Reconstruction of Implied Semantic Relations in Russian Wiktionary

Serge Klimenkov, Evgenij Tsopa, Alexey Pismak and Alexander Yarkeev
Computer Science Department, ITMO University, Saint Petersburg, Russia

Keywords: Semantic Analysis, Semantic Network, Semantic Web, Natural Language Processing, Wiktionary, Russian Language.

Abstract: There were several attempts to retrieve semantic relations from free, online Wiktionary for Russian language. Previous works combine automatic parsing of wiki snapshot with experts' assistance. Our main goal is to create machine readable lexical ontology from Russian Wiktionary, maximally close to its online state. This article provides approach to automatic creation of explicit and implicit semantic relations between words (lexemes) and meanings (senses) to provide exact relations from sense to sense. Explicit semantic relations are constructed comparatively easy. For example, if the lexeme contains single sense, then all relations that point to the lexeme will point to this single sense. Reconstruction of implicit relations relies on logical conclusions from already created explicit ones. Several algorithms for implicit semantic links were developed and tested on Russian Wiktionary. There were parsed more than 550000 online pages, containing about 250000 Russian lexemes with about 500000 senses in them, but only about 20% of these senses were linked with at least one external lexeme. About 47% of explicitly existing links were resolved as “sense-to-sense” relations and about 28% of new implicit “sense-to-sense” links were reconstructed. 53% of lexemes’ references could not be resolved to exact sense.

1 INTRODUCTION

One of the most important parts in modern semantic computer applications is the lexical ontologies or thesauri, created by professional linguists. The best scientific resource currently known is WordNet (Miller and George, 1995). There were several attempts for Russian language to create similar resources, such as RussNet (Azarowa, 2008), RuThes (Loukachevitch and Dobrov, 2014), automated WordNet translation (Balkova, Suhonogov and Yablonsky, 2008). Important disadvantage of manual or semi-manual ontologies creation is that natural languages are continuously changing. YARN (Braslavski, Ustalov and Mukhin, 2014) is important attempt to resolve this issue via platform that supports crowdsourced articles creation. Authors took raw data from Russian Wiktionary and Small Academic Dictionary and have provided applications that simplify and distribute routine tasks for making WordNet-like synsets. YARN uses Wikokit (Krizhanovsky and Smirnov, 2013) to parse Russian and English Wiktionary dump for extracting information and convert data to machine readable format (Bessmertny, 2010). This approach also has disadvantages: available dumps are rarely taken from online dictionary and there are limited numbers of volunteers that execute actual information processing (most of them are professional linguists and students of chairs linguistics).

In contrast, Russian Wiktionary is popular free web resource in Slavic speaking countries and is being changed every day by thousand enthusiasts that follow the lexicon changes. Of course, Russian Wiktionary is popular but not well structured source of the information (Klimenkov, Tsopa, Kharitonova and Pismak, 2016). We can successfully find semantic information about the word, represented by lexeme, its meanings (senses) and various lexicographic information created by experts. A set of methods intended for semantic references creation for lexicographic dictionary was proposed in (Wandmacher, Ovchinnikova and Krumnack, 2007) including German Wiktionary analysis. It is really important for context- and knowledge-aware applications to have modern and consistent lexical ontology with actualized semantic relations in it.

This article describes an approach to the creation of automatically retrieved, up-to date clone

Klimenkov, S., Tsopa, E., Pismak, A. and Yarkeev, A.
Reconstruction of Implied Semantic Relations in Russian Wiktionary.
DOI: 10.5220/0006038900740080
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 2: KEOD*, pages 74-80
ISBN: 978-989-758-203-5

of Russian Wiktionary, converted to machine readable format, with most important lexicographic information for Russian words. We use our developed algorithms to retrieve explicit and implicit semantic relations.

The remainder of this paper is organized as follows. Section 2 notes important difference between article structure in English and Russian Wiktionary. Section 3 presents rules and explains reasons that were used for references reconstruction in developed algorithms. Software architecture for online references extractions and extractions results were described in section 4. There are brief conclusion and future research directions in section 5.

