Method of Reconstruction of Semantic Relations using Translingual Information

Viktor Osika, Sergey Klimenkov, Evgenij Tsopa, Alexey Pismak,
Vladimir Nikolaev 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: The article is devoted to the problem of developing a method of restoring semantic relations using translingual information. Wiktionary article may contain a translation section (a translingual section) – an important element that allows to link a sense from one language to a sense from another language that is expressed with a reference to the lexeme of that sense. One of our tasks is the design and implementation of Wiktionary-based ontology for the semantic analysis. In this article we present a set of rules that can be used to establish the coincidence of the same senses in different languages, and consequently map links restored from Russian Wiktionary to English nodes. The algorithm for restoring inter-sense references includes the selection of candidate links across senses of different language sections and set of rules to accept the candidate to be included in the list of senses' links. As a result, 69,309 potential links-transfers (excluding duplicated links) were selected. More than 16,000 links between the nodes of semantic senses of the Russian and English sections of Wiktionary were confirmed that allowed the creation of a generalized ontology.

1 INTRODUCTION

The ability of applications to perform semantic analysis of natural language texts becomes especially important in modern information systems that process unstructured and semi-structured data. One of the most important parts of such systems is the graph of meanings or concepts, related to each other by semantic links. In the field of information processing such graph is called ontology or thesaurus. The most famous thesaurus for the English language is WordNet(Miller, George., 1995), which has been created by professional linguists. For the Russian language several similar thesauri have been constructed, including the development of St. Petersburg University RussNet (Azarowa, 2008) and RuThes (Loukachevitch, and Dobrov, 2014), which is a part of the commercially used ontology converted to WordNet. In addition, a partial semantic translation of WordNet into Russian has been made in (Balkova, Suhonogov, Yablonsky, 2008).

An important feature of the lexicon is its variability. Although the thesauri described above represent the core of concepts, they can not take into account continuous changes that occur with language. One of authors' tasks is the development and usage in the semantic analysis of the ontology, based on Russian Wiktionary (Klimenkov, Tsopa, Pismak., Yarkeev., 2016), (Pismak, Kharitonova, Tsopa, Klimenkov, 2016). Russian Wiktionary is a platform for thousands of enthusiasts that fill it with new information and change the dictionary every day. It should be noted that Wiktionary contains all the information that is necessary for the computer ontology construction: lexemes, word forms based on the classification of A.A. Zaliznyak (Zaliznyak, 2008), important morphological and lexicographic information that includes meanings (senses) of words and their semantic links. In their previous works authors transformed Wiktionary into a machine-readable structure, organized its automatic synchronization with the online resource of Wiktionary, and restored explicit and implicit semantic links between the senses of words (Klimenkov, Tsopa, Pismak, Yarkeev., 2016).

Osika V., Klimenkov S., Tsopa E., Pismak A., Nikolaev V. and Yarkeev A.
Method of Reconstruction of Semantic Relations using Translingual Information.
DOI: 10.5220/0006516602390245
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2017)*, pages 239-245
ISBN: 978-989-758-272-1

This study represents the evolution of the earlier described approach that reconstructs links using translingual information from the dictionary and verifies the correctness of created links with English and German versions of Wiktionary.

2 STRUCTURE OF THE ONTOLOGY AND REPRESENTATION OF SEMANTIC LINKS

The structure of articles in Wiktionary for Russian language (and also for some other Slavic languages) is fundamentally different from the majority of other Wiktionaries. In Russian Wiktionary lexeme can contain several senses that it expresses, and each of these senses may contain one or more semantic links to other lexemes. In English Wiktionary it is impossible to distinguish a specific meaning, the semantic link exists only between lexemes. This property of the Russian dictionary allows to create a semantic network using explicit and implicit structural dependencies.

In addition to the sense-to-lexeme links in Wiktionary, there is a translation section that links Russian Wiktionary to other languages. Such section is called translingual. Translations are important elements that allow to link a sense from one language to sense from another, expressed through the link to the lexeme of some sense. In this article we describe a set of rules that can be used to establish the coincidence of identical senses in different languages, and, therefore, map links reconstructed from Russian Wiktionary to senses of English Wiktionary.

