National Research University Higher School of Economics

as a manuscript

Daniil Andreyevich Alexeyevsky

METHODS FOR AUTOMATIC WORDNET RELATION EXTRACTION FROM DICTIONARY DEFINITIONS

PhD Thesis Summary
for the purpose of obtaining
Philosophy Doctor in Philology and Linguistics HSE

Academic Supervisor
PhD in Linguistics
Svetlana Toldova

Moscow 2018

General overview of the thesis

The thesis presents a methodology for semi-automated extraction of thesaurus relations from a corpus of dictionary definitions. It is suggested that this methodology may be used as a tool for electronic thesauri development. Automated relations extraction is one of the prioritized areas of contemporary linguistics both as an independent research field and as incorporated into the process of designing ideographic dictionaries, thesauri and ontologies.

Within the scope of this work we approach automated relations extraction as a tool for building electronic thesauri, which are used for numerous semantically oriented tasks in text processing, among which are fact extraction, text tone analysis, disambiguation, question-answering systems, etc. The aim of the present research is to develop an approach that facilitates the extraction of relationships for building such thesauri.

Electronic thesauri are based on a multitude of concepts and relations that bind concepts and words. We call these relations thesaurus relations. Sets of these relations are formed differently. They may be created by a lexicographer, as well as extracted from ontologies and dictionaries of different types, text corpora, or a database designed for any semantic model. As a method for the present research we accept extracting relations from a dictionary definition corpus.

Thus, we define **the subject** of this thesis as a set of methods for automated and semi-automated extraction of thesaurus relations.

The data for the present work comes primarily from the Big Russian Explanatory Dictionary (BRED) by S.A. Kuznetsov, supported by auxiliary materials. The dictionary has a rich structure and includes morphological, derivational, grammatical, phonetic, etymological information, three-level sense hierarchy, usage examples and quotes from classical literature and proverbs. The electronic version of the dictionary is produced by OCR and proofreading with very high quality (less than 1 error in 1000 words overall). The version also has sectioning markup of lower quality, with FPR in the range of 1~10 in 1000 tag uses for the section tags of our interest. We developed specific preprocessor for the dictionary that extracts word, its definition and usage examples (if any) from each article. We call every such triplet word sense, and give it unique numeric ID. An article can have reference to derived word or synonym instead of text definition. Type of the reference is not annotated in the dictionary. We preserve such references in a special slot of word sense.

To successfully apply the methods proposed for extracting thesaurus relations, we need a morphological annotation tool and a tool for the assessment of semantic proximity. Morphological markup was accomplished with Mystem tagger. Several methodologies for semantic proximity assessment are analysed within the framework of this study: Serelex database, vector models word2vec and AdaGram, of which the latter need to be trained. The following corpora were used as golden standard datasets: RuTenTen11, RuWac, lib.ru, RuWiki.

The scientific novelty of the thesis stems from the new means and methods for adding new relations to a thesaurus. The methodology proposed demands only a limited expert contribution and is well applicable for languages that are not supported by a large number of linguistic resources.

At the moment of writing this paper, electronic thesauri are available for less than 200 languages. Thus, the research is highly relevant for numerous languages, such as Moksha for instance, that are not supported by electronic thesauri.

The theoretical significance of the present study is defined by the development and further research into the set of thesaurus relations which uncovers taxonomic structure of the basic concepts in Russian and investigates the set of linguistic features relevant for extracting thesaurus relations from explanatory dictionaries.

The practical significance of the thesis consists in the designing an approach to building a set of thesaurus relations, defining the set of relations for Russian and analysing the algorшthms used for their extraction.

Public demonstration of the results.

- The 9th Russian Summer School in Information Retrieval (RuSSIR 2015), Saint-Petersburg, Russia, August 24–28 2015. Alexeyevsky D., Toldova S. “Key noun phrases for biological fact extraction”,
- The Eighth Global WordNet Conference 2016, Bucharest, Romania, January 27–30 2016. Alexeyevsky D. A., Temchenko A. V. “Word sense disambiguation in monolingual dictionaries for building russian wordnet.”
- The 10th Russian Summer School in Information Retrieval (RuSSIR 2016), Saratov, Russia, August 22–26 2016. Alexeyevsky D. A., Tregubova M. A. “Semi-supervised Relation Extraction from Monolingual Dictionary”,
- The 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017), Budapest, Hungary, april 17–23 2017. Alexeyevsky D. A. “Semi-supervised Relation Extraction from Monolingual Dictionary for Russian WordNet.”

