WSD in monolingual dictionaries for Russian WordNet

Daniil Alexeyevsky[1], Anastasiya V. Temchenko

[1] School of Linguistics, Faculty of Humanities, National Research
University Higher School of Economics
21/4 Staraya Basmannaya Ulitsa, Moscow, 105066, Russia

dalexeyevsky@hse.ru, avtemko@gmail.com

Abstract

Russian Language is currently poorly supported with WordNet-like resources. One of the new efforts for building Russian WordNet involves mining the monolingual dictionaries. While most steps of the building process are straightforward, word sense disambiguation (WSD) is a source of problems. Due to limited word context specific WSD mechanism is required for each kind of relations mined. This paper describes the WSD method used for mining hypernym relations. First part of the paper explains the main reasons for choosing monolingual dictionaries as the primary source of information for Russian language WordNet and states some problems faced during the information extraction. The second part defines algorithm used to extract hyponym-hypernym pair. The third part describes the algorithm used for WSD

1 Introduction

After the development of Princeton WordNet (Fellbaum, 2012), two main approaches were widely exploited to create WordNet for any given language: dictionary-based concept (Brazilian Portuguese WordNet, Dias-da-Silva *et al.*, 2002) and translation-based approach (see for example, Turkish WordNet, Bilgin *et al.*, 2004). The last one assumes that there is a correlation between synset and hyponym hierarchy in different languages, even in the languages that come from distant families. Bilgin *et al.* employ bilingual dictionaries for building the Turkish WordNet using existing WordNets.

Multilingual resources represent the next stage in WordNet history. EuroWordNet, described by Vossen (1998), was build for Dutch, Italian, Spanish, German, French, Czech, Estonian and English languages. Tufis *et al.* (2004) explain the methods used to create BalkaNet for Bulgarian, Greek, Romanian, Serbian and Turkish languages. These projects developed monolingual WordNets for a group of languages and aligned them to the structure of Princeton WordNet by the means of Inter-Lingual-Index.

Several attempts were made to create Russian WordNet. Azarova *et al.* (2002) attempted to create Russian WordNet from scratch using merge approach: first the authors created the core of the Base Concepts by combining the most frequent Russian words and so-called “core of the national mental lexicon”, extracted from the Russian Word Association Thesaurus, and then proceeded with linking the structure of RussNet to EuroWordNet. The result, according to project’s site[1], contains more than 5500 synsets, which are not published for general use. Group of Balkova *et al.* (2004) started a large project based on bilingual and monolingual dictionaries and manual lexicographer work. As for 2004, the project is reported to have nearly 145 000 synsets (Balkova *et al.* 2004), but no website is available (Loukachevitch and Dobrov, 2014). Gelfenbeyn *et al.* (2003) used direct machine translation without any manual interference or proofreading to create a resource for Russian WordNet[2]. Project RuThes by Loukachevitch and Dobrov (2014), which differs in structure from the canonical Princeton WordNet, is a linguistically motivated ontology and contains 158 000 words and 53 500 concepts at the moment of writing. YARN (Yet Another RussNet) project, described

[1] http://project.phil.spbgu.ru/RussNet/, last update June 14, 2005

[2] Available for download at http://www.wordnet.ru

by Ustalov (2014), is based on the crowdsourcing approach towards creating WordNet-like machine readable open online thesaurus and contains at the time of writing more than 46 500 synsets and more than 119 500 words, but lacks any type of relation between synsets.

This paper describes one step of semi-automated effort towards building Russian WordNet. The work is based on the hypothesis that existing monolingual dictionaries are the most reliable resource for creating the core of Russian WordNet. Due to absence of open machine-readable dictionaries (MRD) for Russian Language the work involves shallow sectioning of a non machine-readable dictionary (non-MRD). This paper focuses on automatic extraction of hypernyms from Russian dictionary over a limited number of article types. Experts then evaluate the results manually.