To reconstruct semantic links it is required to convert the content of Wiktionary articles into a set of semantic nodes with a certain structure. Taking into account the feature of the conducted research, it is necessary to provide storage in the semantic graph of translational information.

The ontology was implemented in the form of Neo4j graph database (Van Bruggen, 2014). Its scheme is described by the UML-model shown in Diagram 1. Nodes of types Gloss, MorphItem and Lexeme are translingual.

3 ALGORITHM FOR LINKS RECONSTRUCTION

The block of translingual links in Wiktionary is allocated in a separate section. In such sections for Russian and some Slavic languages the correspondence of sense and its translations to other languages is explicit and shown through the successive numbering of senses (Fig. 1). In other

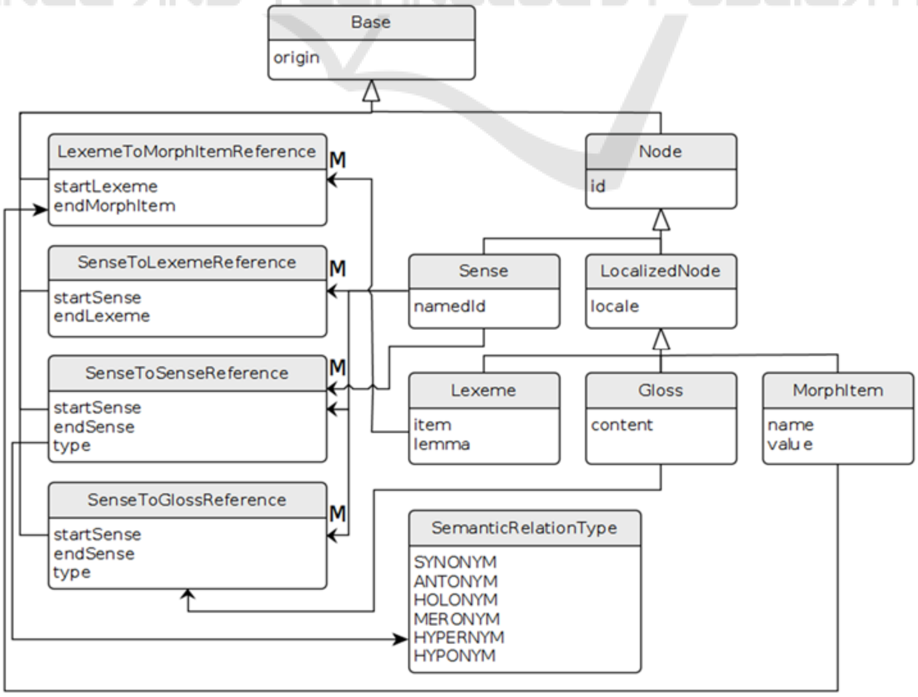

Diagram 1: The structure of nodes and links of the semantic graph.

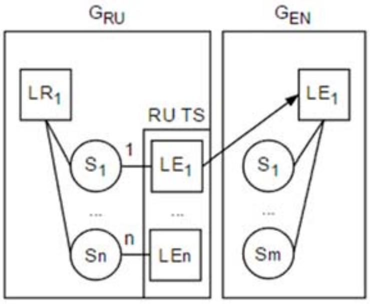

Figure 1: The structure of translingual link.

sections the correspondence of the sense to its translation can be determined probabilistically based on the method described in (Pismak., Kharitonova, Tsopa, Klimenkov, 2016). The translation block contains lexemes with links to lexemes from another language section of Wiktionary.

The developed method allows to establish links between senses of two (and more) language sections based on the developed set of rules. As initial data we will use the results of the study (Klimenkov, Tsopa, Pismak., Yarkeev., 2016), which contains the reconstruction of inter-sense links for the Russian section, and a dictionary for the third language. Based on the number of lexemes and senses, as well as the presence of advanced programming interface for data extraction (Miller, George, 1990), the German section was chosen.