The following propositions are submitted for the defence

- a new method of semi-automated thesaurus relation extraction developed within the framework of this thesis generates pairs of lexeme meanings bound with the same thesaurus relation; those are primarily hypernym and hyponym relations; using this methodology considerably facilitates expert work;
- a new method of grouping dictionary definitions proposed in the thesis enables classification of their structural properties. The method is based on clusterization of definitions with the use of lexico-grammatical n-gram properties;
- using lexico-grammatical trigrams as clustering features increases precision in defining definition types, as compared to homogenous (only lexical or only POS) trigrams and unigrams;

- various structural definition types defined with clusterization match different patterns for hypernym extraction. These patterns are extracted semi-automatically on the basis of previous expert annotation;
- investigating different automated disambiguation techniques for the lexemes holding hypernym relations with their definitions shows that the tasks in question can be solved by standard algorithms based on Lesk methodology, as well as by semi-supervised machine learning (Label Propagation) and methods coming from distributional semantics. The latter outperform the others in quality.

The contents of the thesis

The thesis includes an introduction, four chapters, conclusion and references.

The first chapter «Building semantic networks: motivation, approach, sources» defines the basic terminologe accepted in the study, gives a brief description of the history of thesauri develpment and the evolution of tools used for their design, with the latter being the main motivation for this work. The chapter presents the aim of the study and an overall description of the accepted approach. **Chapter two** «Data» is a brief review of explanatory dictionaries available for the Russian language. It explains theoretical and practical premises for developing the corpus of dictionary definitions and grounds the choice of the dictionary that serves as the main data source. In **the third chapter** «Relation extraction» we describe experiments on annotating definitions of lexical units that are bound with the defined word by a thesaurus relation. **Chapter four** «Word sense disambiguation (WSD)» thoroughly describes two experiments on automatic disambiguation of annotated lexical units. The results of the research and further discussion are presented in the **Conclusion**.

Summary of the thesis.

Chapter I «Building semantic networks: motivation, approach, sources» outlines the terminology accepted in the thesis, reviews the previous work done within the field, puts forward the aim of the research and gives a brief overview of the proposed solutions. The chapter also introduces the further structure of the thesis.

Princeton WordNet project (Fellbaum, 2012) gave a feasible impetus to the development of electronic thesauri. After the project was launched for English, two main approaches were widely exploited to create WordNet for any given language: dictionary-based concept (Brazilian Portuguese WordNet, Dias-da-Silva et al., 2002) and translation-based approach (see for example, Turkish WordNet, Bilgin et al., 2004). The last one assumes that there is a correlation between synset and hyponym hierarchy in different languages, even in the languages that come from distant families. Bilgin et al. employ bilingual dictionaries for building the Turkish WordNet using existing WordNets.

Multilingual resources represent the next stage in WordNet history. EuroWordNet, described by Vossen (1998), was build for Dutch, Italian, Spanish, German, French, Czech, Estonian and English languages. Tufis et al. (2004) explain the methods used to create

