РУТЕЗ-LITE, ОПУБЛИКОВАННАЯ ВЕРСИЯ ТЕЗАУРУСА РУССКОГО ЯЗЫКА РУТЕЗ

Лукашевич Н. В. (louk_nat@mail.ru),
Добров Б. В. (dobrov_bv@mail.ru),
Четверкин И. И. (ilia2010@yandex.ru)

Московский государственный университет
им. М. В. Ломоносова, Москва, Россия

Ключевые слова: автоматическая обработка текстов, WordNet, тезаурус, толкования

RUTHES-LITE, A PUBLICLY AVAILABLE VERSION OF THESAURUS OF RUSSIAN LANGUAGE RUTHES

Loukachevitch N. V. (louk_nat@mail.ru),
Dobrov B. V. (dobrov_bv@mail.ru),
Chetviorkin I. I. (ilia2010@yandex.ru)

Lomonosov Moscow State University, Moscow, Russia

The paper presents RuThes-lite, a publicly available version of RuThes linguistic ontology, which has been developed for more than fifteen years and is intended for automatic document processing. RuThes has considerable similarities with WordNet: inclusion of concepts based on senses of real text units, representation of lexical senses, detailed coverage of word senses. At the same time the differences include attachment of different parts of speech to the same concepts, formulating names of concepts, attention to multiword expressions, intentional inclusion of terms of the sociopolitical domain, a set of conceptual relations. RuThes-lite was generated from RuThes on the basis of the most frequent words in a contemporary news collection. Besides, we describe additional data, which have been specially prepared for RuThes-lite publication: morph-syntactic labeling of thesaurus text entries and assignment of glosses to concepts.

Keywords: natural language processing, WordNet, thesaurus, part-of-speech tagging, gloss

Introduction

WordNet is one of popular resources used for natural language processing and information-retrieval applications (Fellbaum, 1998). For many languages projects on developing national wordnets have been initiated. At least four attempts to create a Russian wordnet are known (Azarowa, 2008; Gelfenbeyn et al., 2003; Balkova et al., 2008; Braslavski et al. 2013).

In spite of its popularity in computational linguistics applications, WordNet initially was created as a justification of a psycholinguistic theory (Miller, 1998), its structure and relations were based on psycholinguistic experiments and were not initially intended for natural language processing tasks. So, some constructive features of WordNet hinder its applications in automatic text processing.

These problems include: the initial absence of relations between different parts of speech with the same meaning (*adopt—adoption*—now this problem is corrected with special relations (Clark et al., 2008)); the absence of links between semantically related senses of derivate words (*initiation—initiator*); so-called "tennis problem", indicating the absence of relations between synsets of the same domain (*plane—airport*); problems in introducing synsets for multiword expressions. Some of these problems are partially corrected with the generation of additional data. For example, for "tennis problems"—the system of WordNet domains has been developed (Bentivogli et al., 2004; Bhatt et al., 2014), in many wordnets derivational links between synsets for labeling word derivations have been introduced (Azarova et al., 2002; Koeva et al., 2008).

Research on better structures of computer-oriented language resources is not a simple task because one should not only create a quite large resource, but also demonstrate its quality and characteristics of its structure in various NLP applications.

In this paper we will describe the structure and the current state of newly published RuThes-lite linguistic ontology, which is intended for use automatic text processing of Russian documents. RuThes-lite is a public part of RuThes ontology, which has been developed since 1994 and was applied in several tasks of natural language processing and information retrieval (Loukachevitch, Dobrov, 2014). In contrast to WordNet, in RuThes we implemented a unified representation for different parts of speech, lexical units and domain terms, single words and multiword expressions, adopted a set of conceptual relations, tested in applications.

The structure of this paper is as follows. In Section 1 we briefly describe the structure of RuThes linguistic ontology. Section 2 explains how RuThes-lite was generated from RuThes. In Section 3 we describe additional linguistic information that was specially prepared and provided for RuThes-lite. Section 4 reports some details on RuThes-lite publication.

1. RuThes Linguistic Ontology

RuThes Thesaurus of Russian language can be called a linguistic ontology for natural language processing, i.e. an ontology, where the majority of concepts are introduced on the basis of actual language expressions. RuThes is a hierarchical network of concepts. Each concept has a name, relations with other concepts, a set of language expressions (words, phrases, terms), whose meanings correspond to the concept.