2 WIKTIONARY ARTICLE AND SEMANTIC RELATIONS

Wiktionary article describes a lexeme which consists of one or more senses. Each sense has a description, translation links and set of semantic options. Semantic option is the link to other lexeme labelled with option type that represents one of well defined basic semantic relations. There are 6 types of semantic options defined for Russian Wiktionary:

- synonym;
- antonym;
- hyponym;
- hypernym;
- holonym;
- meronym.

The problem is that relations do not link one of the word's sense definitions with other sense. Wiktionary semantic relations point to a whole lexeme. Moreover, article structure and relations creating rules depend on Wiktionary language section. For example, English Wiktionary article describes lexeme with a couple of senses and links from one lexeme to another lexeme (Fig.1).

In a contrast, Russian Wiktionary article (Fig.2) contains references from the sense to other lexeme.

Basic relations in semantic networks are always bidirectional. If the sense S_1 has basic semantic reference to the sense S_2, it implies that the sense S_2 has backward reference to the sense S_1. Backward references do not always exist in articles as it is not easy to find all related articles out. Therefore authors often forget to create backward references. Semantic relation R can also be symmetric or transitive. The backward reference of symmetric relations has the same type R as the direct one. Transitive reference labels the link back with complementary $\bar{R}$ relation type. Synonymy and antonymy are symmetric relations. Hypernymy – hyponymy and meronymy – holonymy are transitive.

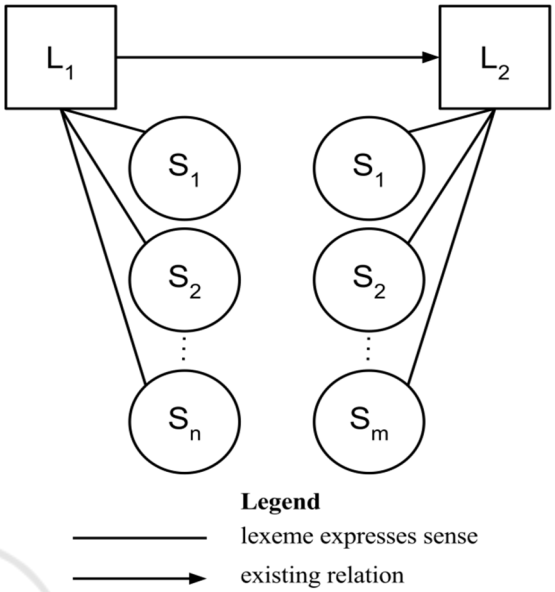

Figure 1: References between articles in English Wiktionary.

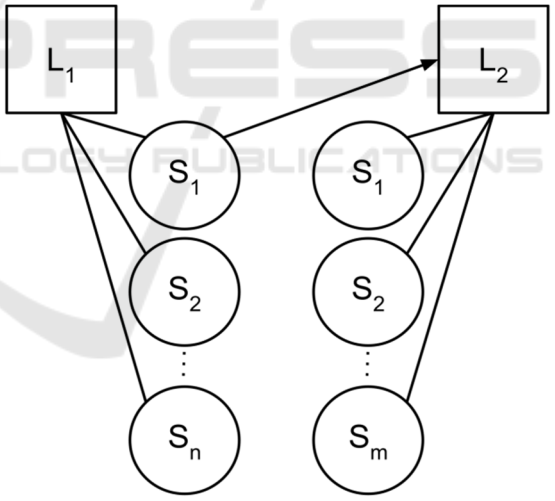

Figure 2: References between articles in Russian Wiktionary.

3 CREATING SEMANTIC RELATIONS

As mentioned above, Wiktionary article for the lexeme can contain one or more senses, and those senses point to a lexeme, not to a sense. We have defined following set of rules that can extend existing relations not explicitly defined in lexeme.

3.1 Lexeme with Single Sense

If the lexeme contains only one sense (Fig.3), semantic relation that points to the lexeme (L_1S_1 to L_2) may be replaced by the link to that single sense (L_1S_1 to L_2S_1). Russian Wiktionary has around 70% Cyrillic lexemes with single sense only. So it is really obvious to create such explicit semantic references. At this time we can reconstruct implicit backward relation from L_2S_1 to L_1S_1, which is absent in most cases.