BalkaNet for Bulgarian, Greek, Romanian, Serbian and Turkish languages. These projects developed monolingual WordNets for a group of languages and aligned them to the structure of Princeton WordNet by the means of Inter-Lingual-Index. Following the creation of PrincetonWordNet in 1998, similar lexical ontologies with a similar set of node types and relations found their use in NLP. These are generally called WordNets and are developed for many languages. Methods for building new WordNets range from those based on human labor to mostly automated methods. Typically automated approaches involve either extracting relations from machine-readable dictionaries or translating an existing wordnet, although other approaches were attempted too. While translation-based approaches are the most simple, they have a presumption that ontological structure of different languages is similar, which is a questionable statement, especially for languages of different families. Several attempts were made to create Russian WordNet. (Azarova et al. 2002) attempted to create Russian WordNet from scratch using merge approach: first the authors created the core of the Base Concepts by combining the most frequent Russian words and so-called "core of the national mental lexicon", extracted from the Russian Word Association Thesaurus, and then proceeded with linking the structure of RussNet to EuroWordNet. The result, according to project's site, contains more than 5500 synsets, which are not published for general use. Group of (Balkova et al. 2004) started a large project based on bilingual and monolingual dictionaries and manual lexicographer work. As for 2004, the project is reported to have nearly 145 000 synsets (Balkova et al. 2004), but no website is available (Loukachevitch and Dobrov, 2014). (Gelfenbeyn et al. 2003) used direct machine translation without any manual interference or proofreading to create a resource for Russian WordNet. Project RuThes by (Loukachevitch and Dobrov 2014), which differs in structure from the canonical Princeton WordNet, is a linguistically motivated ontology and contains 158 000 words and 53 500 concepts at the moment of writing. YARN (Yet Another RussNet) project, described by (Ustalov 2014), is based on the crowd-sourcing approach towards creating WordNet-like machine readable open online thesaurus and contains at the time of writing more than 46 500 synsets and more than 119 500 words, but lacks any type of relation between synsets. According to the view accepted within this thesis, the most efficient method for building an electronic thesaurus is extracting relations from explanatory dictionaries, which makes it possible to develop a large part of a thesaurus with the least contribution from an expert's side. Thus, previously introduced attempts to create a WordNet for the Russian language didn't lead to freely-available complete lexical ontology conforming to a WordNet definition, and a place for the Russian resource of this kind still remains vacant.

The aim of the present work is to create a methodology that results into a corpus of thesaurus relations which can eventually be used for compiling a full-fledged thesaurus. Another question that we seek to find answers for is whether low resource – both expert and electronic – is enough to obtain reliable results. A tool enabling to develop a corpus of thesaurus relations on the basis of limited resources will extensively broaden the range of languages supported by electronic thesauri.

The accepted approach serves a basis for investigating the typology of taxonomic relations between the basic concepts among speakers of different languages and includes the following steps:

- building up a corpus of definitions;
- extracting triples «meaning – relation – lexeme» from definitions;
- word sense disambiguation.

This process generates chains of word meanings linked with thesaurus relations. Although these chains might require verification, they significantly simplify manual contribution in the process of thesaurus development. They may also be used as an independent linguistic tool for numerous tasks that any thesaurus is apt for, as well as a means of verification and enrichment of existing resources.

***

Chapter II «Monolingual dictionaries: a semi-structured resource» presents a brief overview of explanatory dictionaries available for the Russian language, it also gives a thorough description of how the definition corpus is developed.

With the development of lexicography, the structure of dictionary entries was becoming more and more homogeneous. Thus, some contemporary dictionaries define a closed set of words that can be used in definition entries, enumerate all meanings and ascribe corresponding indices to words (LDOCE). Lexicography today has a number of challenges: dictionaries are to be electronically based and their layout should be separated from their purpose and markup. Such dictionaries are referred to as machine readable (e.g. Der Danske Oordbog). Contributing new words to a dictionary is another challenge. An illustration of a dictionary that is constantly enriched with new words is Wiktionary, however, due to its nature the uniformity of its entries' structure is relatively low.

The present thesis focuses on data extraction from dictionary definitions. This choice is motivated, among other reasons, by the fact that the language of dictionary entries is a limited subset of a natural language. As a source of data for the present work we took the Big Russian Explanatory Dictionary (BRED) by S.A. Kuznetsov. As BRED describes itself, word meanings are presented in three levels, however in the course of our work we will show that it proposes a more extended hierarchy of meanings. The present research thoroughly describes the typical stages, necessary for developing a definition corpus on the basis of BRED, the most important of which are building a hierarchical structure of a definition and extracting necessary information from a structured entry.

Chapter II also gives an account of the difficulties that are connected with morphological properties assignment on the basis of Kuznetsov's dictionary. Thus, we have conducted an experiment for Mystem tagging tool to evaluate its precision in POS annotation for title words in the dictionary. The results show that within the subset of 1000 words the precision of POS markup is 98.0%. Finally we obtained a corpus of noun definitions that is used in all the experiments set within the framework of the present study.

***

Chapter III «Relation extraction» describes the experiments on annotation in definitions of words that hold thesaurus relation with the defined term.