1.1 Parsing the Dictionary

As far as our knowledge extends, there is no Russian monolingual dictionary that was designed and structured according to machine-readable dictionary (MRD) principles and is also available for public use.

There exist two Russian Government Standards that specify structure for machine readable thesauri (Standard, 2008), but they are not widely obeyed.

Some printed monolingual dictionaries are available in form of scanned and proof-read texts or online resources. For example, http://dic.academic.ru/ offers online access to 5 monolingual Russian dictionaries and more than 100 theme-specific encyclopedias. Each dictionary article is presented as one unparsed text entry. Resource http://www.lingvoda.ru/dictionaries/, supported by ABBYY, publishes user-created dictionaries in Dictionary Specification Language (DSL) format. DSL purpose is to describe how the article is displayed. DSL operates in terms of *of italic, sub-article, reference-to-article* and contains no instrument to specify type of relations. This seems to be closest to MRD among available resources. Fully automated information extraction is out of the question in this case. When using non-MRD we have faced with number of problems that should be addressed before any future processing can be started:

1. Words and word senses at the article head are not marked by unique numeric identifiers.
2. Words used in article definitions are not disambiguated, so creating a link from a word in a definition to article defining the word sense is not trivial task.
3. Many contractions and special symbols are used.
4. Circular references exist; this is expected for synonyms and base lexicon, but uncalled for in sister terms, hypernyms, and pairs of articles with more complex relations.
5. The lexicon used in definitions is nearly equal to or larger than the lexicon of the dictionary.

In general, ordinary monolingual dictionaries, compiled by lexicographers, were not intended for future automated parsing and analysis. As stated in Ide and Véronis (1994), when converting typeset dictionaries to more suitable format researchers are forced to deal with:

1. Difficulties when converting from the original format, that often requires development of complex dedicated grammar, as previously showed by Neff and Boguraev (1989).
2. Inconsistencies and variations in definition format and meta-text;
3. Partiality of information, since some critical information in definitions is considered common knowledge and is omitted.

Research by Ide and Véronis (1994) gives us hope that using monolingual dictionaries is the best source of lexical information for WordNet. First they show that one dictionary may lack significant amount of relevant hypernym links (around 50-70%). Next they collect hypernym links from merged set of dictionaries and in the resulting set of hypernym links only 5% are missing or inconsistent as compared with expert created ontology.

Their work is partly based on work by Hearst (1998) who introduced patterns for parsing definitions in traditional monolingual dictionaries.

One notable work for word sense disambiguation using text definitions from articles was performed by Lesk (1986). The approach is based on intersecting set of words in word context with set of words in different definitions of the word being disambiguated. The approach was further extended by Navigli (2009) to use corpus bootstrapping to compensate for restricted context in dictionary articles.

In this paper we propose yet another extension of Lesk's algorithm based on semantic similarity databases.

2 Building the Russian WordNet

Specific aim of this work is to create a bulk of noun synsets and hypernym relations between them for further manual filtering and editing. To simplify the task we assume that every word sense defined in a dictionary represents a unique synset. Furthermore we only consider one kind of word definitions: such definitions that start with nominative case noun phrase. E. g.: *rus. ВЕНТИЛЯ́ЦИЯ: Процесс воздухообмена в лёгких. eng.'VENTILATION: Process of gas exchange in lungs'*. We adhere to hypothesis that in this kind of definitions top noun in the NP is hypernym. In order to build a relation between word sense and its hypernym we need to decide which sense of hypernym word is used in the definition. This step is the focus of this work.

2.1 The Dictionary

The work is based on the Big Russian Explanatory Dictionary (BRED) by Kuznetsov S.A. (2008). The dictionary has rich structure and includes morphological, word derivation, grammatical, phonetic, etymological information, three-level sense hierarchy, usage examples and quotes from classical literature and proverbs. The electronic version of the dictionary is produced by OCR and proofreading with very high quality (less than 1 error in 1000 words overall). The version also has sectioning markup of lower quality, with FPR in range 1~10 in 1000 tag uses for the section tags of our interest.

