RuWordNet - Russian WordNet by Extraction from RuThes

German Lashevich, Vladimir Ivanov, Shagit Ziganshin
Kazan (Volga region) Federal University
Kazan, Russian Federation
german.lashevich@gmail.com, nomemm@gmail.com, darkrevan13@gmail.com

***Abstract*—We describe the Russian WordNet developed on the basis of the published linguistic ontology ”Thesaurus RuThes”. RuWordNet is a WordNet-like thesaurus produced by converting the RuThes concepts and text entries to synsets with further relations restoration.**

I. Introduction

Contemporary tasks of natural language processing are quite different and require various program tools and linguistic resources, which cannot be created in a single study. To overcome these drawbacks, a lot of tools and resources have been presented.

From this point of view, Russian is a language with a very small number of published resources necessary for natural language processing. The aim of the project is to create and publish a large thesaurus of Russian constructed in compliance with the structure of the famous WordNet[1] thesaurus which could be employed in numerous natural language processing applications.

At least four attempts to create the Russian WordNet are known[2], [3], [4], [5], but the result has not been achieved yet. The RussNet[2] is known to be developed from scratch and at the moment it still to be quite small (not more than 20 thousand synsets). Two other Russian wordnets were generated using automated translation[3], [4]. The first of the two above is publicly available (http://wordnet.ru/) but represents the direct translation from the Princeton Wordnet without any interpreters revision. The webpage of the latter ceased to exist. Project[5] is aimed at creating a large open thesaurus for the Russian language using crowdsourcing. It started in 2013 and it is still not available for practical employment. According to information on the projects webpage there are more than 45 thousand of synsets and around 120 thousand of words but there are no any relations between synsets above.

The paper mainly deals with the semi-automatic generation of the Russian WordNet on the basis of the data of other Russian thesaurus, namely the RuThes-lite, which is the smaller version of the RuThes and has been already published (http://www.labinform.ru/ruthes/index.htm). It contains around 100 thousand words and expressions which have been transformed into more than 40 thousand synsets connected by 50 thousand meronymy and hyponymy relations in the resource obtained.

The emergence of a large Russian WordNet will allow to do the following: to solve the problem of the Russian WordNet absence, to facilitate links with other wordnets created for other languages, to compare thesauri with different structures and to choose of an appropriate thesaurus for a specific application.

II. The RuThes structure

The RuThes thesaurus is a linguistically motivated ontology, based on the denotational part of lexical senses and conceptbased relations, which is a hierarchical network of concepts connected with text entries. The WordNet is a hierarchical network too, but the main element of it is a synset – set of synonyms. The WordNet consists of four sub-networks for each of the parts of speech with various relations, including relations between sub-networks. The RuThes that was created especially as resource for automatic text processing, unlike the WordNet was created as the model of human memory (separated definition of POSes, special types of relations, etc.)[1].

The RuThes-lite distributed by four files in xml format: a file contains the concepts, a file contains relations between the concepts, a file contains the the text entries and a file contains relations between the concepts and the text entries.

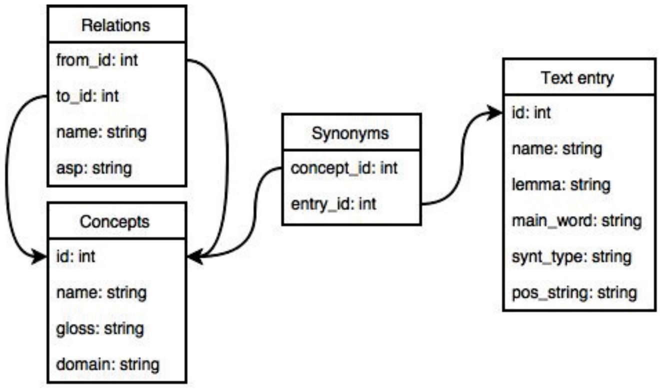

Fig. 1. The RuThes structure

The file of the concepts contains entries with the following information:

- an identifier (`id`);
- a name (`name`). Names are unique;
- an interpretation (`gloss`). The interpretations was extracted from Wikidictionary automatically with hand corrections afterwards. Only third part of concepts has interpretations;

ISBN 978-952-68397-0-7 (paperback), ISBN 978-952-68397-1-4 (PDF)

- a subject area (domain). In the current version of the thesaurus there are three types of the domains: GL (general lexicon), GEO (geographical objects) and SOC-POL (social and political concepts of modern society).