It gives a review of various methods used in relation extraction from dictionary definitions and based on a limited set of lexical and grammatical rules. The main problem associated with this approach stems from discriminating power of these rules. The chapter briefly describes two experiments carried out within the framework of the present study.

Both experiments are based on the corpus of noun definitions described in the previous chapter.

The first experiment shows that the extraction of hypernym relations based on a single rule is not universally applicable. The rule designed within the experiment framework is: the first noun in the nominative is also a hypernym for the defined one. A test corpus is tagged for checking the rule. It is shown that the accuracy of extracting hyponym relations within this rule is 0,5. The second experiment focused on the possibility to improve the results with the help of preliminary clusterization.

The following three steps are carried out:

- cluster word sense definitions,
- annotate each cluster (as a whole) by a human expert,
- summarize annotation results.

The aim of clustering step is to reduce the amount of work for human annotators. Thus each cluster should have as few clusters as possible, while each cluster has as regular syntactic and semantic coherence as possible.

Given a cluster of word sense definitions an annotator has to answer three questions:

- is it possible to extract WordNet relation from most definitions in the cluster, and if so, which relation it is,
- what morphosyntactic rule can extract the relation,
- for what fraction of definitions in the cluster does this rule give the correct answer?

To measure the rule quality the expert assesses the result of rule application for 25 cases in each cluster (or the whole cluster, if the cluster is smaller than 25 word senses).

The annotation guideline strongly suggests to the expert to annotate exactly one rule per each cluster. This means that it is more harmful to merge unsimilar clusters in clustering step than to split similar definitions into several clusters.

The aim of clustering in the work is to group together definitions that have the same presentation style and will likely be parsed using the same morphosyntactic rule. Author used the following assumptions about definition structure:

- a few first words are usually enough to guess the article style
- style manifests itself in syntactic structure
- some styles manifest in presence of specific genus terms and have specific
- coordination structure for them
- dictionary authors strive to have a few standardized wording styles, and hence features defining every style are frequent within the dictionary.

This was accomplished by using the following set of features:

- lexical unigrams: word-form, lemma

- morphological unigrams: part of speech, every morphological feature as a tag
- compound morphological unigrams: full morphological parse (gr), immutable morphological features (immutable_gr, e.g. part of speech, gender and animacy for nouns), mutable morphological features (mutable_gr , e.g. case and number for nouns)
- mixed trigrams with templates:
    - (lemmas, immutable_gr, immutable_gr) ,
    - (immutable_gr, lemmas, immutable_gr) ,
    - (immutable_gr, immutable_gr, lemmas).

For each feature type it's frequency list was built and only the top 200 most frequent were used. This restriction follows two aims: both to alleviate the dimensionality curse and to reduce amount of noise features. Vector representation of features for clustering and analysis was produced using bag-of-words model for the first three words concatenated with the average for the whole definition.

Classical n-gram is a n-tuple of sequential elements from a list. Similarly, given n different lists let us define mixed n-gram as a n-tuple of sequential elements from the list, each element in the tuple corresponds both to sequential list and to sequential position in the list. In the linguistic domain let us call the set of lists used n-gram template.

To assess the quality of the rules obtained, we grouped clusters by relation that can be extracted from each cluster and counted number of definitions in the group and combined estimate of precision.

For each group of clusters an overall number of definitions:

	Russian WordNet		**Onto.PT**	
relation	**amount**	**precision**	**amount**	**precision**
hypernym	53246	85.54%	29,563	59.10%
synonym	10044	75.69%	11,862	86.10%
junk	7175	100.00%		
hypernym synonym	4160	76.11%		
hyponym	2761	53.71%		
part of	1017	100.00%	1,287	52.60%
domain	495	51.72%		
instance of	253	61.26%	253	61.26%
hypernym hypernym	125	100.00%		
has part	105	92.38%		
dictionary	58621	83.93%	37898	76.64%

Table 2. Estimate on number of extracted relations and extraction precision as compared to Onto.PT.

The chapter contains the results of the experiment and their discussion. It is obvious that clusterization helps improve the quality of relation extraction.