We developed specific preprocessor for the dictionary that extracts word, its definition and usage examples (if any) from each article. We call every such triplet word sense, and give it unique numeric ID. A article can have reference to derived word or synonym instead of text definition. Type of the reference is not annotated in the dictionary. We preserve such references in a special slot of word sense. The preprocessor produces a CSV table with senses.

2.2 Hypernym candidates

Given a word sense *W* we produce a list of all candidate hypernym senses.

Ideally under our assumption the first nominative case noun in *W*'s definition is a hypernym. However, due to variance in article definition styles and imperfect morphological disambiguation used, some words before the actual hypernym are erroneously considered candidate hypernym. To mitigate this we consider each of the first three nominative nouns candidate hypernyms. For each such noun we add each of its senses as candidate hypernym senses.

If sense *W* is defined by reference rather than by textual definition, we add both every sense of referenced word and each of its candidate hypernym senses to the list of candidate hypernym senses of *W*.

2.3 Disambiguation pipeline

We have developed a pipeline for massively testing different disambiguation setups. The pipeline is preceded by obtaining common data: word lemmas, morphological information, word frequency.

For the pipeline we broke down the task of disambiguation into steps. For each step we presented several alternative implementations. These are:

1. Represent candidate hyponym-hypernym sense pair as a Cartesian product of list of words in hyponym sense and list of words in hypernym sense, repeats retained.
2. Calculate numerical metric of words similarity. This is the point we strive to improve. As a baseline we used: random number, inverse dictionary definition number; classic Lesk algorithm. We also introduce several new metrics described below.
3. Apply compensation function for word frequency. We assume that coincidence of frequent words in to definitions gives us much less information about their relatedness than coincidence of infrequent words. We try the following compensation functions: no compensation, divide by logarithm of word frequency, divide by word frequency.
4. Apply non-parametric normalization function to similarity measure. Some of the metrics produce values with very large variance. This leads to situations where one matching pair of words outweighs a lot of outright mismatching pairs. To mitigate this we attempted to apply these functions to reduce variance: linear (no normalization), logarithm, Gaussian, and logistic curve.
5. Apply adjustment function to prioritize the first noun in each definition. While extracting candidate hypernyms the algorithm retained up to three candidate nouns in each article. Our hypothesis states that the first one is most likely the hypernym. We apply penalty to the metric depending

on candidate hypernym position within hyponym definition. We tested the following penalties: no penalty, divide by word number, divide by exponent of word number.

6. Aggregate weights of individual pairs of words. We test two aggregation functions: average weight and sum of best N weights. In the last case we repeat the sequence of weights if there were less than N pairs. We also tested the following values of N: 2, 4, 8, 16, 32.

Finally, the algorithm returns candidate hypernym with the highest score.

2.4 Testing setup

For testing the algorithms we selected words in several domains for manual markup. We determined domain as a connected component in a graph of word senses and hypernyms produced by one of the algorithms. Each annotator was given the task to disambiguate every sense for every word in such domain. Given a triplet an annotator assigns either no hypernyms or one hypernym; in exceptional cases assigning two hypernyms for a sense is allowed.

One domain with 175 senses defining 90 nouns and noun phrases was given to two annotators to estimate inter-annotator agreement. Both annotators assigned 145 hypernyms within the set. Of those only 93 matched, resulting in 64% inter-annotator agreement.

The 93 identically assigned hyponym-hypernym pairs were used as a core dataset for testing results. Additional 300 word senses were marked up to verify the results on larger datasets. The algorithms described were tested on both of the datasets.

2.5 Our Approach to Disambiguation

In this section we describe various alternatives to metric function on step 2 of the pipeline.