In the file of the relations between the concepts each row contains identifiers of two concepts and a name of the relation. The relations are specified from concept "from" to concept "to". For each relation the opposite relation is specified.

There are four kinds of relations:

- hyponymy (IS-A) relations;
- meronymy relations;
- non-symmetric association relations. (If pointed that there is such relation from A to B, then it means that existence of the concept A is depends to existence of the concept B. For example, existence of the concept "CAR FACTORY" depends to existence of the concept "CAR".);
- symmetric association relations.

The file of the text entries contains list of entries with the following information:

- the identifier (entry_id);
- the dictionary form (name);
- the lemmatic form, i.e. composed from dictionary word forms (lemma);
- the main word (main_word);
- the syntactic type (synt_type);
- the list of parts of speech for each word contained in the text entry (pos_string).

In the file of the relations between the concepts and the text entries each row specifies relation between identifier of concept (concept_id) and identifier of text entry (entry_id). Many text entries may be related to one concept (synonyms), one text entry may be related to many concepts (lexical ambiguity). Concepts not separated by parts of speech, therefore words of different parts of speech and many collocations may be related to one concept.

Current version of the RuThes-lite contains various types of text entries described in Table I.

III. The RuThes to the WordNet converting

The main idea of converting concepts and relations of the thesaurus RuThes to the WordNet-like format is to do the following steps: to separate the text entries by parts of speech, to unite them in synsets by relations within the concepts, and to repair relations between the synsets using relations between the concepts.

To uniting text entries in synsets we used the information stored in the "syn_type" fields of the text entries file. For now we picked out the text entries for three parts of speech: noun (syntactic type is "N", "NG" or "NGprep"), verb (syntactic type is "V", "VG" or "VGprep") and adjective (syntactic type is "Adj", "AdjG" or "AdjGprep"). In case of the RuThes structure one or more synsets of different parts of speech from each concept may be extracted. The relations between the concepts are used to establishing relations between the synsets. If in the chain of the concepts (A → B → C) for some concept (B) not exists synset of some part of speech, then the relation establishing across it (A → C).

TABLE I. Syntactic types of text entries in the RuThes-lite

Text entry syntactic type	Designation	Count of entries
noun group	NG	33252
noun	N	24265
verb	V	12368
adjective	Adj	12317
verb group	VG	9979
not specified		986
wrong value (may be repaired)	10 or 20	558
adjective group	AdjG	412
verb group with preposition	VGprep	321
preposition group	PrepG	300
adverb group	AdvG	265
noun group with preposition	NGprep	112
adverb	Adv	78
adjective group with preposition	AdjGprep	24
preposition	Prep	11
misc	Misc	9
pronoun	Pron	4
particle	Prtc	2

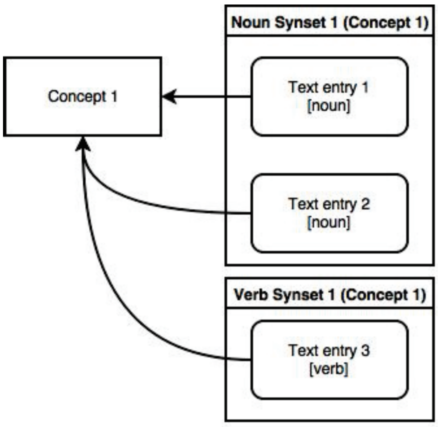

Fig. 2. Synset extracting method

IV. Used tools

For convenience of manipulating data of the the RuThes the special program (https://github.com/Zebradil/RuWordNet) was written. It contains two utilities. The first imports data from xml files to relational database (we choose PostgreSQL for this purpose). Some inconsistent relations and wrong values of important fields were found in the RuThes-lite, so we did some SQL-patches to correct that. For example, there was mixed up values of syn_type and pos_string fields in the text entries file.

