perkussiya (percussion)	As percussion musical instrument	Newly appeared, Borrowing from English, Music	*klavishniy* (key-based) 0.73	146

Table 3. Examples of found ambiguous words with missed senses

Currently, about 90 evident missed senses (different from named entities), which are most frequent senses of the word in the collection, are identified. Among them, 10 words are in the Top-100 of the ProblemList. Table 3 presents the examples of found ambiguous words with missed senses that should be added to RuWordNet.

4.5 Other Cases

In some cases, paths longer than 3 should be used to provide better correspondence between thesaurus-based and corpus-based similar words (10 words in the top 100 words of ProblemList), for example, such 4-step paths as two hypernyms, then two hyponyms.

Four words in the top-100 have strange corpus-based similarities. We suppose that it is because of the presence of some news articles in Ukrainian.

5 Conclusion

In this paper we discuss the usefulness of applying a checking procedure to existing thesauri. The procedure is based on the analysis of discrepancies between corpus-based and thesaurus-based word similarities. We applied the procedure to more than 30 thousand words of Russian wordnet RuWordNet, classified sources of differences between word similarities and found some serious errors in word sense description including inaccurate relationships and missing senses for ambiguous words.

We highly recommend using this procedure for checking wordnets. It is possible to find a lot of unexpected knowledge about the language and the thesaurus.

In future, we plan to develop an automatic procedure of finding thesaurus regularities in DBag of problematic words, which can make more evident what kind of relations or senses are missed in the thesaurus.

Acknowledgments

The reported study was funded by RFBR according to the research project N 18-00-01226 (18-00-01240). The author is grateful to Ekaterina Parkhomenko for help with programming the approach.

.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations* Association for Computational Linguistics, pages 7-12. http://www.aclweb.org/anthology/S07-1002.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Edinburgh, Scotland, pages 1–11. http://www.aclweb.org/anthology/W11-2501.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages. 1352-1362. http://www.aclweb.org/anthology/P13-1133.

Francis Bond and Shan Wang. 2014. Issues in building English-Chinese parallel corpora with WordNets. In *Proceedings of the Seventh Global Wordnet Conference,* pages 391-399. http://www.aclweb.org/anthology/W14-0154.

Jose Camacho-Collados, Claudio Bovi, Luis Espinosa-Anke, Siergio Oramas, Tomasso Pasini, Enriko Santus, Vered Schartz, Roberto Navigli and Horacio Saggion. 2018. SemEval-2018 Task 9: hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation,* pages 712-724. http://www.aclweb.org/anthology/S18-1115.

Jose Camacho-Collados. 2017. Why we have switched from building full-fledged taxonomies to simply detecting hypernymy relations. arXiv preprint arXiv:1703.04178

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1), pages 109-120. https://doi: 10.1515/pralin-2017-0013.

Paul Cook and Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. https://www.aclweb.org/anthology/Y11-1028

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Lea Fremann and Mirella Lapata. 2016. Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistic*s. V. 4. pages 31-45. https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00081

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics, pages 67-71. http://www.aclweb.org/anthology/W11-2508.

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 the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pages. 259-270. http://www.aclweb.org/anthology/P14-1025

Natalia Loukachevitch and Boris Dobrov. 2002. Development and Use of Thesaurus of Russian Language RuThes. In *Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation*.(LREC 2002), pages 65-70.

Natalia Loukachevitch, German Lashevich and Boris Dobrov, Boris. 2018. Comparing Two Thesaurus Representations for Russian. In *Proceedings of Global WordNet Conference GWC-2018*, pages 35-44.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, *21*(5), 773-798. https:// doi:10.1017/S135132491500011X

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*. V. 41, №. 2, pages 10. http://doi.acm.org/10.1145/1459352.1459355

Roberto Navigli and Simone Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pages 217-250. https://doi.org/10.1016/j.artint.2012.07.001

Alexander Panchenko, Natalia Loukachevitch, Dmitrii Ustalov, Denis Paperno, Christian Meyer, and Natali Konstantinova. 2015. Russe: The first workshop on russian semantic similarity. In *Proceeding of the Dialogue 2015 Conference,* pages 89-105.

Alexander Panchenko, Anastasiya Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch.

.

2018. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In *Proceedings of Intern. conference Dialogue-2018*, pages 547--564.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 801-808. http://www.aclweb.org/anthology/P06-1101.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. http://www.aclweb.org/anthology/W04-0811