One known problem with Lesk algorithm is that it uses only word co-occurrence when calculating overlap rate (Basile *et al.*, 2004) and does not extract information from synonyms or inflected words. In our test it worked surprisingly well on the dictionary corpus, finding twice as many correct hypernym senses as the random baseline. We strive to improve that result for dictionary definition texts.

Russian language has rich word derivation through variation of word suffixes. The first obvious enhancement to Lesk algorithm to account for this is to assign similarity scores to words based on length of common prefix. In the results we refer to this metric as advanced Lesk.

Another approach to enhance Lesk algorithm is to detect cases where two different words are semantically related. To this end we picked up a database of word associations Serelex (Panchenko *et al*, 2013). It assigns a score on a 0 to infinity scale to a pair of noun lemmas roughly describing their semantic similarity. As a possible way to score words that are not nouns in Serelex we truncate a few characters off the ends of both words and search for the best pair matching the prefixes in Serelex. (See prefix “serelex” in Table 1).

We tested several hypotheses on how these two metrics can be used to improve the resulting performance. The tests were: to use only Lesk; to use only Serelex; to use Serelex where possible and fallback to advanced Lesk for cases where no answer was available; and to sum the results of Serelex and Lesk. Since Serelex has a specific distribution of scores we adjusted the advanced Lesk score to produce similar distribution.

For each estimator we performed full search through available variations on steps 3-6 of the pipeline and selected the best on the core set and estimated again on the larger dataset.

Test results are given in the Table 1:

Algorithm	CoreSet	LargeSet
random	30.8%	23.9%
first sense	38.7%	37.7%
naive Lesk	51.6%	**41.3%**
serelex	49.5%	38.0%
advanced Lesk	53.8%	33.3%
serelex with adjusted Lesk fallback	52.7%	36.3%
serelex + adjusted Lesk	52.7%	38.3%
prefix serelex	**53.8%**	38.0%

Table 1. Precision of different WSD algorithms.

3 Discussion

The low resulting quality of disambiguation seems to be a result of several factors: overall difficulty of the task (inter-annotator agreement is 64%), quality of input dictionaries, quality of used similarity database. We also seem to have missed some important linguistic or systemic features of text as well. Notably, the algorithms presented are still generically-applicable and do not use hypernym information.

Despite the low precision in determining the exact hypernyms, the pipeline produces thematically related chains of words. Examples of chains, extracted by *prefix Serelex* algorithm are given below with English translation and comparison to Princeton WordNet (here “>>” symbolises *IS_A* relation):

- *rus. спираль >> кривая >> линия* eng.‘spiral >> curve >> *line’* compared to *PWN spiral >> curve, curved shape >> line >> shape >> attribute >> abstraction >> entity*
- *rus. передняя >> комната >> помещение* eng. ‘anteroom >> room >> premises’ compared to PWN *anteroom >> room >> area >> structure >> artifact >> whole >> object >> physical entity >> entity*
- *rus. рост >> высота >> расстояние* eng. ‘stature, height >> height >> distance’ compared to PWN *stature, height >> bodily property >> property >> attribute >> abstraction >> entity*

Dictionary parsing quality appears to be crucial for the current work, and the dictionary we selected provides us with a huge set of difficulties: abbreviations; alternating language in sense definitions; not all head words are lemmas (e.g. plural for nouns that have singular); poor quality of sectioning in OCR. Sectioning within BRED presents a large problem due to underspecified vaguely nested nature of sections. Properly digitized openly published Russian dictionary is really wished for.

Another problem with the dictionary is presence of nearly-identical definitions for the same term. Due to restricted context in dictionary in some cases it is difficult even for a human annotator to guess correctly whether a given pair of definitions describes the same concepts or two very distinct ones. This is especially true with abstract terms like *time (rus.: время),* but physical entities like *field (rus.: поле)* also present such troubles.

One further step to building the Russian WordNet is to differentiate hypernyms from synonyms and co-hyponyms. Currently we hope to achieve this through classification of definitions and developing morphosyntactic templates to match different relation types within them. This is out of the scope of the current article though.