The general algorithm for the reconstruction of inter-sense links in the developed method is as follows:

1. Candidate links between the senses of different language sections are chosen based on the developed rules;
2. The candidate-link is verified and included in the general list of links between the senses using another set of rules.

Below we present the rules for the search and reconstruction of the inter-sense links.

3.1 Bilateral Links

If there are symmetrical cross-links between lexemes and senses of two dictionaries, that is, the sense of Russian Wiktionary links to the lexeme of English Wiktionary, and one of the senses of English language, in turn, points back to the Russian lexeme that expresses this sense, then we can identify a potential candidate for the link reconstruction.

$$\forall \begin{cases} LR_1 \in G_{RU} \\ LE_1 \in G_{EN} \end{cases}. \quad (1)$$

$$\exists(LR_1S_1, LE_1) \& \exists(LE_1S_1, LR_1) \rightarrow (LR_1S_1, LE_1S_1) \quad (2)$$

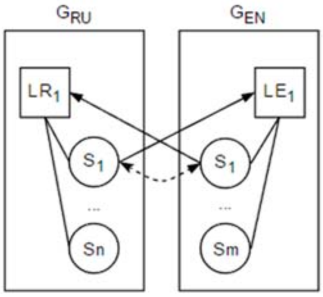

Figure 2: The usage of bilateral communication.

Example (Figure 2):

LR_1 (словарь) → LR_1S_1 (список, перечень, собрание языковых единиц одного уровня, обычно упорядоченных по алфавиту) → LE_1 (dictionary) → LE_1S_1 (A reference work with a list of words from one or more languages, normally ordered alphabetically) → LE_1 (словарь).

3.2 Lexemes with a Single Sense

If the sense of the lexeme of the Russian dictionary links to the lexeme of the English dictionary and this lexeme contains only one sense, then we can identify a potential candidate for the reconstruction of the link (Figure 3).

$$\forall \begin{cases} \mathrm{LR}_1 \in \mathrm{G}_{\mathrm{RU}} \\ \mathrm{LE}_1 \in \mathrm{G}_{\mathrm{EN}}. \\ |\mathrm{LE}_1| = 1 \end{cases} \quad (3)$$

$$\exists(LR_1S_1, LE_1) \rightarrow (LR_1S_1, LE_1S_1) \quad (4)$$

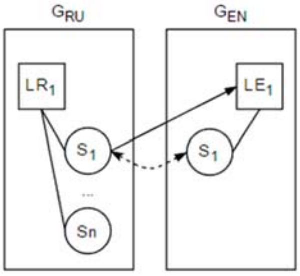

Figure 3: Lexeme with a single sense.

For example, if one of the senses of the Russian lexeme "шельф" links to the only sense of the English lexeme "shelf", these two senses are translations of each other: LR1 (шельф) → LR_1S_1 (прибрежная зона океана) → LE_1 (shelf) → LE_1S_1 (A reef, shoal or sandbar) → LR_1 (шельф).

It does not matter which of the lexemes (English or Russian) has a single sense - the developed principle is symmetric. For example, if one of the senses of the English lexeme "trivially" is linked with the single sense of the Russian lexeme "тривиально", these two meanings are translations of each other: LE_1 (trivially) → LE_1S_1 (In a trivial manner) → LR_1 (тривиально) → LR_1S1 (банально, неоригинально) → LE_1 (trivially).

3.3 The Usage of the Third Dictionary

For the search of potential links we can use the third language dictionary (Figure 4). The German section of Wiktionary has been used in the work.

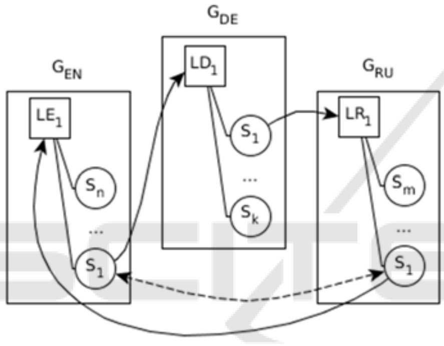

Figure 4: The usage of the third dictionary.

