ВИРТУАЛЬНЫЙ РУССКИЙ КОРПУС С СЕМАНТИЧЕСКОЙ РАЗМЕТКОЙ И ПОИСК ДЕФЕКТОВ В СЛОВАРЕ-ПОСРЕДНИКЕ

Диконов В. Г. (dikonov@iitp.ru)
ИППИ РАН, Москва, Россия

Порицкий В. В. (v.poritski@gmail.com)
БГУ, Минск, Беларусь

Ключевые слова: векторное пространство, корпус, семантическая разметка, лексические ресурсы, словарь-посредник, семантика

A VIRTUAL RUSSIAN SENSE TAGGED CORPUS AND CATCHING ERRORS IN A RUSSIAN ↔ SEMANTIC PIVOT DICTIONARY

Dikonov V. G. (dikonov@iitp.ru)
IITP RAS, Moscow, Russia

Poritski V. V. (v.poritski@gmail.com)
BSU, Minsk, Belarus

There are areas in computational linguistics, where a word-sense tagged corpus becomes a necessary prerequisite or gives a significant boost to research. Unfortunately, publicly available corpora of this kind are extremely rare and making them from scratch is a very long and costly process. No corpus of Russian with unambiguous word-sense tags has been published so far. This paper describes an experimental approach of creating a virtual equivalent of a Russian sense tagged corpus and putting it to some real use. The virtual corpus was created using two public resources: the English SemCor corpus and our free multilingual semantic pivot dictionary, called the “Universal Dictionary of Concepts”. The dictionary provides information sufficient to find sense-specific translations for nearly all sense-tagged words in SemCor. However, the pivot dictionary itself is under development and we are looking for the ways to improve it. We used the existing Russian volume of the pivot dictionary to calculate lexical context vectors for individual senses of 13,832 Russian words, supposedly equivalent to the

vectors that could be obtained from a real Russian translation of SemCor. Another set of vectors representing real usage of the same Russian words was extracted from a medium-size corpus of Russian without any semantic markup. The vector similarity score proved to be a useful factor in judging the correctness of links between Russian words and word senses similar to ones registered in the Princeton Wordnet. It helped to rank over 21,000 of such links out of 56,000 known and significantly reduce the amount of the manual work required to proofread the dictionary.

Keywords: vector space, corpus, sense tagging, lexical resources, pivot, semantics, dictionary

1. Introduction

The motive for using the approach outlined in the abstract was absence of a free Russian corpus with semantic markup and scarcity of publicly available lexical and semantic resources that would formally describe the meanings of Russian words in a machine-readable form. One of the authors has already invested some effort in plugging the latter gap while developing an open multilingual semantic resource "Universal Dictionary of Concepts", described in [Dikonov, 2013], [Boguslavsky, Dikonov, 2009]. It is further referred to as pivot dictionary. The goal of the work presented here is two-fold. Firstly, we try to ensure good quality of the pivot dictionary by correcting most of the eventual errors. Secondly, we do what is possible to prepare raw data that could be used for developing and improving Russian word sense disambiguation tools (rule-based and statistical models). We shall briefly describe the resources and devices we used in sections 2–5, explain the process of result evaluation in section 6 and present the results in section 7.

2. Pivot dictionary

The Universal Dictionary of Concepts is a repository of fine-grained semantic concepts, which are equivalent to word senses, with translations into several natural languages. The senses are organized into semantic classes, supported by SUMO ontology [Pease, 2011], and linked by a network of semantic relations. The dictionary serves as a lexicon of an artificial computer interlingua called UNL and uses UNL "Universal Words" as unique identifiers of the senses. Its structure makes it a good neutral semantic pivot dictionary, which is not limited to the lexicon of any single natural language.

The first versions were bootstrapped by integrating available free lexical and ontological data, including Princeton Wordnet, by various automatic methods. However, further development along the lines of simple data merging was hampered by the fact that every imported error in the links between words and abstract concepts tends to multiply and produce even more entropy, as soon as already known links get used to classify new data.