4 Conclusion

In this work we present a new pipeline for disambiguating and testing disambiguation frameworks for building WordNet relations from raw dictionary data in Russian language[3].

We described new algorithm for hypernym disambiguation which performs somewhat better than baseline in cases where annotators agree. The possibility for better disambiguation of specific relation types within dictionaries to be still open.

The resulting network, though noisy, is very suitable for rapid manual filtering.

References

Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., and Oparin, I. 2002. *Russnet: Building a lexical database for the russian language.* In Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas: 60-64.

Basile, P., Caputo, A., and Semeraro, G. 2014. *An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model*. In Proceedings of COLING: 1591-1600.

Bilgin, O., Çetinoğlu, Ö., and Oflazer, K. 2004. *Building a wordnet for Turkish*.Romanian Journal of Information Science and Technology, 7(1-2):163-172.

Balkova, V., Sukhonogov, A., and Yablonsky, S. 2004. *Russian wordnet. From UML-notation to Internet/Intranet Database Implementation.* In Proceedings of the Second Global Wordnet Conference.

Dias-da-Silva, B. C., de Oliveira, M. F., and de Moraes, H. R. 2002. *Groundwork for the development of the Brazilian Portuguese Wordnet.* In Advances in natural language processing:189-196.

Fellbaum, C. 2012. *WordNet.* The Encyclopedia of Applied Linguistics.

Gelfenbeyn, I., Goncharuk, A., Lehelt, V., Lipatov, A. and Shilo, V. 2003. *Automatic translation of WordNet semantic network to Russian language*. In Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003.

Hearst, M. A. 1998. *Automated discovery of WordNet relations.* WordNet: an electronic lexical database: 131-153.

Ide, N., Véronis, J. 1994. *Machine Readable Dictionaries: What have we learned, where do we*

[3] Available at http://bitbucket.org/dendik/yarn-pipeline

go. In Proceedings of the International Workshop on the Future of Lexical Research, Beijing, China: 137-146.

Ide, N., Véronis, J. 1993. *Refining taxonomies extracted from machine-readable dictionaries*. In Hockey, S., Ide, N. Research in Humanities Computing 2, Oxford University Press.

Kuznetsov S.A. Кузнецов, С. А. 2008. *Новейший большой толковый словарь русского языка*. СПб.: РИПОЛ-Норинт.

Lesk, M. 1986. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.* In Proceedings of the 5th annual international conference on Systems documentation: 24-26

Loukachevitch, N., Dobrov, B. 2014. *RuThes linguistic ontology vs. Russian Wordnets.*GWC 2014: Proceedings of the 7th Global Wordnet Conference: 154–162.

Navigli, R. (2009, March). *Using cycles and quasi-cycles to disambiguate dictionary glosses.* In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: 594-602.

Neff, M. S., and Boguraev, B. K. 1989. *Dictionaries, dictionary grammars and dictionary entry parsing.* In Proceedings of the 27th annual meeting on Association for Computational Linguistics: 91-101.

Panchenko, A., Romanov, P., Morozova, O., Naets, H., Philippovich, A., Romanov, A., and Fairon, C. 2013. *Serelex: Search and visualization of semantically related words.* In Advances in Information Retrieval: 837-840.

Standard, G. O. S. T. 2008. Standard 7.0.47-2008, *Format for representation on machine-readable media of information retrieval languages vocabularies and terminological data*.

Tufis, D., Cristea, D., Stamou, S. 2004. *BalkaNet: Aims, Methods, Results and Perspectives.* A General Overview In: D. Tufiş (ed): Special Issue on BalkaNet. Romanian Journal on Science and Technology of Information.

Ustalov, D. 2014. *Enhancing Russian Wordnets Using the Force of the Crowd.* In Analysis of Images, Social Networks and Texts. Third International Conference, AIST 2014. Springer International Publishing: 257-264.

Vossen, P. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Network*. Dordrecht.