If the sense of the lexeme of the Russian dictionary has a link to the lexeme of the German dictionary, that in turn has the sense that contains a link to the lexeme of the English dictionary, and one of the senses of this lexeme links back to the Russian lexeme, then we can identify a potential candidate for the reconstruction of the link.

$$\forall \begin{cases} LR_1 \in G_{RU} \\ LE_1 \in G_{EN}. \\ LD_1 \in G_{DE} \end{cases} \quad (5)$$

$$\exists(LE_1S_1, LD_1) \& \exists(LD_1S_1, LR_1) \\ \& \exists(LR_1S_1, LE_1) \rightarrow (LR_1S_1, LE_1S_1) \quad (6)$$

That is, if the sense LR1S1 in one Wiktionary language section contains a translingual link to the lexeme LD_1 in another language section and the sense LD_1S_1 contains a translingual link to the lexeme LE_1 of the third language section, and the sense LE_1S_1 contains a translingual link to the lexeme LR_1, then you can reconstruct a biliteral link of the type "Translation" between the sense LR_1S_1 and the sense LE_1S_1.

Similarly to 3.2, this rule is also symmetric - the original lexeme can be found in Russian Wiktionary or in English Wiktionary:

LR_1 (смазать) → LR_1S_1 (нанести смазку) → LD_1 (schmieren) → LD_1S_1 (etwas mit etwas bestreichen) → LE_1 (oil) → LE_1S_1 (to lubricate with oil) → LR_1 (смазать).

LE_1 (owner) → LE_1S_1 (one who owns something) → LD_1 (Besitzer) → LD_1S_1 (die Person, die tatsächliche Herrschaft über eine Sache ausüben) → LR_1 (хозяйка) → LR_1S_1 (собственница) → LE_1 (owner).

3.4 Verification with the Usage of Linked Lexemes

After the analysis of the obtained links, we found that there are incorrect links in the list of candidates. To check and eliminate incorrect links we developed a set of correction rules (Figure 5).

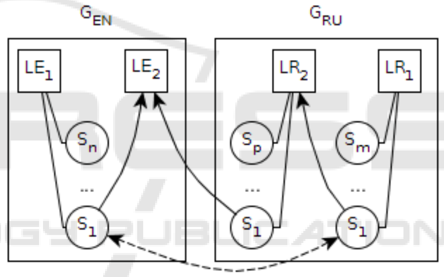

Figure 5: The usage of common linked lexemes.

If the sense-candidate of the Russian dictionary has one of the semantic links known in the dictionary (synonym, antonym, hyperonim, etc.) with a lexeme, and one of the senses of this lexeme has a link to the lexeme of the English dictionary, to which the sense-candidate from the same dictionary is connected with the same type of the link, then we can confirm candidates as a reconstructed link.

$$\forall \begin{cases} LR_1 \in G_{RU} \\ LE_1 \in G_{EN}. \\ LD_1 \in G_{DE} \end{cases} \quad (7)$$

$$\exists(LE_1S_1, LD_1) \& \exists(LD_1S_1, LR_1) \\ \rightarrow (LR_1S_1, LE_1S_1) \quad (8)$$

For example, if as a result of applying the proposed rules the link with the type “translation” has been restored between senses of lexemes "close" and "завершать", then we can verify its correctness in the following way. One of the senses of the English-speaking lexeme "close" is the antonym of the lexeme "begin", one of its senses, in turn, linked in Wiktionary by the link of type "translation" with the Russian lexeme “начинать”: LE_1 (close) → LE_1S_1 (To finish) → LE_2 (begin) → LE_2S_1 (To start) → LR2 (начинать).

Due to the fact that the lexeme "начинать" in Russian section of Wiktionary is an antonym of one of the senses of the original lexeme "завершать", we can conclude that the link was reconstructed correctly: LR_1 (завершать) → LR_1S_1 (оканчивать) → LR_2 (начинать).