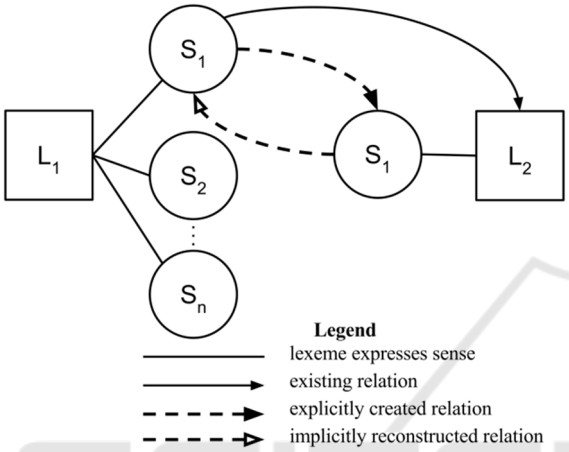

Figure 3: Lexeme with single sense references reconstruction.

For example, article “рубероид” (ruberoid, L_2) contains single sense definition “гидроизоляционный материал” (bituminous waterproofing, L_2S_1). Another article “стройматериал” (building material, L_1) contains multiple senses, one of them is “строительный материал” (material used for construction purposes, L_1S_1). Sense L_1S_1 have a hyponymy reference to L_2 article (ruberoid). As it contains only single sense, we can conclude that sense L_2S_1 (bituminous waterproofing) is a hyponym for sense L_1S_1 (material used for construction purposes). Additionally, since all semantic relations are bidirectional, we can reconstruct implicit hypernym backward reference between senses L_2S_1 (bituminous waterproofing) and L_1S_2 (material used for construction purposes).

3.2 Mutual Sense Reference

The same principle can also be used for relations reconstruction in case when two lexemes are mutually cross-referenced from its senses (Fig.4). In that case when the sense L_1S_1 references to the lexeme L_2 and the sense L_2S_1 references to the lexeme L_1 and references are the same (or complementary) type, we can reconstruct two explicit references from L_1S_1 to L_2S_1 and backward from L_2S_1 to L_1S_1. Original links to lexemes became redundant and must be removed.

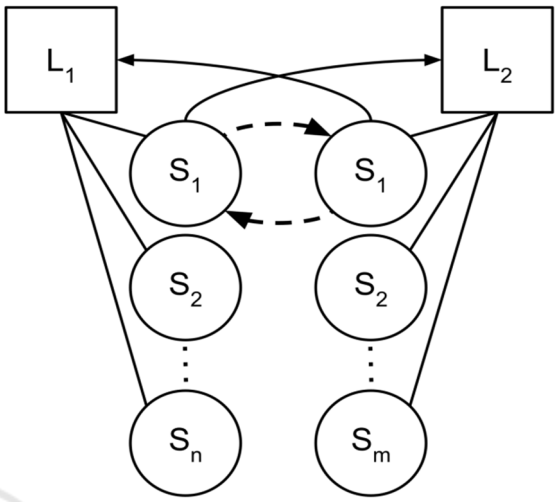

Figure 4: Mutual sense references reconstruction.

For example, Russian Wiktionary article “ребенок” (baby, L_1) contains sense with definition “человек до начала полового созревания” (a very young human, L_1S_1). Another article “старик” (old man, L_2) contains sense with definition “старый мужчина” (an elderly man, L_2S_1). So, because of sense L_1S_1 (a very young human) contains antonymy reference to L_2, (old man) and sense L_2S_1 (an elderly man) contains the same relation to L_1 (baby), senses L_1S_1 (a very young human) and L_2S_1 (an elderly man) are antonyms.

It is important to note that the only one reference from sense to lexeme must exist in both lexemes. In the case when we have two or more such references, we can not choose the sense that the reference should be created for (Fig.5). Developed algorithm checks and ignores these ambiguous references.