In RuThes, a unit is presented not by a set of similar words or terms, as it is done in the WordNet thesaurus, but by a concept—as a unit of thought, which can be associated with several synonymic language expressions. Every concept should have distinctions from related concepts that are independent from context and should be expressed in a specific set of relations or associated language expressions—text entries.

Each concept should have a concise and unambiguous name. Such names often help to express, delimit the denotational scope of the concept. Besides, the names facilitate the analysis of the results of natural language processing. If necessary, a concept may have a gloss, which is not a part of the concept name.

Words and phrases, which meanings are represented as references to the same concepts of the thesaurus, are called ontological synonyms. Ontological synonyms can comprise:

- words belonging to different parts of speech (*стабилизация* (*stabilization*), *стабилизировать* (*stabilize*), *стабилизационный* (*stabilizing*))—therefore the number of RuThes concepts is approximately 2.5 times less than in a wordnet-like resource of the same size;
- language expressions relating to different linguistic styles, genres;
- idioms and even free multiword expressions (for example, synonymous with single words).

A row of ontological synonyms can include quite a large number of words and phrases. So, a concept *ДУШЕВНОЕ СТРАДАНИЕ* (*wound in the soul*) has more than 20 text entries including such as: *боль*, *боль в душе*, *в душе наболело*, *душа болит*, *душа саднит*, *душевная пытка*, *душевная рана*, *душевный недуг*, *наболеть*, *рана в душе*, *рана в сердце*, *рана души*, *саднить* (several English translations may be as follows: *wound*, *emotional wound*, *pain in the soul* etc.).

Introducing a concept linguists specially search for multiple lexical variants (especially multiword ones) that can express the same sense. An introduced text entry should have the sufficient frequency in contemporary text collections. With this aim usually Yandex.news service or Yandex search engine are used. We do not use Russian National corpus for this check because it does not comprise necessary volumes of lexical data.

An ambiguous word is assigned to several concepts—this is the same approach as in WordNet. For example, word *коса* is assigned to three different concepts:

- *КОСА* (*ВЫСТУП ЗЕМЛИ*) (*tongue of land*)
- *КОСА ВОЛОС* (*braid of hair*)
- *КОСА* (*СЕЛЬСКОХОЗЯЙСТВЕННОЕ ОРУДИЕ*) (*scythe*)

Language expressions whose sense can serve as a basis for a separate concept in RuThes belong not only to the general vocabulary, but also can be terms of specific subject domains within the broad scope of social life (economy, law, international relations, politics, transport, banks, etc.), so-called *sociopolitical domain* (Loukachevitch and Dobrov, 2004). This was done because many professional concepts, terms, and slang of these domains penetrate easily into the general language, and can be widely discussed in mass media: news reports and newspaper articles. The appearance of these terms in general news is not accidental. People interact with professionals and professional domains in everyday life and therefore should possess relevant terminology. In addition, such a scope of concepts facilitates the application of RuThes in specialized subdomains of the broad socio-political domain. Examples of such concepts in RuThes include: *EMERGENCY LOAN*, *TAX EXEMPTION*, *IMPORT TAX*, *DEMOGRAPHIC INDICATOR* etc.

The relations in RuThes are only conceptual, not lexical (as antonyms or derivational links in wordnets). They are constructed as more formal, ontological relations of traditional information-retrieval thesauri (Z39.19, 2005). The set of conceptual relations includes:

- the class-subclass relation;
- the part-whole relation applied with the following restriction: the existence of the concept-part should be strictly attached to the concept-whole (so tree can grow in many places therefore concept *TREE* cannot be directly linked to concept *FOREST* with the part-whole relation, the additional concept *FOREST TREE* should be introduced);
- the external ontological dependence when the existence of a concept depends on the existence of another concept (in such a way forests depend on the existence of trees) (Guarino, Welty, 2002). In RuThes we denote this relation as association with indexes: *asc1* is directed to the main concept, *asc2*—to the dependent concept;
- In the very restricted number of cases symmetric associations between concepts can be established.

The main idea behind this set of relations is to describe the most essential, reliable relations of concepts, which are relevant to various contexts of concept mentioning.

Thus, RuThes has considerable similarities with WordNet: the inclusion of concepts based on senses of real text units, representation of lexical senses, detailed coverage of word senses. At the same time the differences include attachment of different parts of speech to the same concepts, formulating names of concepts, attention to multiword expressions, intentional inclusion of terms of the sociopolitical domain, the set of conceptual relations. The more detailed description of RuThes and RuThes-based applications can be found in (Loukachevitch, Dobrov, 2014) or (Lukashevich, 2011).