***

Chapter IV «Word sense disambiguation (WSD)» presents two experiments on automatic disambiguation of annotated words. It gives an overview of existing disambiguation alorithms, which can be divided into three classes:

1. algorithms based on simple heuristics;
2. algorithms based on machine learning;
3. algorithms using distributional semantics models.

Special attention was given to two groups of algorithms:

1. Lesk algorithm and its modifications
2. algorithms using results obtained by predicting vector models Word2Vec and AdaGram as features for machine learning.

The chapter analyses the possibility to apply the algorithms that we chose for thesaurus relation extraction from definition corpora and their potential modifications that could improve WSD results. Here we focus on hypernym and hyponym relations.

The first part focuses on Lesk algorithm and its modifications. We set an experiment which tests different approaches to feature extraction, weight modifications and options to improve the results with the help of Serelex – a word associations database.

We have developed a pipeline for massively testing different disambiguation setups. The pipeline is preceded by obtaining common data: word lemmas, morphological information, word frequency. For the pipeline we broke down the task of disambiguation into steps. For each step we presented several alternative implementations. These are:

- Represent candidate hyponym-hypernym sense pair as a Cartesian product of list of words in hyponym sense and list of words in hypernym sense, repeats retained.
- Calculate numerical metric of words similarity. This is the point we strive to improve. As a baseline we used: random number, inverse dictionary definition number; classic Lesk algorithm. We also introduce several new metrics described below.
- Apply compensation function for word frequency. We assume that coincidence of frequent words in to definitions gives us much less information about their relatedness than coincidence of infrequent words. We try the following compensation functions: no compensation, divide by logarithm of word frequency, divide by word frequency.
- Apply non-parametric normalization function to similarity measure. Some of the metrics produce values with very large variance. This leads to situations where one matching pair of words outweighs a lot of outright mismatching pairs. To mitigate this we attempted to apply these functions to reduce variance: linear (no normalization), logarithm, Gaussian, and logistic curve.
- Apply adjustment function to prioritize the first noun in each definition. While extracting candidate hypernyms the algorithm retained up to three candidate nouns in each article. Our hypothesis states that the first one is most likely the hypernym. We apply penalty to the metric depending on candidate hypernym position within hyponym definition. We

tested the following penalties: no penalty, divide by word number, divide by exponent of word number.

- Aggregate weights of individual pairs of words. We test two aggregation functions: average weight and sum of best N weights. In the last case we repeat the sequence of weights if there were less than N pairs. We also tested the following values of N: 2, 4, 8, 16, 32.
- Algorithm returns candidate hypernym with the highest score.
- The data for these experiments comes from the BRED dictionary described in Chapter II earlier.

For testing the algorithms we selected words in several domains for manual markup. We determined domain as a connected component in a graph of word senses and hypernyms produced by one of the algorithms. Each annotator was given the task to disambiguate every sense for every word in such domain. Given a triplet an annotator assigns either no hypernyms or one hypernym; in exceptional cases assigning two hypernyms for a sense is allowed.

One domain with 175 senses defining 90 nouns and noun phrases was given to two annotators to estimate inter-annotator agreement. Both annotators assigned 145 hypernyms within the set. Of those only 93 matched, resulting in 64% inter-annotator agreement. The 93 identically assigned hyponym-hypernym pairs were used as a core dataset for testing results. Additional 300 word senses were marked up to verify the results on larger datasets. The algorithms described were tested on both of the datasets.

One known problem with Lesk algorithm is that it uses only word co-occurrence when calculating overlap rate (Basile et al., 2004) and does not extract information from synonyms or inflected words. In our test it worked surprisingly well on the dictionary corpus, finding twice as many correct hypernym senses as the random baseline. We strive to improve that result for dictionary definition texts.

Russian language has rich word derivation through variation of word suffixes. The first obvious enhancement to Lesk algorithm to account for this is to assign similarity scores to words based on length of common prefix. In the results we refer to this metric as advanced Lesk.

Another approach to enhance Lesk algorithm is to detect cases where two different words are semantically related. To this end we picked up a database of word associations Serelex (Panchenko et al, 2013). It assigns a score on a 0 to infinity scale to a pair of noun lemmas roughly describing their semantic similarity. As a possible way to score words that are not nouns in Serelex we truncate a few characters off the ends of both words and search for the best pair matching the prefixes in Serelex. (See “prefix serelex” in the Table).