Currently the Russian part of the dictionary covers about 33,000 entries, not counting most proper names and multiword terminology. These words are linked to over 56,000 senses, including some specific to the Russian language. The target

quality level at the early automatic data acquisition phase was set at no less than 90% of correct Russian word↔sense links. A lot of work has already been done to improve it by proofreading critical parts of the dictionary. As a result, the estimated percentage of wrong links decreased to approx. 2.5% with about 3% of additional questionable or vague translations[1], such as *гореть (burn)* instead of *полыхать (flare, burn up)* in the sense "Burn brightly", which is a hyponym of *burn* "Undergo combustion". The remaining errors were randomly scattered among more than 56,000 word↔sense pairs. We needed a way to concentrate the errors in a smaller section of the data to reduce labor intensive straightforward editing. One possibility to do it is to employ a vector space model and a corpus.

3. Virtual corpus

Our source of sense-tagged text was SemCor [Mihalcea, 1998]—a subset of Brown corpus with manual tagging by Wordnet 2.1 senses. It was supplemented by SensEval [Kilgarriff, 1998] 2 and 3 benchmark files converted to SemCor format. This gave us 37,698 English sentences containing 724,207 words.

In 2010 we evaluated the potential of sense tags in SemCor to improve the quality of syntactic parsing and made dependency trees for 37,136 (98.5%) sentences in SemCor+SensEval [Dikonov, D'jachenko, 2010]. We used an experimental build of the ETAP-3 parser, which was modified to use external tagging, either manual or from another parser. The use of semantic annotation helped us to build better syntactic trees.

The tree-tagged SemCor+SensEval corpus contains both the original sense tagging and extra tags given to words, which can have only one meaning according to the pivot dictionary. The total number of sense tagged instances of English words in our corpus is 144,723 (21,978 unique senses) and they make 782,009 unique pairs. Unfortunately, large portions of SemCor have very sparse annotation, e.g. only verbs are disambiguated in 166 files out of 352. Sentences with only one tagged word or excessive linear distance between tags are useless for measuring co-occurrence of senses. We used both linear window sized 1, 2 or 3 words on two sides and syntactic dependencies to learn co-occurrence statistics for pairs of senses. The largest set of such pairs was built by combining dependencies with a 3-word window. This option was used for all subsequent steps.

Our pivot dictionary provides sense-tagged words in the corpus with sense-specific translations into Russian. For example, the English noun *bill* has many senses and each is translated into Russian in a different way:

- "A sign posted in a public place"—*афиша*
- "A statute in draft before it becomes law "—*законопроект*
- "A statement of money owed for goods shipped or services rendered"—*счет*
- "A piece of paper money"—*купюра, банкнота*
- "A list of particulars "—*список*
- "A male given name "—*Билл*

[1] This estimation was done by taking random samples of 200 links and counting defects found during proofreading of the samples.

- "A brim that projects to the front to shade the eyes"—*козырек*
- "Horny projecting mouth of a bird "—*клюв.*

In figure 1 below it is translated as *законопроект* because the corpus has a tag indicating that *bill* means "A statute in draft before it becomes law". Almost all sense-tagged words receive one or more Russian equivalents is this way. Russian translation equivalents which are multiword phrases are converted into sequences of independent lemmas.

From such data we can compute mutual co-occurrence frequencies of Russian words which should be close enough to the frequencies, that would be observed in a real Russian translation of the same text. The numbers would never actually match for many reasons. One of them is that some words in SemCor lack sense tags, like the words *school*, *student*, *reduce* in figure 1. Another is that the pivot dictionary is incomplete and some words remain untranslated. Nevertheless, at this step we obtain potentially useful data approximating the data that could be obtained from a non-existent Russian SemCor-like corpus.