There is a special tool named "grind" which generates WordNet database files (wndb-files). On input it uses human-readable source files written by lexicographers (lex-files). To generate those files second utility was used. It exports data from relational database to lex-files. It does main part of

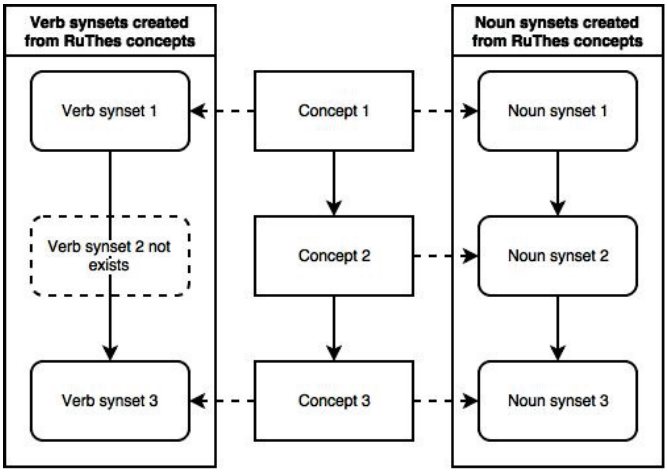

Fig. 3. Method of repairing relations

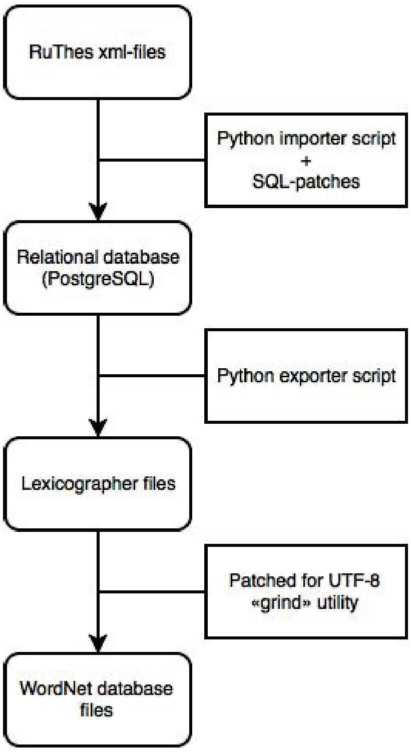
RuThes xml-files
Python importer script + SQL-patches
Relational database (PostgreSQL)
Python exporter script
Lexicographer files
Patched for UTF-8 «grind» utility
WordNet database files

Fig. 4. Flow of the data converting

work - extracts synsets, repairs relations, cleans up gloss from unsupported chars and other. That program implements algorithm of converting data structure and relations.

To generate wndb-files from cyrillic lex-files we use version (WNgrind-3.0-FiWN-20111201) of ”grind” modified to support UTF-8 encoding which required for cyrillic alphabet (http://www.ling.helsinki.fi/en/lt/research/finnwordnet).

V. Conclusion and Future Work

For now the generated RuWordNet contains hyponymy and meronymy relations for three parts of speech: nouns, verb and adjectives. It contains almost 42 thousand synsets. The wndb-files of the RuWordNet may be used with tools developed for the Princeton WordNet. In the future we plan to establish cross-POS relations, to compare thesauri with different structures, to choose an appropriate thesaurus for a specific application, and to publish the RuWordNet in the internet. There is a new utility to converting lex-files to wndb-files written in Java and integrated with the extjwnl library. It named jGrind and it is in development for now.

TABLE II. Counts of synsets of different parts of speech

Part of speech	Count of synsets
nouns	24279
verbs	6696
adjectives	10965
total	41940

VI. Acknowledgements

This work is partially supported by Russian Foundation for Humanities grant 15-04-12017.

References

[1] Fellbaum, C. (1999). WordNet. Blackwell Publishing Ltd.

[2] Irina Azarowa. 2008. RussNet as a Computer Lexicon for Russian. In Proceedings of the Intelligent Information systems IIS-2008: 341-350.

[3] Valentina Balkova, Andrey Suhonogov, and Sergey Yablonsky. 2008. Some Issues in the Construction of a Russian WordNet Grid. In Proceedings of the Forth International WordNet Conference, Szeged, Hungary:44-55.

[4] Ilia Gelfenbeyn, Artem Goncharuk, Vlad Lehelt, Anton Lipatov, and Viktor Shilo. 2003. Automatic translation of WordNet semantic network to Russian language. In Proceedings of International Conference on Computational Linguistics and In-tellectual Technologies Dialog-2003.

[5] Braslavski P.I., Mukhin M.Y., Lyashevskaya O.N., Bonch-Osmolovskaya A.A., Krizhanovsky A.A., Egorov P. (2013), Yarn Begins.

[6] Loukachevitch, N. and Dobrov, B. 2014. RuThes Linguistic Ontology vs. Russian Wordnets. In Proceedings of Global WordNet Conference GWC-2014, Tartu