3.5 Verification using Common Translations into a Third Language

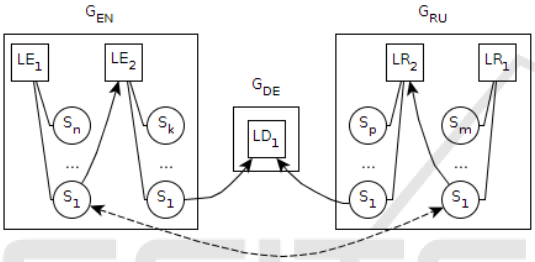

Figure 6: Verification of common linked lexemes that have a translation into a third language.

$$\forall \begin{cases} LR_1 \in G_{RU} \\ LR_2 \in G_{RU} \\ LE_1 \in G_{EN} \\ LE_2 \in G_{EN} \end{cases}. \qquad (9)$$

$$\exists(LR_1S_1, LR_2) \& \exists(LR_2S_1, LE_2) \\ \& \exists(LE_1S_1, LE_2) \\ \rightarrow (LR_1S_1, LE_1S_1) \qquad (10)$$

If the sense-candidate of the Russian dictionary is connected with a semantic link to a lexeme that has a sense with a link to the lexeme of the German dictionary, and the sense-candidate of the English dictionary is connected with a link of the same type with a lexeme that has sense that is linked to the same lexeme of the German dictionary, then candidates should have a reconstructed link.

For example, the correctness of the reconstruction of the link between senses of the lexemes "important" and "важный" can be verified as follows. One of the senses of the English-speaking lexeme "important" is an antonym of the lexeme "petty", one of which senses, in turn, refers to the German-speaking lexeme "unbedeutend": LE_1 (important) → LE_1S_1 (Having relevant and crucial value.) → LE_2 (petty) → LE_2S_1 (Little, small, secondary in rank or importance) → LD_2 (unbedeutend).

At the same time, one of the senses of the Russian-speaking lexeme "важный" is the antonym of the lexeme “незначительный”, which similarly refers to the German-speaking lexeme "unbedeutend": LR1 (важный) → LR_1S_1 (имеющий большое значение) → LR_2 (незначительный) → LR_2S_1 (небольшой по количеству, размерам, силе проявления) → LD_2 (unbedeutend).

Hence, we conclude that the link between the senses of the lexemes "important" and "важный" was correctly defined (Figure 6).

3.6 Verification using a Sequence of Translations

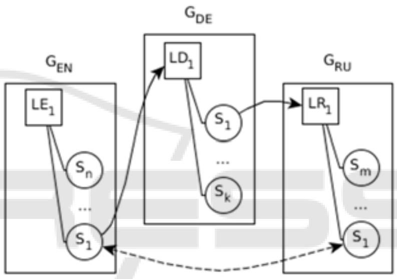

Figure 7: Sequence of translations.

If the sense-candidate from English dictionary has a link to the lexeme of the German dictionary, that in turn has a link to sense, containing a link to the lexeme of the Russian dictionary that contains the sense-candidate, then the link between candidates is a restored link (Figure 7).

$$\forall \begin{cases} LR_1 \in G_{RU} \\ LR_2 \in G_{RU} \\ LE_1 \in G_{EN} \\ LE_2 \in G_{EN} \\ LD_1 \in G_{DE} \end{cases}. \qquad (11)$$

$$\exists(LE_1S_1, LE_2) \& \exists(LE_2S_1, LD_1) \& \\ \exists(LR_1S_1, LR_2) \& \exists(LE_2S_1, LD_1) \\ \rightarrow (LR_1S_1, LE_1S_1) \qquad (12)$$

For example, the verification of the correctness for the reconstruction of the link between the senses of the English lexeme “rivalry” and the Russian lexeme “соперничество” can be done as follows.

The original sense of the lexeme "rivalry" has a link-translation to the lexeme "Rivalität" (from German Wiktionary), one of the senses of which, in turn, has a similar link to one of the senses of the Russian lexeme "соперничество": LE_1 (rivalry) → LE_1S_1 (The relationship between two or more rivals who regularly compete with each other) → LD_1 (Rivalität) → LD_1S_1 (das Verhältnis von Rivalenzueinander) → LR_1 (соперничество) → LR_1S_1 (ситуация, при которой кто-либо стремится превзойти, победить кого-либо другого в чём-либо). Thus, the correctness of the reconstruction of the link is verified.