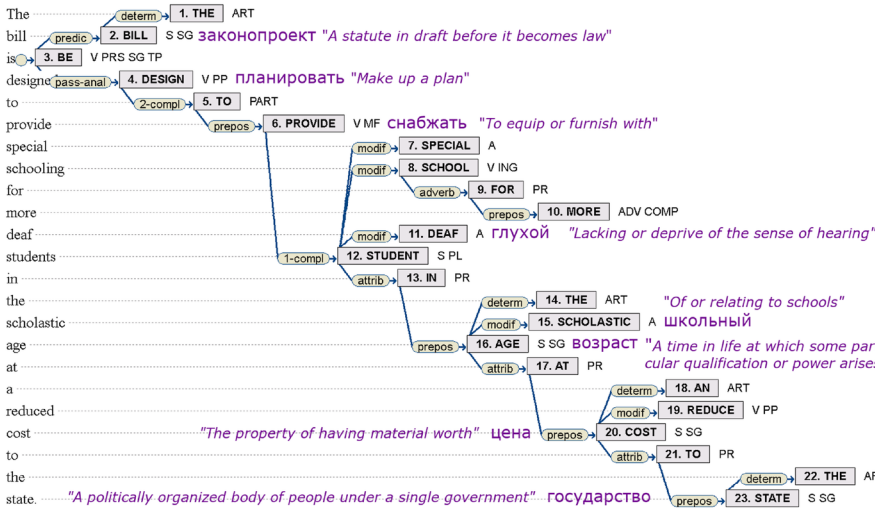

Fig. 1: A sample of SemCor data with extra annotation: dependency tree produced by ETAP3 automatic parser, guided by semantic tags, and Russian translations of semantically tagged words

4. Semantic vectors

The set of numbers associated with a word sense and showing, how many times different other words occur in the context of the word used to express that sense, makes up a numerical vector. The vectors for different senses of the same word are different, because the context neighbors of the senses usually differ. We call such vectors semantic, as opposed to lexical vectors associated with non-disambiguated words.

Predicted semantic vectors are sensitive to defects of the pivot dictionary used to produce them. It is possible to identify vectors based on wrong translations by comparing them to other vectors built from some benchmark data and representing correct use of the Russian words. Ideally, the benchmark should provide semantic vectors representing the same word senses, but in our case this was not possible.

However, the semantic vectors produced from the virtual corpus can still be compared with lexical vectors representing all available contexts of Russian translations for the semantic vector's base sense. Any false translations in the semantic vectors reduce similarity with the benchmark lexical vectors. A false translation of the vector base sense causes the comparison to be made with entirely different set of contexts, belonging to the word which does not have this sense. This results in a very noticeable difference. For example, one of the real detected errors was that the sense *weld(icl>join>do,agt>thing,obj>thing) "Join together by heating*" was wrongly linked to the Russian word *сплачивать*, which means "Unite closely or intimately" *weld(icl>unite>do,cob>thing,agt>volitional_thing,obj>thing)*. The first sense is likely to be found in phrases like *сваривать панели (weld panels)* but the wrong translation meant that the virtual corpus offered **сплачивать панели* instead. The latter phrase never occurs in real Russian texts and the corresponding semantic vector fails comparison with the benchmark vector of the word *сплачивать*.

A bunch of correct semantic vectors, representing all senses of some word, put together is likely to show close semblance to the benchmark lexical vector of the word. Our hypothesis is that a single correct semantic vector still has enough similarity with its benchmark to be distinguishable from random errors.

5. Benchmark corpus

We used a benchmark Russian corpus of approximately 17 mln tokens. The corpus contains samples of present-day Russian fiction (10 mln tokens) and newspaper articles (the rest). To obtain lemmas, we merged the output of MyStem [Segalovich 2003] and TreeTagger with Russian parameter file [Schmid 1992; Sharoff et al. 2008]. In most cases TreeTagger works as a disambiguator over the output of MyStem, but its lexical coverage is rather narrow, since the parameter file has been trained on the disambiguated portion of Russian National Corpus. For some of the wordforms not recognized by TreeTagger, MyStem produces a unique lemma, so that a simple fallback strategy is available. MyStem also helps to deduce lemmas for several trickier classes of tokens: compound nouns, age designations like *23-летний* (23 years old) etc.

The key idea was to design a high-dimensional vector space, such that both senses and their purported Russian equivalents could be represented as points thereof. The basis of this space was made up, rather straightforwardly, of lemmas attested both in the virtual sense-tagged Russian corpus and in the benchmark corpus. This amounts to ca. 10^4 distinct lemmas. To compute a suitable similarity score between two 10^4-dimensional vectors is a tractable task, so no further dimensionality reduction was done. Note however, that moderate size of the basis brings not only

computational ease, but also scalability issues: no matter how large the benchmark corpus is, most of the co-occurrence statistics collected in it will remain unused.