At present RuThes includes 54 thousand concepts, 158 thousand unique text entries (75 thousand single words), 178 thousand concept-text entry relations, more than 215 thousand conceptual relations.

2. Generating RuThes-Lite

We decided to publish partially RuThes creating RuThes-lite version, which includes approximately 100,000 unique text entries. Such a resource should contain the most frequent words of contemporary Russian and at the same time include the upper levels of the RuThes hierarchy to preserve its property to be a connected net.

Frequency estimation of words is based on a news collection. Automatic news flow processing is one of the most important directions of natural language processing technologies. News and newspaper articles are categorized, clustered, from them named entities, relations, facts, opinions are extracted, special news services collect, process them and provide access to news data. In addition to news, such collections also contain newspaper articles, legal acts, and even literature pieces published in newspapers and journals.

The used news collection comprised 2 millions newspaper articles and news reports from around 2000 news sources. So RuThes text entries were matched with texts of this text collection and the revealed text entries were ordered by frequency decrease.

The beginning of the obtained list was cleaned up from compositional text expressions (usually synonymic variants of single words), names of persons and organizations, professional legal or economy terms. From this cleaned list we selected approximately 30 thousand the most frequent text entries (in fact, it was an iterative procedure), most of them were single words.

The frequency list begins quite traditionally: *быть*, *год*, *сообщать*, *мочь*, *время*, *стать*, *слово.* At the end of the list the following words are situated: *биофизика*, *абонементный*, *чаевые*, *спиваться*, *распашной* etc.

These selected text entries were used as seeds for concept extraction. In RuThes-lite the following concepts were included:

- All concepts having text entries from the seed list—seed concepts,
- Upper level concepts to seed concepts, that is concepts, which have a path of hyponymy or part relations to the selected concepts.
- For extracted concepts all their text entries and relations between each other are also extracted. The current version of RuThes-lite contains 26,365 concepts, around 96,941 unique words and expressions, 115,349 senses (concept-text entry links), 108,000 relations between concepts.

3. Preparing Additional Data for RuThes-Lite

The basic data of RuThes comprise:

- the list of thesaurus concepts including the concept identifier and its name,
- the list of text entries of thesaurus in the dictionary form and in the lemmatized form (each word in a text entry is lemmatized);
- the list of relations between text entries and concepts;
- the list of relations between concepts.

For the public version we prepared additional data useful for applications. In this paper we describe two type of additional information: morpho-syntactic labels of thesaurus text entries and glosses extracted from Wiktionary and assigned to thesaurus concepts.

Below we will describe techniques utilized for preparing these data for RuThes-lite.

3.1. Morpho-Syntactic Labeling

As indicated above, in RuThes all parts of speech, single words and multiword expressions are presented as text entries to the same concept. Each text entry is provided with the representation as a sequence of lemmas—words in dictionary forms (lemmatic representation): for example, *голубые фишки—голубой фишка*. This information was introduced manually. The part-of-speech tags of text entries were absent because it was supposed that part-of-speech labeling is produced during automatic text processing with a morphological tagger. However, for many applications information about the part of speech of a text entry, the head word of a multiword word expression can be essential.

In RuThes-lite we provide additional morphological and syntactic information about a text entry: the part of speech of a single word; the head of a phrase and the part of speech of a text entry as a whole (= part of speech of its head word) for a multiword expression.

The labeling was fulfilled automatically with morphological processing of a text entry and its lemmatic representation—the use of the both types of information decreases potential morphological ambiguity.

So, now such text entry as *уголовное дело* (*criminal case*) has the following information about own structure:

уголовное дело:
уголовный дело (words in lemmatic forms)
NG (noun group)
дело (head word)
Adj N (parts of speech for every word in the text entry).

It should be noted that word *"дело"* is morphologically ambiguous but the description eliminates the ambiguity.

3.2. Assignment of Glosses to Concepts

In RuThes names of concepts play an important role. They should be clear and unambiguous and should inform a native speaker about the meaning of a concept. Only for a small number of concepts some additional explanations are provided.

