Verification of lexic ontologies by the method of using semantic proximity of words calculated by large corpus of texts

V R Bayrasheva¹, N V Loukachevitch²

¹Kazan Federal University, Kazan, Russia
²Moscow State University, Moscow, Russian

¹vbayrasheva@gmail.com ²louk_nat@mail.ru

Abstract. The paper analyzes the concepts and relations between them in the RuWordNet digital thesaurus of the Russian language. Semantically related words automatically extracted from the corpus of texts are compared with the words of the thesaurus. The characteristic cases of information gap (omission of word meanings and semantic relationships) in the thesaurus are revealed, their causes of a computer and cognitive nature are described. Recommendations on improving the RuWordNet thesaurus are given.

1. Introduction

Recently, ontologies and thesauruses are being increasingly applied in a number of tasks of natural language processing. For the English language, the WordNet lexical ontology created more than 30 years ago is well known. Several attempts have been made to create a similar lexical resource in the Russian language [1-3]. Currently, only one version is being actively developed and used - the RuWordNet thesaurus [4]. High-quality language resources of this kind are created only manually, and this work is extremely labor-intensive. Therefore, the availability of at least semi-automatic means of their development and verification would be highly appreciated.

This study applies the approach proposed in the paper [5]. Essentially, it involves finding words by applying modern methods of vector representation for semantics, such as word2vec, that have close distribution each other in a large representational corpus. According to the well-known distribution hypothesis [6], such words should be semantically related. It is analyzed whether these words are already present in RuWordNet, and how they close are each other.

We continue the study commenced in the work [5]. However, instead of the mass data analysis carried out therein, this paper makes a thorough analysis of a limited number of examples, serving the basis for identifying the gaps in RuWordNet (omission of words, semantic relations), analyzing the causes of the gaps occurrence; as well offering general recommendations for detecting errors and identifying weak points in RuWordNet and potential sources of errors. This approach is applicable to other lexical ontologies and thesauri.

2. Methods and Data

The RuWordNet thesaurus contains more than 110 thousand words and phrases of the Russian language, united in synonymic sets and abbreviated as synsets. The following semantic relations were established between synsets: hypo-hypernymy, relation to the domain, the whole – the part (for nouns), antonymy (for adjectives), causative-consecutive relations (for verbs). In addition, synsets of various parts of
speech are interconnected by the part-of-speech synonymy relation, for example, стол — столовый (table, n – table, adj). The relationship with their constituent components is indicated for phrases, derivational relations are established between the root words.

The newsletter corpus constituting 7.5 million articles was used for clarifying the content and structure of the thesaurus. Vector representations were calculated for 40 thousand text inputs (words and phrases) from RuWordNet occurring in the assembled corpus for at least 100 times. 200 words were selected, their closest 20 words in the corpus being located in RuWordNet at a distance of more than 4 steps in semantic relations. The proximity of words in the corpus was calculated using the standard word2vec model [7]. Thus, for each of the 200 words, all 20 words corresponding to them seem to be semantically related in the corpus, but semantically remote in the thesaurus. This may, although not necessarily, be an indication of errors in the thesaurus. Let us name the selected 200 words targeted.

It appears, that in many cases the algorithm finds words in the corpus that have, by some chance, a similar distribution with the target word, though being not semantically related to it. This is especially true for rare words. For example, for the word чаща (thicket), the algorithm produces the following 20 words: частый (frequent); редкий (rare); худой (thin); предпочитать (to prefer); простой (plain, adj); быстрый (fast); медленный (slow); давление (pressure); остерегаться (to beware); запойный (hard drinking); малозначительный (insignificant); прозаичный (unpoetic); маленький (little); стыдливый (shamefaced); яствственный (explicit); клясть (to curse); попеременно (alternately); легкий (easy); гиповитаминоз (hypovitaminosis); соблазнять (to tempt). Only the first of them have semantic links with the target word. The case with the word чаща (thicket) is also explained by the morphological ambiguity in the wordform “чаще” (“more often”) (derivative of чаща (thicket) and частый (frequent)).

It turns out that the lemma (чаща) simply looks like the closest words of another lemma (частый). The automatic division of different meanings of a word is a complex, still not completely resolved problem that is beyond the scope of the present study.