Four sets of benchmark vectors have been computed, with symmetric linear context window size ranging from 1 to 4 tokens. We used cosine similarity which had performed well in earlier experiments on coarse-grained synonym identification [Poritski, Volchek 2013]. Pairwise similarity computations were run on raw co-occurrence counts as well as on PMI weighted vectors (for the definition of PMI weighting scheme see, e.g., [Manning, Schütze 2003, p. 178]). The similarity score value are numbers between 1 and 0. The value of 1 is given to pairs of vectors which are elementwise proportional (high similarity). Absence of similarity (totally different vectors) is marked with 0 (for raw frequency counts) or −1 (with PMI applied). However negative scores under PMI weighting turned out to be quite rare and were counted as zeros

As a result we built several versions of similarity scores, using different ways of finding word pairs and calculating vector similarity. Each version was presented as a table containing three columns: a Russian word, pivot word sense designation (UNL universal word) and the similarity score of the semantic and benchmark vectors, as shown in figure 2. The number of lines was 22,874, which corresponds to the number of word senses occurring in SemCor+SensEval and translatable through the pivot dictionary.

беспокоить	bother(icl>trouble> cause>do,agt>thing,obj> person,met>uw)	0.756537995710645	*To cause inconvenience or discomfort to*
камень	stone(icl>material> thing,equ>rock)	0.0501434171547867	*Material consisting of the aggregate of minerals*
камень	stone(icl>natural_ object>thing,equ>rock)	0.0438500966889403	*A lump or mass of hard consolidated mineral matter*
~~гореть~~ → полыхать	burn_up(icl>burn>occur, equ>flare,obj>thing)	0.0178584089774577	*Burn brightly*
~~сплачивать~~ → сваривать	weld(icl>join>do,agt> thing,obj>thing)	0	*Join together by heating*

Fig. 2. Lines from the similarity score table with comments and corrections

6. Evaluation

The links in the similarity tables were sorted by decreasing of the similarity score. At this point we needed to find, which word↔sense links were wrong. It was done in several iterations.

At first, one of the tables was deemed most promising by comparing the score and position of a couple of already known errors and subjected to selective manual examination. All bad links found were marked as errors or overly vague translations by different symbols. Samples of 200 lines were taken from different parts of the table,

starting from lines 0 (highest scores, 1 error), 8,000 (scores of ~0.02, 3 errors), 10,100 (scores of ~0.009, 13 errors), 12,127 (scores of ~0.001, 7 errors) and 18,600 (zero similarity scores, 16 errors). This first attempt produced a test set of 63 defects (40 errors and 23 vague translations). It also showed that the probability of errors in word↔sense pairs increased as vector similarity score dropped and the concentration of errors in different parts of the table changed from 0.5% to 8% per 200 line sample.

This allowed us to do a better numerical estimation and choose another table, which seemed more likely to have the optimum parameters. We tried to find a threshold in similarity score. Again, 200 line samples were taken starting from lines 4,800 (scores of ~0.05, 8 errors), 6,800 (scores of ~0.04, 8 errors), 8,500 (scores of ~0.03, 10 errors), 10,000 (scores of ~0.024, 17 errors), 11,800 (scores of ~0.015, 10 errors) and further 580 lines with zero score (46 errors). This time the overall error distribution curve, similar to ones shown in figure 3, actually got flatter and some fluctuations became visible.

The same word↔sense links receive different scores in tables built with different settings, so the errors found in the first chosen table were scattered around the second table randomly. Combining the error sets produced a more evenly distributed test set. A review of the combined set confirmed that there are certain other factors helpful in selecting likely errors. Our pivot dictionary allows to differentiate between polysemic and monosemic words in all supported languages, including Russian. It also has technical flags showing the amount of attention given to each word↔sense link. It is rather obvious that polysemic words and less reviewed links are more suspicious and our data confirmed it. As a result, further samples were gathered and all non-reviewed polysemic words occurring in the zero similarity zone were checked. This gave us a test set of 1,141 bad links (659 errors and 512 vague translations).