But in WordNet-like resources glosses are often used as prominent information in various applications, for example, generation of a sentiment vocabulary (Bacianella et al., 2010), calculation of similarity measures (Pedersen et al., 2004), lexical disambiguation (Agirre, Soroa 2009) and others. Therefore some wordnet developers try to mine glosses from lexical resources (Henrich et al., 2011).

For RuThes-lite we also made the first step in providing concepts with glosses explaining their intended meanings—we automatically extracted glosses from Russian Wiktionary, matched glosses and concepts and selected the most appropriate gloss for a concept. The problem here is how to select the best gloss describing the meaning of a concept, provided that:

- a concept can have several text entries;
- each of these text entries can have several senses in Wiktionary and in RuThes, these two sets of senses for a text entry can be different in size. For example, word *стрелка* is related to seven concepts in RuThes-lite and has eleven senses in Wiktionary.

  To extract a gloss for a given concept the following procedure was implemented:
- for all text entries of a concept, candidate glosses from Wiktionary are extracted. Glosses are cleaned from examples because examples can accidentally influence on matching. Then glosses are lemmatized, functional words are removed. So for every gloss we obtain the vector of lemmas.

For a concept we also create a vector. The vector includes all text entries of a concept, text entries of super-class concepts and whole-concepts. If a word is met several times in these text entries then its frequency in the vector is enhanced.

For example, concept *СТРЕЛКА РЕК* (*river spit*) has the following text entries and relations:

СТРЕЛКА РЕК (*river spit*)

(Syn: *стрека*, *стрелка рек*, *стрелка между реками*, *стрелочный*)
class: *КОСА* (*ВЫСТУП ЗЕМЛИ*) (Syn: *береговая коса*, *коса*, *коса берега*, *намывная коса*, *песчаная коса*)
asc1: *ВПАДЕНИЕ РЕКИ*, *ПОТОКА* (*stream inflow*) (Syn: *впадать*, *впадание*, *впадение*).

Therefore the following concept vector is generated for matching:
((*коса* 5) (*стрелка* 3) (*река* 2) (*стрелочный* 1) (*берег* 1) (*береговой* 1) (*намывной* 1) (*песчаный* 1) (*впадение* 1) (*впадать* 1) (*впадание* 1))

The relevant gloss from Wiktionary is as follows: "*узкий продолговатый участок суши, окружённый с трёх сторон водой, особенно на слиянии двух рек*". Its vector looks like:
((*узкий* 1) (*продолговатый* 1) (*участок* 1) (*суша* 1)
(*окруженный* 1) (*окружить* 1) (*особенно* 1) (*особенный* 1)
(*сторона* 1) (*вода* 1) (*слияние* 1) (*река* 1).

To this vector synonyms and hypernyms described in Wiktionary are added. In this case hypernyms *мыс* and *полуостров* are indicated for this sense in Wiktionary and therefore they are added with 1 count to the vector.

The matching weight between the concept vector and a gloss vector is equal to the scalar product of vectors without normalization, that is in the above-mentioned example the weight is equal 3 and this is the largest weight for all candidate glosses. As a result of this procedure around 60% of RuThes-lite concepts have obtained glosses.

Random testing of assigned glosses showed that 91% glosses were matched correctly to relevant concepts. Some glosses were missed, so recall is equal to 85%, F-measure—87.9%. In similar experiments on linking of GermaNet and Wiktionary, authors of (Henrich et al., 2012) report that the best matching results (84.3%) were achieved using all relations of GermaNet with weights tuned for every type of relations. We did not use subclass relations (hyponyms) and parts, so some improvement of matching can be possible.

Currently, the list of extracted glosses is checked out by linguists, which remove irrelevant glosses and correct glosses with some problems of extraction.

4. Publication of RuThes

At present, RuThes thesaurus is partially involved in several commercial projects with other organizations and therefore it cannot be published as a whole. But the interest in a large thesaurus of Russian language is considerably growing therefore we decided to publish RuThes partially.

The first publicly available version of RuThes (RuThes-lite) is available from http://www.labinform.ru/ruthes/index.htm. We plan to distribute RuThes-lite as free for noncommercial use (Attribution-NonCommercial-ShareAlike 3.0 Unported license).

Conclusion

In this paper we presented RuThes linguistic ontology. This resource has been developed for a long time (more than fifteen years) and was used as a resource in various applications of NLP and information retrieval such as conceptual indexing, semantic search, query expansion, automatic text categorization and clustering, automatic summarization of a single document and multiple documents.