Here’s another example of words that are related in respect the corpus but not in the thesaurus. The word херес (sherry) according to the corpus is close to the words Валенсия (Valencia), Барселона (Barcelona) and other cities of Spain. This is natural, because sherry is produced in Spain. However, it seems, that here we are dealing with the encyclopaedic information (the place of a particular wine production), not supposed to be present in a thesaurus.

Therefore, we select only some words from the output of the algorithm, which intuitively have a semantic link with the target word, and analyse the causes for the absence of this link in RuWordNet.

There are two main causes for the lack of small distance between the words in RuWordNet: omission of word meanings and omission of semantic relations.

3. Results
Let us consider in turn these types of discrepancy causes, analysing the data for the target words.

3.1. Omission of word meanings.
Однобокий (one-sided). According to the corpus, the word однобокий is related to необъективный (biased). RuWordNet has now two meanings for однобокий. The first of them has the hypernym of a synset {ограниченный, узкий} (limited, narrow), and this in its turn, has the hypernym of маленький (little). The second meaning has the hypernym of асимметричный (asymmetric), and the latter has a hypernym of расположенный (located).

Moving along these links, we can see, that we are in the plane of physical (spatial) relations; at the same time, the word однобокий is often used in the mental sphere: однобокий взгляд (one-sided view), однобокий подход (one-sided approach). Therefore, it seems appropriate to highlight the third meaning with a hypernym of необъективный (biased). Currently, in RuWordNet, the word необъективный has an only hyponym – ангажированный (engaged). With the proposed addition between the considered words in RuWordNet, the shortest one step path appears between однобокий and необъективный in terms of hypo-hypernymic relationship.
Придыхание (aspiration). According to the corpus, придыхание is close to the words благоговение (awe), трепет (trepidation). In RuWordNet, the word придыхание has the only meaning with the hypernym призыв (glide sound) which refers придыхание to the acoustic sphere indicated. At the same time, придыхание also has a different meaning, in which it is often used, and indicated, for example, in Wiktionary: “священный трепет, раболепие” (sacred awe, servility). For example: “А это частенько случается за кулисами Большого театра, на этой ярмарке тщеславия, когда с придыханием лицемерно скажут: «Ты — гений, старик!»” [8] (And this often happens behind the scenes of the Bolshoi Theatre, at this vanity fair, when, breathily (с придыханием), they will hypocritically say: “You are a genius, old man!”) [8].

It makes sense to add this meaning, indicating the emotional sphere as the domain and поклонение (reverence) as hypernym. Поклонение (reverence) already has such words as поклонение (worship) and почетное (esteem) as the hyponyms.

Прошивка (firmware). According to the corpus, this word is close to the words андэйт (update); браузер (browser); утилита (utility), субд (database system) and other computer terms. RuWordNet offers a single meaning: “seaming”. It is proposed to add a second meaning - “contents of computer memory” with a hypernym of scientific software.

At present, this phrase has hyponyms: браузер (browser), субд (database system). Obviously, the recently appeared highly specialized meaning of the word – прошивка (firmware) – is missing. Currently, this meaning is already present in Wikipedia, Wiktionary. A short path in two steps is going to appear upon adding the proposed meaning between the words прошивка and браузер.

3.2. Omission of semantic relations

Целлюлозный (cellulose). According to the corpus, this word is close to кожевенный (leather, adj.). In RuWordNet, the word целлюлозный is considered only in the context of the material – it has the only hypernym – полимерный (polymer, adj). At the same time, the word кожевенный has the hypernym промышленный (industrial). The word целлюлозный naturally belongs with the words завод (factory), промышленность (industry). It is proposed to also add the hypernym промышленный, after which the distance according to RuWordNet between целлюлозный и кожевенный will be equal to 2.

Перерегистрировать (to re-register). According to the corpus, перерегистрировать is related to переоформить (to re-issue). For переоформить, a hypernym sunset {пере делать, сделать заново} – {redo 2 (redo, remake)} is proposed. Hyponyms of this word are переоформить (renew) and переоборудовать (re-equip), перегруппировать (regroup), etc.

At the same time, a number of other words with the prefix пере- (re-) have two hypernyms. For example, for перегруппировать (regroup), it shall be группировать (to group) and переделать (to redo).

General recommendation: look through the words prefixed with пере- for the appropriateness of adding a new link as a hyponym to переделать.