7. Results

The following two diagrams show the distribution of all defects discovered until now in the cosine similarity score tables, as shown in figure 2. Figure 3 illustrates the dependency between the score (X) and the number of defects (Y) in one of the best ranking tables which was computed with the following settings:

- co-occurrence of senses in the virtual corpus within linear window of width 3 and within the range of syntactic dependencies;
- co-occurrence of words in the benchmark corpus within linear window of width 2;
- PMI weighting applied.

Higher score value means better alignment between semantic and benchmark vectors. The first defects start to appear when the score drops below 0.16. The three colored bars correspond to all defects, errors properly and vague translations.

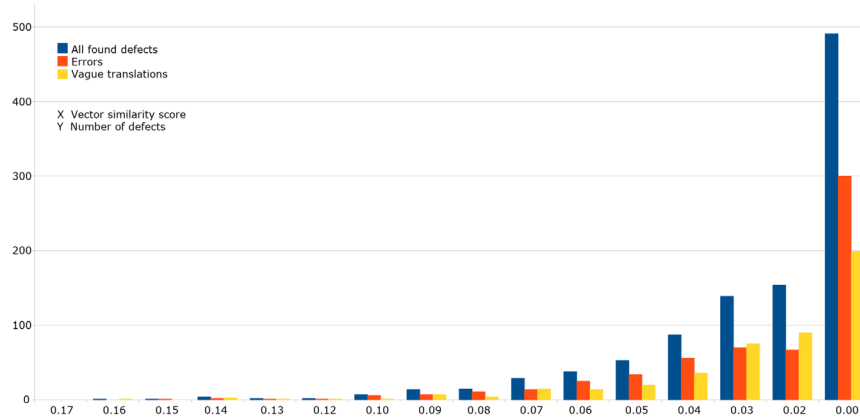

Fig. 3. Distribution of defects according to the similarity score

The diagram in figure 4 shows, how the number of defects per 1,000 lines changes from beginning to the end in several versions of the table, all ranked by cosine similarity score in decreasing order. Each version represents a different combination of options used to produce the benchmark vectors and is shown by a different curve in the diagram. The possible options are:

- linear window width (1–4) in the benchmark corpus;
- PMI weighting (yes/no);
- frequency count strategy for vector elements, which are known to be synonyms (sum all / take maximum).

Here PMI weighting is always on, because it makes the result consistently better. The horizontal line shows calculated average level, when all defects are scattered randomly.

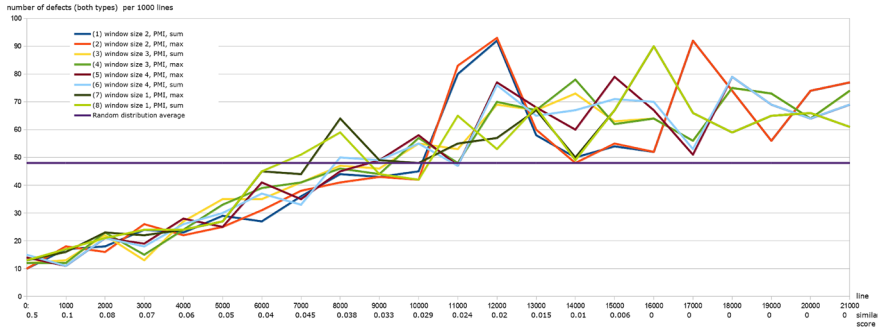

Fig. 4. Distribution of defects by line number in the similarity score tables

Already reviewed data confirms previous quality estimations of the pivot dictionary. Before this work, the predicted total number of errors in its Russian part ran

between 2,070 and 2,275 out of 56,000 links. The same number for vague translations was about 2,500. Defects already exposed by the procedures described in section 6 constitute approximately 29% of the predicted total number of errors and about 20% of vague translations. At the same time, the reviewed portion of the vector similarity table at the time of writing (6,218 links) is only 10.6% of the total contents of current Russian dictionary.

8. Conclusion

The virtual corpus proved to be a reasonably useful tool for narrowing down the search for anomalies in relations between Russian words and pivot word senses. It can make the process of discovering and fixing dictionary defects almost 3 times faster than baseline. This is a sound practical outcome.