4 RESULTS

The experiments are based on Wiktionary dump from 17.04.2017. The Russian version of Wiktionary contained 569,120 dictionary entries (lexemes), 1,298,654 individual senses and 305,024 links of type "sense-lexeme".

The algorithm for the reconstruction of links with the usage of translingual links has been implemented as follows. The candidates for links were reconstructed (result of applying rules from Sections 3.1-3.3). Links in the list of candidates were checked for correctness (applying rules from Sections 3.4-3.6). If none of the rules from previous step were applied to the candidate, such link was not included in the dictionary.

As a result of applying the rule from Section 3.1 we get 25,400 candidate links. The rule from Section 3.2 restored 23,851 links (when searching for translations from English Wiktionary); the same rule applied for translations from Russian Wiktionary restored 18,023 links. The rule from 3.3 is also symmetric. It was applied twice: searching for translations from English Wiktionary allowed to reconstruct 193 links, and similar search for Russian Wiktionary allowed to restore 2,737 links. The result of the usage of rules from Sections 3.1-3.3 is 69,309 potential links-transfers that were reconstructed (excluding duplicated links).

Using rules from Sections 3.4-3.6 we verified the correctness of reconstructed links from candidate list. The rule from Section 3.4 allowed to confirm the correctness of 4,448 links: 1,732 links using translations of Russian articles (1,224 synonyms, 450 antonyms, 41 hyponyms and 17 hyperonims) and 2,716 - using translations of English articles (1,736 synonyms, 618 antonyms, 258 hyponyms and 104 hyperonims). The rule from point 3.5 confirmed the validity of 737 links (501 synonyms, 187 antonyms, 15 hyponyms and 34 hyperonims). The rule from Section 3.6 allowed to confirm the correctness of 17,027 reconstructed links – 5,807 translations from Russian to English and 11,220 from English to Russian.

Consistent application of the rules from Sections 3.4-3.6 allowed to confirm the correctness of the reconstruction of 16,664 links obtained in the previous stage.

5 CONCLUSIONS

After applying the developed method, more than 16,000 links between the nodes of semantic senses of Russian and English sections of Wiktionary were reconstructed. These links allowed to create a generalized ontology. Inter-semantic links expand the connections of the sense with a set of synonyms in synset (Miller, George, 1990), (Gross, and Miller, 1990), (Fellbaum, Christiane, 1990) for the lexeme. As a result we can merge the English semantic node of ontology with Russian semantic node. In the implementation of presented approach we decided to keep two semantic nodes separated, but we connected them with a special type of the link called "synonym-translation".

For further research we plan to combine the developed method with results of our previous researches to reconstruct more sense-to-sense links. Another direction of our further researches is focused on study the ability to apply the information from temporal content of articles in Wiktionary in context of link reconstruction, namely the history of article changes.

REFERENCES

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

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.*

Bruggen, R., 2014. Learning Neo4j. *Packt Publishing Ltd.*

Fellbaum, C., 1990. English verbs as a semantic net. *International Journal of Lexicography.*

Gross, D., Miller, K., 1990. Adjectives in wordnet. *International Journal of lexicography.*

Klimenkov, S., Tsopa, E., Pismak, A., Yarkeev, A., 2016. Reconstruction of Implied Semantic Relations in Russian Wiktionary. *Proceedings of the 8th International Joint Conference on Knowledge*

Discovery, Knowledge Engineering and Knowledge Management (KDIR).

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

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

Miller, George A., 1990. Nouns in WordNet: a lexical inheritance system. *International journal of Lexicography.*

Pismak, A., Kharitonova, A., Tsopa, E., Klimenkov, S., 2016. Method of automatic construction of semantic network from weakly structured sources. *Software products and systems.*

Torsten, Z., Müller, C., Gurevych, I., 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *LREC.*

Zaliznyak, A., 2008. Grammatical dictionary of the Russian language *M .: AST-Press.*