Скалака (skipping rope). According to the corpus, the word скалака is related to words обруч (hoop), батут (trampoline). In RuWordNet, скалака has only the word шнур (cord) for a hypernym. It is proposed to indicate another hypernym - спортивный инвентарь (sports equipment). Then there will be a close link with the words батут, гимнастический (gymnastic) and обруч (via гимнастический снаряд (gymnastic apparatus)). In the current version of RuWordNet, only the physical appearance of the skipping rope is taken into account, but not the function of the object.

Сайга (саида) (saiga antelope). According to the corpus, the word сайга is close to косуля (roe deer), вепрь (wild boar). In RuWordNet, косуля и вепрь have дикое животное (wild animal) as hypernym. At the same time, сайга has only антилопа (antelope) as hypernym, and further on, according to biological systematics, this goes back to mammals. It is logical to add the hypernymic relationship of сайга – дикое животное. As a matter of fact, the list of wild animal in RuWordNet is far from complete. It makes sense to systematically review such general categories, which are usually one level higher than the base, and replenish them.
Сапфировый. (sapphire, adj). According to the corpus, this word is close to перламутр (mother of pearl). Mother of pearl in RuWordNet is classified only as a substance. However, by the nature of its use and physical properties, it should also be classified as a gemstone (see, for example, [9]). In this case, a short path appears between the words in question: сапфировый – сапфир – драгоценный камень – перламутр (sapphire, adj – sapphire, n – gemstone – mother of pearl).

4. Conclusion
In this article, we have analysed several particular examples of the discrepancy in word affinity in the RuWordNet thesaurus and in the corpus of news articles. The method of comparing the affinity of words in the corpus with affinity in the thesaurus proved to be very effective, as was demonstrated in [5] of the references and confirmed in the present paper. An alternative to this semi-automatic method may be a fully manual expert analysis of the structure of the thesaurus with a comparison of its data with the data of synonyms dictionaries, which was done in [10]. However, the latter method is a very time-consuming one and not free from the influence of errors in the synonyms dictionaries, and in addition, it does not tell anything about the correctness of semantic relations in the thesaurus.

The analysis of a series of case studies made it possible to identify some special aspects of the thesaurus and formulate recommendations for its improvement not only concerning these case studies, but also in a wider class of cases. The most typical case is the omission of some meaning in a polysemantic word. Such case may occur in explanatory dictionaries or using special programs of new meanings of words identification in the corpus of texts [11]. Another typical case is the omission of relations, in particular, hypo-hypernymic ones. It is possible to systematically check words of a certain structure, for example, with the prefix пере- (re-), that tend to have several hypernyms.

Acknowledgements
This research was financially supported by RFBR, grants N 18-00-01238, 18-00-01226, as part of the grant N 18-00-01240 (K).

References
[1] Braslavski P, Ustalov D, Mukhin M and Kiselev Y 2016 Proc. 8th Global Wordnet Conf. (Bucharest, Romania) pp 58–65
[2] Balkova V, Sukhonogov A and Yablonsky S 2004 Proc. 2nd Int. WordNet Conf. (Masaryk University, Brno) pp 31–8
[3] Gelfenbeyn I, Goncharuk A, Lekhelt V, et al. 2003 Proc. Conf. on Computational linguistics and Intellectual Technologies Dialogue-2003 (Moscow) pp 193–8
[4] RuWordNet 2020 https://ruwordnet.ru/ru/
[5] Loukachevitch N 2019 Proc. 57th Annual Meeting of the Association for Computational Linguistics pp 5773–9
[6] Sahlgren M Rivista di Linguistica 2008 20(1) pp 33–53
[7] Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg and Dean Jeffrey 2013 Preprint arXiv:1310.4546
[8] Magomaev M 1999 My love is a melody quote from the Russian National Corpus
[9] Finesell 2020 https://finesell.ru/vse-kamni/perlamutr.html.
[10] Soloviev V, Gimaletdinova G, Khalitova L and Usmanova L 2020 Communications in Computer and Information Science 1086 pp 174–83
[11] Kutuzov A, Øvrelid L, Szymanski T and Velldal E 2018 Proc. 27th Int. Conf. on Computational Linguistics (Santa Fe, USA) pp 1384–97