Although the two sets of vectors are based on different things—individual word senses and non-disambiguated words—comparing them was fruitful. It is possible to merge semantic vectors representing all registered senses of the same word to obtain a predicted lexical vector. That would make a completely fair "apples to apples" comparison. It has not been done because we cannot state at this point that all words in our pivot and Wordnet have complete description of their polysemy. We may do it at a later stage to facilitate the search of Russian words, which lack certain key senses in our pivot dictionary.

The amount of publicly available sense disambiguated corpus data is dismally limited for English and is simply zero for Russian. There are some Russian resources though, which are not public in the sense that they cannot be freely downloaded and used. One example is the semantic annotation layer within the Russian National Corpus (RNC) [Lashevskaja, Shemanaeva 2008]. It is different from the SemCor data used in this work in several respects. RNC does not label individual instances of words with any concrete word senses. Instead, they receive a set of taxonomic, mereological and derivational tags, assigned by software according to the RNC's internal semantic dictionary. Unlike SemCor, no manual disambiguation has been done in RNC. The tags, however, were filtered with manually formulated rules to remove tags violating known contextual restrictions. The resulting partially disambiguating semantic markup is used by the online RNC search engine.

Even if a real manually sense-tagged Russian corpus will be developed, we can hardly expect it to be larger than SemCor. It is possible to improve the situation by supplementing one small corpus with another small corpus made for a different language. It requires a reliable pivot, which allows to match or relate different sets of word sense labels. Such supplementing may work for projects that generalize the senses to a coarser grain level or rely on statistics to smooth over small problems. The current version of the pivot dictionary is available for download from the git repository at https://github.com/dikonov/Universal-Dictionary-of-Concepts. The tree-tagged virtual corpus files with Russian translations of the sense tagged English words will be published when the process of proofreading the links will be near completion. Snapshots can be found at https://github.com/dikonov/SemCorRus.

References

1. *Boguslavsky I., Dikonov V.* (2009), Universal Dictionary of Concepts [Universal'nyj slovar' konceptov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferentsii "Dialog 2009"], Bekasovo, pp. 91–96.
2. *Dikonov V., Boguslavsky I.* (2009), Semantic Network of the UNL Dictionary of Concepts, Proceedings of the SENSE Workshop on Conceptual Structures for Extracting Natural Language Semantics, Moscow, available at: http://ceur-ws.org/Vol-476/paper2.pdf.
3. *Dikonov V., D'jachenko P.* (2010), An Experiment in Automatic Building of English Dependency Trees Governed by Externally Provided Incomplete Tagging [Eksperiment po postroeniju sintaksicheskoj struktury anglijskih predlozhenij s ispol'zovaniem zaranee izvestnyh fragmentarnyh dannyh], Proceedings of ITaS'10, Gelendzhik, pp. 310–319.
4. *Kilgarriff A.* (1998), SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs, Proceedings of LREC'98, Granada, pp. 581–588.
5. *Lashevskaja O. N., Shemanaeva O. Yu.* (2008), Semantic Annotation Layer in Russian National Corpus: Lexical Classes of Nouns and Adjectives, Proceedings of LREC'08, Marrakech, pp. 3355–3358.
6. *Manning C. D., Schütze H.* (1999), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA.
7. *Mihalcea, R.* (1998), SemCor semantically tagged corpus, SenseEval 2 & 3 data in SemCor format. http://www.cse.unt.edu/~rada/downloads.html
8. *Pease A.* (2011), Ontology: A Practical Guide, Articulate Software Press, Angwin, CA.
9. *Poritski V. V., Volchek O. A.* (2013), Building a Vector Space Model of Meaning for Russian: A Preliminary Study [Postroenie vektornoj semanticheskoj modeli na osnove russkojazychnyh tekstov: pervye eksperimenty], Proceedings of ITaS'13, Svetlogorsk, pp. 114–119.
10. *Schmid H.* (1992), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, pp. 44–49.
11. *Segalovich I.* (2003), A Fast Morphological Algorithm With Unknown Word Guessing Induced by a Dictionary for a Web Search Engine, Proceedings of MLMTA'03, Las Vegas, pp. 273–280.
12. *Sharoff S., Kopotev M., Erjavec T., Feldman A., Divjak D.* (2008), Designing and Evaluating a Russian Tagset, Proceedings of LREC'08, Marrakech, pp. 279–285.