The article “дерево” (tree, L_1) may be used to illustrate this case. This article contains the sense “многолетнее, как правило, крупное растение” (a large plant typically over four meters in height, L_1S_1) with hyponymy relation to the lexeme “сосна” (pine, L_2). But the article “сосна” (pine) contains two senses “дерево из рода вечнозелёных голосеменных растений” (any coniferous tree of the genus Pinus, L_2S_1) and “древесина сосны” (the wood of pine tree, L_2S_2) with hypernymy references to the lexeme “дерево” (tree, L_1). So we cannot choose any of these senses for “sense-to-sense” reference reconstruction and the relation must be ignored by our algorithm.

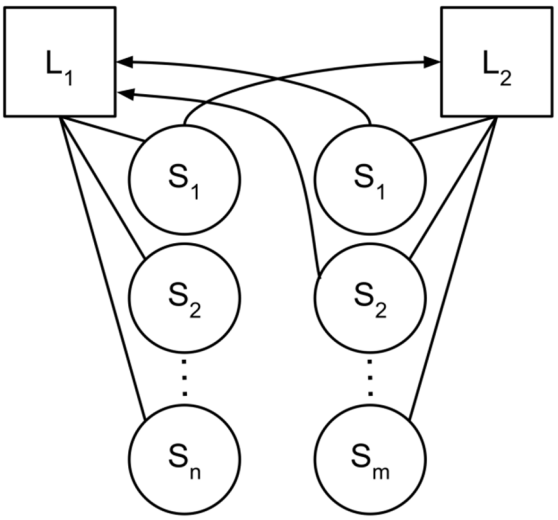

Figure 5: Ambiguous references in mutual sense references reconstruction.

3.3 References Reconstruction via Common Synonym

If senses are linked as synonyms we can mirror other already existing references between senses to those synonyms. More formally, we can conclude that senses L_1S_1 and L_2S_1 are linked by the bidirectional semantic relation of type R (Fig.6), if following criteria are met:

1. Sense L_1S_1 has semantic relation of type R to lexeme L_2.
2. Senses L_1S_1 and L_3S_1 are linked with synonymy relation.
3. Lexeme L_2 has sense L_2S_1 that is linked with L_3S_1 via semantic relation of the same type R.

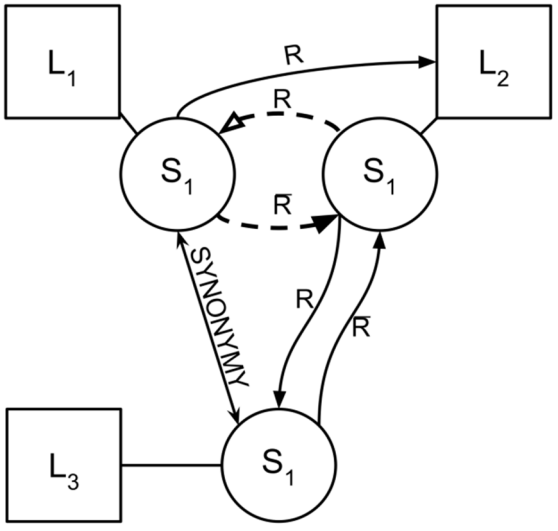

Figure 6: References reconstruction via common synonym.

Type R in this case can be symmetric or transitive. In case of transitive relation we must reconstruct complementary references for given type. That is, if R is the hyponymy than two complementary links, hyponymy R and hypernymy R have to be created. The first of reconstructed relations is always explicit (as there is already the link that points to the lexeme). Backward relation may be implicit if corresponding relation doesn't exist in Wiktionary.

This rule may be illustrated by the article "мишка" (a little bear, L_1) that contains the sense "медведь" (a bear, L_1S_1). This sense contains hypernym reference to the lexeme "животное" (an animal, L_2). Also the sense L_1S_1 (a little bear) contains synonymy reference to the sense "крупное мохнатое хищное млекопитающее" (big furry predatory mammal, L_3S_1) of the lexeme "медведь" (a bear, L_3). At the same time the sense L_3S_1 (big furry predatory mammal) has hypernym reference to the sense "представитель фауны" (any member of the kingdom Animalia, L_2S_1) of the lexeme L_2 (an animal). So it is possible to replace the "sense-to-lexeme" hypernym link from the sense L_1S_1 (a little bear) to the lexeme L_2 (an animal) by the "sense-to-sense" antonymy reference from the sense L_1S_1 (a little bear) to the sense L_2S_1 (any member of the kingdom Animalia). Antonymy is symmetric relation and we can reconstruct the implicit backward antonymy reference from the sense L_2S_1 (any member of the kingdom Animalia) to the sense L_1S_1 (a little bear).

3.4 Synonymy Reconstruction via Common Reference

Similar conclusions can be made when the pair of senses is connected to the third through identical references type. If one of the pair has the link to the lexeme of the second as synonym then we can create synonymy reference between that pair of senses. Formally, senses and links between them must satisfy these three criteria:

1. Sense L_1S_1 references as synonym of lexeme L_2.
2. Sense L_1S_1 is linked via semantic relation of type R with sense L_3S_1 of lexeme L_3.
3. Any sense of lexeme L_2 (for example, L_2S_1) is linked via the R-reference with sense L_3S_1.

When all these conditions are met, we can conclude that senses L_1S_1 and L_2S_1 are synonyms (Fig.7). At this time we can reconstruct implicit backward relation from L_2S_1 to L_1S_1.

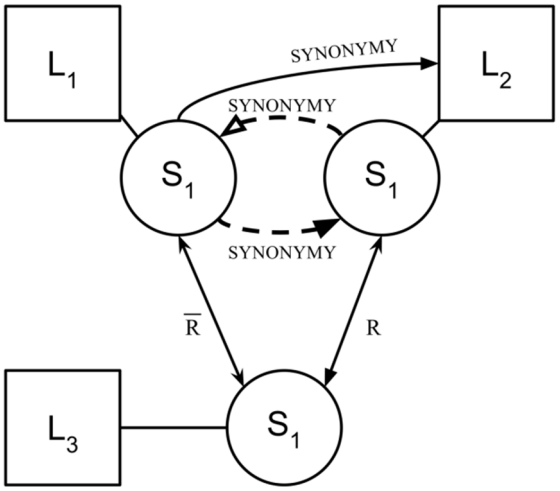

Figure 7: Synonyms reconstruction via common reference.

For example it could be illustrated by the article “нежный” (gentle, L_1) that contains the sense “хрупкий, уязвимый” (fragile, vulnerable, L_1S_1). This sense contains the synonymy reference to the lexeme “мягкий” (soft, L_2). In addition, the sense L_1S_1 (fragile, vulnerable) contains the antonymy reference to the sense “плохо поддающийся деформации или разделению” (poorly amenable to deformation or separation, L_3S_1) of the lexeme “жёсткий“ (hard, L_3) and the sense “легко поддающийся нажиму, деформации” (easily amenable to pressure and strain, L_2S_1) of the lexeme L_2 also have the antonymy reference to the same sense. So it is possible to replace the “sense-to-lexeme” synonymy link from the sense L_1S_1 (fragile, vulnerable) to the lexeme L_2 (soft) by the “sense-to-sense” synonymy reference from the sense L_1S_1 (fragile, vulnerable) to the sense L_2S_1 (easily amenable to pressure and strain).

4 SOFTWARE ARCHITECTURE AND ALGORITHM RESULTS

Developed algorithm was tested on online Russian Wiktionary. At the time of testing online resource contained 568910 pages. These pages contain in total 1285926 senses and 302358 references to other lexemes. There were found only 101993 Russian senses that have had at least one reference with another lexeme. As mentioned above, Russian Wiktionary contains lexemes from other languages (English, German, etc., created by robots) and they were omitted. Found Russian senses have 206994 links to other lexemes and 70% of 101993 senses (70202) are single sense lexemes, i.e. can be subject of rule, described in section 3.1.

Software was developed using Java programming language and Spring framework. Its architecture consists of three layers (Fig. 8).

First layer loads mark-up contents of Wiktionary articles using online REST API, parses it and converts to graph. This layer executes initial dictionary bulk data import and provides daily dictionary synchronization with online version to maintain created by crowd article in actual state.

Second layer is a software storage that is based on OrientDB (Tesoriero, 2013) DBMS. OrientDB is an open source database based on distributed graph engine. It provides support of HTTP REST and JSON APIs to properly represent and visualise deducted semantic relation in the browser-oriented application. Developed software use Gremlin API (based on SpringData) to provide connection to OrientDB. Lexemes, senses and semantic references implementation is based on directed multigraph (Harary, 1994) model.

Explicit and implicit semantic references are created on the third layer. It gets results of mark-up parsing, inserts it into graph model and sequentially applies algorithm's rules. Order of rules is not important. Rules based on synonymy (3.3 and 3.4) can be applied multiple times. Algorithm repeatedly applies these rules and stops if new references were not created.

During the first run the program downloads all existing pages in Russian Wiktionary for 30 hours. Than it applies developed rules in single thread of 3.3 GHz CPU for a little bit more than 4 hours.

First step of algorithm creates mutual sense references as described in section 3.2. We found 16% (32562) references between Russian lexemes’ senses. Next we run rule that finds single sense lexemes (3.1). It converts sense-to-lexeme to sense-to-sense references and creates 57814 explicit references. As mentioned above this rule also creates implicit (complement) references based on explicit one. In total the rule creates 115628 references that cover 56% of all Russian lexemes. Execution time of rule 3.2 was less then 15 second. The third step rule (3.3) reconstructs 5037 references between senses that is 2.4% of initial Russians senses’ links count. Next rule (3.4) adds 0.7% (1470) references more.

We applied rules form 3.3 an 3.4 again, and got 390+10 references per 1 and 30 seconds respectively. At the third iteration of rules 3.3 and 3.4 we got 28+4 references for approximately same time. Fourth iteration had no any new references.

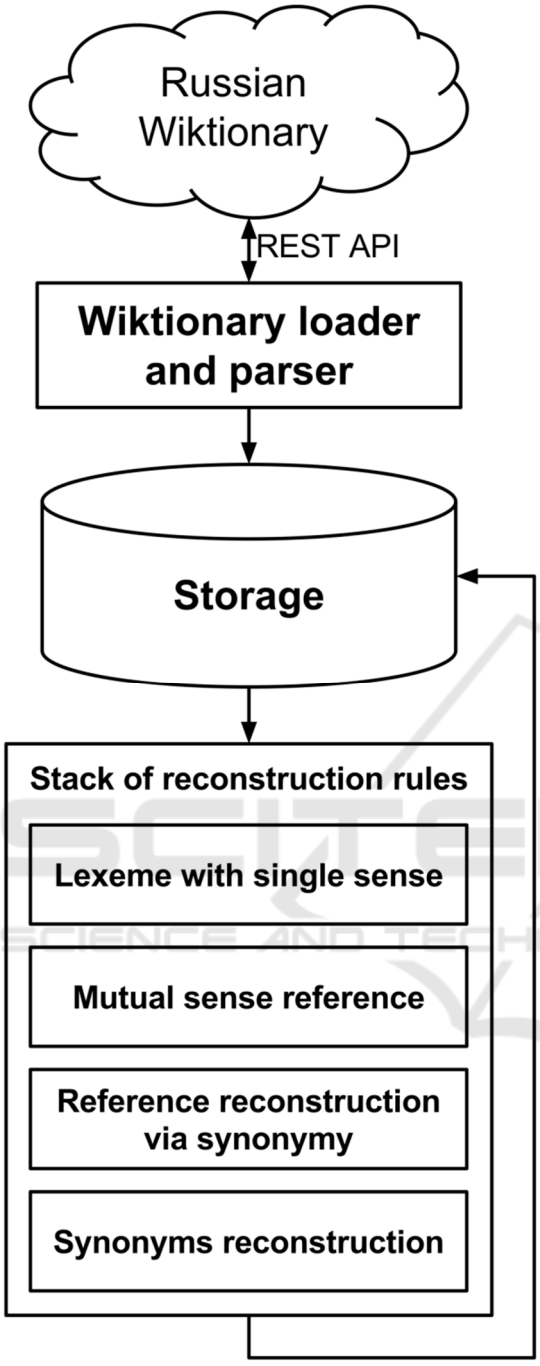

Figure 8: Developed software architecture.