Now the first version of RuThes—RuThes-lite has been published. In this paper we described its structure and current state. We hope that this resource, having the broad and detailed lexical and terminological coverage of contemporary Russian news articles and official documents, will facilitate development of NLP techniques and research for Russian language.

In addition to publication of RuThes, we plan to automatically generate a resource in WordNet-like form (RuWordNet) including such relatively new information as WordNet domains and derivational links, which is widely discussed in the WordNet community. We think that RuThes contains enough data for generation such a resource. Its publication will be an important step in developing Russian semantic resources, connection with WordNet community.

Acknowledgements

The work is partially supported by Dmitrii Zimin Dynastia Foundation with financial support of Yandex founders.

References

1. *Agirre E., Soroa A.* (2009), Personalizing pagerank for word sense disambiguation, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.
2. *Azarowa I.* (2008), RussNet as a Computer Lexicon for Russian, Proceedings of the Intelligent Information systems IIS-2008, pp. 341–350.
3. *Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin, I.* (2002), Russnet: Building a lexical database for the Russian language, Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation, Las Palmas, pp. 60–64.
4. *Balkova V., Suhonogov A., Yablonsky S.* (2008), Some Issues in the Construction of a Russian WordNet Grid, Proceedings of the Forth International WordNet Conference, Szeged, Hungary, pp. 44–55.
5. *Baccianella S., Esuli A., Sebastiani F.* (2010), SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, Proceedings LREC-2010, Vol. 10, pp. 2200–2204.
6. *Bentivogli L., Forner P., Magnini B., Pianta E.* (2004), Revising WordNet domains hierarchy: semantics, coverage, and balancing, Proceedings of COLING 2004, Geneva, Switzerland, pp. 101–108.
7. *Bhatt B., Kunnath S., Bhattacharyya P.* (2014), Graph Based Algorithm for Automatic Domain Segmentation of WordNet, Proceedings of Global WordNet Conference GWC-2014.
8. *Braslavski P. I., Mukhin M. Y., Lyashevskaya O. N., Bonch-Osmolovskaya A. A., Krizhanovsky A. A., Egorov P.* (2013), Yarn Begins. http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BraslavskiyP_YARN.pdf
9. *Clark P., Fellbaum Ch., Hobbs J.* (2008), Using and Extending WordNet to Support Question-Answering, Proceedings of Fourth Global WordNet Conference (GWC'08), Hungary, Szeged, pp. 111–119.
10. *Gelfenbeyn I., Goncharuk A., Lehelt V., Lipatov A., Shilo V.* (2003), Automatic translation of WordNet semantic network to Russian language, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003.
11. *Guarino N., Welty Ch.* (2000), Ontological Analysis of Taxonomic Relationships, In Conceptual Modeling (ER-2000), Springer, Berlin Heidelberg, pp. 210–224.
12. *Fellbaum Ch.* (1998), WordNet: An Electronic Lexical Database, Cambridge, MA, MIT Press.
13. *Henrich V., Hinrichs Er., Vodolazova T.* (2012), Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary, Proceedings of LREC-2012.
14. *Loukachevitch N., Dobrov B.* (2004), Sociopolitical Domain as a Bridge from General Words to Terms of Specific Domains, Proceedings of Second International WordNet Conference GWC-2004, Brno, pp.163–168.
15. *Loukachevitch N.* (2009), Concept Formation in Linguistic Ontologies, Conceptual Structures: Leveraging Semantic Technologies. Proceedings of ICCS-2009, Springer Verlag, LNAI-5662, pp. 2–22.

16. *Loukachevitch N.* (2011), Thesauri in information-retrieval tasks, Moscow, Moscow University publishing house.
17. *Loukachevitch N., Dobrov B.* (2014), RuThes Linguistic Ontology vs. Russian Wordnets. Proceedings of Global WordNet Conference GWC-2014, Tartu.
18. *Koeva S., Krstev C., Vitas D.* (2008), Morpho-semantic relations in Wordnet–a case study for two Slavic languages. In *Proceedings of the Fourth Global WordNet Conference*, pp. 239–254.
19. *Pedersen T., Patwardhan S., Michelizzi J.* (2004), WordNet: Similarity: measuring the relatedness of concepts, Demonstration Papers at HLT-NAACL 2004. Association for Computational Linguistics.
20. *Z39.19.* (2005), Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO.