Transforming RuThes Thesaurus to Generate Russian WordNet

Natalia Loukachevitch

Lomonosov Moscow State University,
Leninskie Gory, 1/4, Moscow, Russia
louk_nat@mail.ru

Abstract. In this paper we describe the semi-automatic process of transforming the Russian language thesaurus RuThes to WordNet-like thesaurus, called RuWordNet. In this procedure we attempted to achieve two main characteristic features of WordNet-like resources: division of data into part-of-speech-oriented structures with cross-references between them and providing a set of relations similar to WordNet-like resources.

Keywords: natural language processing, thesaurus, WordNet, synset

1 Introduction

WordNet-like resources [1] are one of the most popular resources used for natural language processing, wordnet projects have been initiated for many languages in many countries.

At least four attempts to create a Russian wordnet are known. RussNet [2] began development from scratch and at this moment appears to be quite small (not more than 20,000 synsets). Two other Russian wordnets were generated using automated translation [3,4]. The first one is publicly available but represents the direct translation from Princeton WordNet without any manual revision. The last Russian wordnet project YARN (Yet Another Russian wordNet) was initiated in 2012 and is being created using a crowdsourcing approach; it currently contains mainly synsets with small number of relations between them [5].

For Russian, there exists the RuThes thesaurus, a linguistic ontology, which structure has differences from the WordNet approach. RuThes is a more ontology-oriented resource: thesaurus concepts have unique names, text entries of all parts of speech can be linked to the same concept. The RuThes relations are more formal conceptual relations. The current size of the published version of RuThes (RuThes-lite 2.0), accessible for non-commercial use, is more than 115 thousand text entries. RuThes was specially created for information retrieval and natural language applications, it can be used in most applications where WordNet is usually utilized, but researchers and practitioners want to have a Russian wordnet.

In this paper, we describe the transformation of RuThes data to WordNet-like resource, called RuWordNet. In this process we try to reproduce two main features of the Princeton WordNet structure such as the organization in the form of part-of-speech lexical nets and the basic set of relations. The current volume of RuWordNet is the same as the published version of RuThes-lite 2.0 (115 thousand entries). It can be seen in Internet and can be obtained in the XML format.

The paper is organized as follows. The second section reviews the related work. The third section considers main features of the WordNet structure. The fourth section describes the main structure of RuThes and its differences from WordNet. The fifth section presents the transformation process from RuThes to RuWordNet and achieved results. The sixth section compares web-representations of RuThes and RuWordNet.

2 Related work

The most straightforward approach to the development of WordNet-like resources from scratch is a difficult task, which usually takes years of work. The approach to fasten the creation of a national wordnet is to translate Princeton WordNet to the target language [6]. Wordnet-like resources obtained with automatic translation can be generated fast enough but also require a lot of efforts to be manually revised.

An intermediate approach between the above-mentioned ultimate points, which can be considered as quite usual, is to translate the top 5000 concepts of the Princeton WordNet (core WordNet) and then extend this hierarchy manually, using local dictionaries. This approach was accepted in the development of EuroWordnet [6] and Danish wordnet DanNet [7].

Analysing previous approaches for national wordnet development, authors of FinnishWordNet (FiWN) decided to use manual translation of Princeton WordNet synsets by professional translators. The direct translation approach was based on the assumption that most synsets in PWN represent language-independent real-world concepts. Thus, the semantic relations between synsets were also assumed mostly language-independent, so the structure of PWN could be reused as well. In such a way, Finnish wordnet, FinnWordNet (FiWN), was created by translating more than 200,000 word senses in the English Princeton WordNet (PWN) 3.0 in 100 days [8].

Braslavski et al [5] intend to create a Russian wordnet (YARN) utilizing Russian Wiktionary and crowdsourcing. Wiktionary is a crowdsourced dictionary and thesaurus that exists for many languages. Wiktionary pages related to a specific word can contain a lot of useful information about word senses, including a list of lexical senses, definition and examples for a lexical sense, lexical relations (synonyms, antonyms, hyponyms, hypernyms), which are represented as links to Wiktionary pages. However, there are also some problems in word senses description, which can hamper creating a WordNet-like resource especially for inexperienced crowdsourcers:

– a lexical link leads not to a specific sense but to the whole word page,
– synonyms can be described as partial synonyms, this is a very vague notion;

– lexical relations are not symmetrical. For example, word w1 is indicated as a synonym to word w_2, but word w_2, is not indicated as a synonym to word w_1. In other examples, word w_1 is indicated as a synonym to word w_2, but word w_2 is indicated as a hypernym to word w_1.

3 Basic Structure of Princeton WordNet

The structure of Princeton University's WordNet (and other wordnets) is based on sets of partial synonyms – synsets, organized in hierarchical part-of-speech-based lexical nets for nouns, adjectives, verbs, and adverbs. Each part-of-speech net has its own system of relations between synsets.

The most frequent relation between noun synsets is the hyponym-hypernym relation. Also since 2006 in Princeton WordNet class-instance relations denoted as Instance Hypernym and Instance Hyponym [9] were introduced. Such relations substituted hyponym-hypernym relations for synsets of proper nouns designating unique entities such as cities, countries, concrete persons, etc. This work was made under the influence of the ontologists' point of view on "confusion between individuals and concepts" [10].

The noun relationships also include part-whole relations, which are subdivided into proper part-whole relations (wing is a part of bird), member parts (tree is a member of forest), and material (snow is a substance of snowball). Parts can have several wholes (wing is a part of bird, bat, insect, or angel).

For all parts of speech, the lexical relation of antonymy can be established. Lexical relations link lexemes, not whole synsets. In Princeton Wordnet, the antonymy relation is the main type of relations for descriptive adjectives [11], which were described only with the relations of antonymy and similarity. For example, for the word *heavy*, the word *light* is indicated as an antonym, such words as hefty, ponderous, massive are linked to heavy with the relation "similar to". Other wordnets, such as GermaNet [12] or Polish WordNet (PlWordNet) [13], changed this approach and introduced taxonomic relations (hyponymy-hyperonymy) for adjectives.

Verbs in WordNet are mainly linked with hyponym-hypernym relations. Besides, they have their own unique relations not described for nouns or adjectives: entailment (buy – pay) and causation (give – have, kill – die). The WordNet entailment relation is a relation between two verbs V_1 and V_2 that holds when the sentence "Someone V_1" logically entails "Someone V_2" and there is the temporal inclusion of eventV_1 into V_2 or vice versa [1]. The causation relation can be also considered as a subtype of a general logical entailment relation but there is not temporal inclusion between corresponding situations [1].

4 RuThes Structure and Relations

RuThes and WordNet are both thesauri that are lexical resources where semantically related words and expressions are collected together into synsets or concepts between which formalized relations are set. When applying both thesauri to natural language processing, the same steps should be made such as matching between a text and a thesaurus and employing the described thesaurus relations if necessary. The most evident differences between the two types of resources are as follows.

First, in RuThes there is no division into subnets according to different parts of speech that is words of any part of speech can be linked to the same concept if they mean the same (so called derivative or part-of speech synonyms).

Therefore, second, in RuThes it is often very difficult or even impossible to apply traditional tests of synonymy detection such as substitution of synonyms in sentences [14,15]. Tests checking the denotational scope of lexemes are usually applied in the following way: "if entity X can be called with word W_1, then we can call it also with word W_2" and vice versa regardless of specific context. The second test consists in formulation of explicit differences of one concept from other concepts. These differences can be fixed in the unique concept name. Thus, above-mentioned issues of RuThes such as denotational tests, denotational distinctions between concepts, and unique names of concepts make RuThes much closer to ontological resources in an imaginary scale from lexical resources to formal ontologies than WordNet-like thesauri. RuThes can be called a linguistic (lexical) ontology for natural language processing.

Third, 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 [16], which were designed to describe very broad, unstructured domains. 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. For example, trees can grow in many places not only in forests 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) [17]. In RuThes we denote this relation as association with indexes: asc_1 is directed to the main concept, asc_2 – 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. Also this set of relations allows us to describe domain terminologies or domain-specific ontologies, combine descriptions of lexical and domain-specific knowledge in the same resource.

The relation of ontological dependence is very convenient for describing conceptual relations between concepts corresponding to multiword expressions and concepts of their component words (such as nature protection and nature), which allows easier introducing such concepts and describing useful "horizontal" relations.

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 of names of concepts, attention to multiword expressions, the set of conceptual relations, etc. The more detailed description of RuThes and RuThes-based applications can be found in [18] or [19].

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. The published version of RuThes, RuThes-lite 2.0, contains 115 thousand text entries. It was singled out from full RuThes on the basis of words and phrases used in current Russian news flows with exclusion of several specific domains [20].

5 Generating RuWordNet from RuThes

According to the guidelines of world-known WordNet thesaurus, the first version of Russian wordnet (RuWordNet) was created.

In our opinion, one of the most distinctive features of WordNet-like resources is their division into synset nets according to parts of speech. Therefore all text entries of RuThes-lite 2.0 were subdivided into three parts of speech: nouns (single nouns, noun groups, or preposition groups), verbs (single verbs and verb groups), adjectives (single adjectives and adjective groups). We have obtained 29,297 noun synsets, 12,865 adjective synsets, and 7,636 verb synsets.

This subdivision was based on the morphosyntactic representation of RuThes-lite 2.0 text entries, which was fulfilled semi-automatically. Therefore, a small number of mistakes because of particle treatment (verbs or adjectives) or substantivated adjectives can appear. Currently all found mistakes are corrected. The divided synsets were linked with the relation of part-of-speech synonymy.

The hyponym-hypernym relations were established between synsets of the same part of speech. These relations include direct hyponym-hypernym relations from RuThes-lite 2.0. In addition, the transitivity property of hyponym-hypernym relations was employed in cases when a specific synset did not contain a specific part of speech but its parent and child had text entries of this part of speech. In such cases the hypernymy-hyponymy relation was established between the child and the parent of this synset.

Similar to the current version of Princeton WordNet, in RuWordNet class-instance relations are also established. By now, they had been generated semi-automatically for geographical objects.

The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like resources. Now RuWordNet contains 3.5 thousand part-whole relations. The part-whole relations include the following subtypes:

- functional parts (nostrils – nose),
- ingredients (additives – substance),
- geographic parts (Sevilia – Andalusia),
- members (monk – monastery),
- dwellers (Moscow citizen – Moscow),
- temporal parts (gambit – chess party)
- inclusion of processed, acitivities (industrial production – industrial cycle)

Adjectives in RuWordNet similarly to German or Polish wordnets are connected with hyponym-hypernym relations. Adjectives often have POS-synonymy links to nouns, but also can have POS-synonyms to verb synsets.

In the current RuWordNet representation of Russian verbs, part-whole relations can be seen. For example, synset видеть во сне, сниться, грезиться, присниться, привидеться во сне, пригрезиться, пригрезиться во сне" [to dream] is linked to synset спать, поспать, доспать, соснуть, досыпать, почивать, проспать, просыпать [to sleep] with the part-whole relation. Such a relation between the translation equivalents [to dream, to sleep] exists also in Princeton WordNet and called 'entailment relation'. Christian Fellbaum wrote in [1] that "the entailment relation between verbs resembles meronymy between nouns, but meronymy is better suited to nouns than to verbs". Thus, the simple renaming of the part-whole relations between verbs in RuWordNet into entailment relations is possible and correct.

Antonymy relations are conceptual relations in RuWordNet, that means they link synsets, not single lexemes. They are introduced for all parts of speech, mainly for synsets denoting properties and states.

6 Publication of RuThes and RuWordNet on the Web

RuThes-lite 2.0 and RuWordNet are published in form of static web-pages. Looking through RuThes[1], the user should select a letter to begin, then choose an initial trigram of a word, and then click on a proper word. For example, selecting Russian word двор [yard] the user can find three concepts associated with this word, relations of these concepts, and other text entries attached to the same concepts. Further, the navigation through concepts or text entries is possible.

In the similar representation of RuWordNet[2], there is the initial division to parts of speech, which the user should select, then the user should find a word. In the RuWordNet representation, there are no concepts, each synset contains text entries belonging to the same part of speech, POS-synonymy links to other parts of speech are indicated. Thus, in the representation RuThes looks more as an ontology, and RuWordNet is presented more as a lexical net.

[1] http://www.labinform.ru/pub/ruthes/index.htm

[2] http://www.labinform.ru/pub/ruwordnet/index.htm

7 Conclusion

In this paper we have described the semi-automatic process of transforming the Russian language thesaurus RuThes (in version, RuThes-lite 2.0) to WordNet-like thesaurus, called RuWordNet. In this procedure we attempted to achieve two main characteristic features of wordnet-like resources: division of data into part-of-speech-oriented structures with cross-references between them and providing a set of relations similar to wordnet-like relations.

Both thesauri, RuThes-lite 2.0 and RuWordNet, are currently published as static web-pages. Also RuWordNet can be seen through web interface[1]. Researchers can obtain both types of thesauri, compare them in applications. In future, we will continue to add new types of relations to RuWordNet including the domain relation, the cause relation, the entailment relation, etc.

Acknowledgments. This work is supported by the Russian Science Foundation (grant N16-18-02074).

References

1. Fellbaum, C.: A semantic network of English verbs, WordNet: An electronic lexical database, 153–178 (1998)
2. Azarowa, I.: RussNet as a Computer Lexicon for Russian, Proceedings of the Intelligent Information systems IIS-2008, 341–350 (2008)
3. Gelfenbeyn, I., Goncharuk, A., Lehelt, V., Lipatov, A., Shilo, V.: Automatic translation of WordNet semantic network to Russian language, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003 (2003)
4. Balkova, V., Suhonogov, A., Yablonsky, S.: Some Issues in the Construction of a Russian WordNet Grid, Proceedings of the Forth International WordNet Conference, Szeged, Hungary, 44–55 (2008)
5. Braslavski, P., Ustalov, D., Mukhin, M.: A Spinning Wheel for Yarn: User Interface for a Crowdsourced Thesaurus, In Proceedings of EACL-2014, Gothenberg, Sweden,101–104 (2014)
6. Vossen, P.: Introduction to EuroWordNet. In EuroWordNet: A multilingual database with lexical semantic networks, Springer Netherlands, 1–17 (1998)
7. Pedersen, B., Nimb, S., Asmussen, J., Sorensen, N., Trap-Jensen, L., Lorentzen, H.: DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary, Language resources and evaluation, 43(3), 269–299 (2009)
8. Linden, K., Niemi, J.: Is it possible to create a very large wordnet in 100 days? An evaluation, Language resources and evaluation, 48.2, 191–201 (2014)
9. Miller, G., Hristea, F.: WordNet nouns: Classes and instances, Computational linguistics, 32(1), 1–3 (2006)
10. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Understanding Top-Level Ontological Distinctions. Proc. of IJCAI 2001 Workshop on Ontologies and Information Sharing, 26-33 (2001)

[1] http://ruwordnet.ru/

11. Gross, D., Miller, K.J.: Adjectives in WordNet, International Journal of Lexicography, 3(4), 265–277 (1990)
12. Kunze, C., Lemnitzer, L.: Lexical-Semantic and Conceptual relations in GermaNet, In Storjohann P (ed) Lexical-semantic relations: Theoretical and practical perspectives, 163–183 (2010)
13. Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisawska, M., Broda, B.: Words, concepts and relations in the construction of Polish WordNet, In Proceedings of the Global WordNet Conference, Seged, Hungary, 162–177 (2008)
14. Cruse, D.: Lexical Semantics. Cambridge. University Press (1986)
15. Miller, G.: Nouns in WordNet. In WordNet: An Electronic Lexical Database, Fellbaum, C (ed). The MIT Press, 23–47 (1998)
16. Z39.19. Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO (2005)
17. Guarino, N., Welty, C.: Evaluating ontological decisions with ONTOCLEAN, Communications of the ACM, 45(2), 61–65 (2002)
18. Loukachevitch, N., Dobrov, B.: RuThes Linguistic Ontology vs. Russian Wordnets. In Proceedings of the Seventh Global WordNet Conference (GWC 2014), 154–162 (2014)
19. Lukashevich, N.: Thesauri in information-retrieval tasks. Moscow (2011) (in Russian)
20. Loukachevitch, N., Dobrov, B. , Chetviorkin, I.: RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes, In proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2014, 340–350 (2014)