Overall experiment results show that 97315 (about 47%) explicit relations from 206994 were reconstructed and additionally 57814 (about 28%) implicit relations were reconstructed. Unfortunately, developed algorithms could not convert 109679 (53%) sense-to-lexeme to sense-to-sense references.

5 CONCLUSIONS

Semantic references reconstruction is an important part of thesaurus creation process. Developed software and set of rules allow to get online Russian Wiktionary pages and to convert it into classes of direct graph model. More than 206000 semantic references were transferred to this model directly from articles and 47% of them were converted to point to the exact sense, and additional 28% were implicitly reconstructed using existing references.

Unfortunately, only several Slavic Wiktionaries have the same structure as Russian articles. Most of existing Wiktionaries are the same as English, i.e. contain only lexeme-to-lexeme references. Therefore described approach could not be used directly. In subsequent studies translation links from English article to Russian can be used to deduct semantic structure for English Wiktionary. Other future research direction is discovering implicit relations comparing sense description

As shown by recent studies (Meyer, C.M. and Gurevych, I., 2010. and Smirnov, Kruglov and other, 2012) freely available, wiktionary-style online dictionaries are continuously advancing and becoming more sophisticated. Many language features and linguistic information are already incorporated from existing human-created lexical ontologies into these dictionaries. If the quality of articles becomes more acceptable, approaches, similar to described, could convert more raw lexemes data in “sense-to-sense” relations. We hope that in the near future it would be enough information in Russian Wiktionary and additional algorithms will be invented to automatically reconstruct well-connected semantic network, which could be integrated in every application that need dictionary with semantically related features.

REFERENCES

Bessmertny I., 2010. Knowledge visualization based on semantic networks. Programming and Computer Software.

Miller, George A., 1995. WordNet: a lexical database for English. Communications of the ACM.

Azarowa I., 2008. RussNet as a computer lexicon for Russian. Proceedings of the Intelligent Information systems IIS-2008.

Loukachevitch N. and Dobrov B., 2014. RuThes linguistic ontology vs. Russian wordnets. Proceedings of Global WordNet Conference GWC-2014.

Balkova V., Suhonogov A. and Yablonsky S., 2008. Some issues in the construction of a Russian wordnet grid.

Proceedings of the Forth International WordNet Conference, Szeged.

Braslavski P., Ustalov D., Mukhin M., 2014. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics.

Krizhanovsky A.A. and A. V. Smirnov, (2013). An approach to automated construction of a general-purpose lexical ontology based on Wiktionary. Journal of Computer and Systems Sciences International. pp. 215-225.

Klimenkov S.V., Tsopa E.A., Kharitonova A.E. and Pismak A.E., (2016). Method of automatic generation of semantic network from semi-structured sources. Software products and system. pp. 40

Wandmacher T., Ovchinnikova E. and Krumnack U., (2007). Extraction, Evaluation and Integration of Lexical Semantic Relations for the Automated Construction of a Lexical Ontology. in Third Australasian Ontology Workshop (AOW). pp. 61–69.

Tesoriero C., (2013). Getting Started with OrientDB. Packt Publishing Ltd.

Meyer C.M. and Gurevych I., (2010). How web communities analyze human language: Word senses in wiktionary.

Harary F. (1994) Graph Theory. Reading. Addison-Wesley, p. 10.

Smirnov A.V.T., Kruglov, V.M., Krizhanovsky, A.A., Lugovaya, N.B., Karpov, A.A. and Kipyatkova, I.S., (2012). A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries. Trudy SPIIRAN. pp.231-253.