We tested several hypotheses on how these two metrics can be used to improve the resulting performance. The tests were: to use only Lesk; to use only Serelex; to use Serelex where possible and fallback to advanced Lesk for cases where no answer was available; and to sum the results of Serelex and Lesk. Since Serelex has a specific distribution of scores we adjusted the advanced Lesk score to produce similar distribution.

For each estimator we performed full search through available variations on steps 3-6 of the pipeline and selected the best on the core set and estimated again on the larger dataset.

Test results are given in the Table:

Algorithm	**CoreSet**	**LargeSet**
random	30.8%	23.9%
first sense	38.7%	37.7%
naive Lesk	51.6%	**41.3%**
serelex	49.5%	38.0%
advanced Lesk	**53.8%**	33.3%
serelex with adjusted Lesk fallback	52.7%	36.3%
serelex + adjusted Lesk	52.7%	38.3%
prefix serelex	**53.8%**	38.0%

Precision of different WSD algorithms.

The second part explores different methods using distributive semantic models. This section gives a brief review of the history of distributive semantics methods: at the moment of writing this text these methods performed best in disambiguation tasks. The review ends up grounding the choice of two distributional models and the methods of their application for solving disambiguation tasks.

Further in the text, we give a thorough description of an experiment that compares different approaches to extract WSD features from dictionary definitions and several machine learning algorithms. The main task of the experiment is whether using unmarked and tagged data improves WSD results. The data that we use in the experiment was described in Chapter II.

The training dataset included 394 definitions with previously matched hypernyms. Annotated dataset consists of 114 hyponym-hypernym pairs marked up by two annotators. The pairs were selected to have hypernyms of different frequency. We employed two metrics for inter-annotator agreement. Weak agreement which is defined by Fleiss kappa metric as $\kappa = 0.57$, and strong agreement as $\kappa = 0.36$. The section grounds the choice of κ metrics and describes criteria for strong and weak agreement.

The goal of WSD task setup is to permit comparison of supervised and semi-supervised machine learning methods under a hypothesis that presence of similar WSD problems helps to find reliable answers with the use of unmarked data. The section is concluded by results discussion. The best disambiguation algorithm generated 0.7 precision value. This value is not sufficient to fully automate hyponym-hypernym chains. However few transformations are needed to eliminate possible errors. Thus, the method may be used as a helpful tool that considerably reduces expert's manual work in building an electronic thesaurus. The chapter also gives guidelines to further algorithm improvement.

***

The «Conclusion» chapter lists the main outcomes of the thesis among which are:

- the thesis proposes a new approach to semantic relation extraction from a corpus of dictionary definitions;
- the thesis explores various disambiguation methods that are relevant for its main tasks;
- the thesis gives a description of a method that generates word meaning chains linked to each other with thesaurus relations;
- the thesis presents a corpus of such chains developed on the basis of BRED by S.A. Kuznetsov and suggests their thorough multifaceted analysis;

The thesis also puts forward a hypothesis that the methods proposed form a sufficient basis for designing thesauri in low-resource languages and minimize the contribution coming from experts. Checking this hypothesis is the subject of further research.

The following papers discuss the main results of the thesis:

- *Alexeyevsky, Daniil Andreevich.* "BioNLP ontology extraction from a restricted language corpus with context-free grammars" // Informatics and its Applications, vol. 10 issue 2, pp. 119128, 2016, Moscow, Russian Academy of Sciences, Branch of Informatics, Computer Equipment and Automatization.
- *Alexeyevsky, Daniil, and Anastasiya V. Temchenko.* "WSD in Monolingual Dictionaries for Russian WordNet." // In Proceedings of the Eighth Global WordNet Conference, 1015. Bucharest, Romania, 2016.
- *Alexeyevsky, Daniil.* "Semi-Supervised Relation Extraction from Monolingual Dictionary for Russian WordNet." // In Proceedings of CICLing17 Conference. LNCS, 2018. *(in print)*
- *Alexeyevsky, Daniil.* "Word sense disambiguation features for taxonomy extraction" // Computacion y Sistemas, vol. 22 issue 3. 2018