Automatic Extraction of Hypernyms and Hyponyms from Russian Texts

Kristina Sabirova, Artem Lukanin

South Ural State University, Chelyabinsk, Russia
{bezadresa.net,artyom.lukanin}@gmail.com

Abstract. The paper describes a rule-based approach for hypernym and hyponym extraction from Russian texts. For this task we employ finite state transducers (FSTs). We developed 6 finite state transducers that encode 6 lexico-syntactic patterns, which show a good precision on Russian DBpedia: 79.5% of the matched contexts are correct.

Keywords: text mining, wordnet, hypernym, hyponym, noun.

1 Introduction

These days there is no established Russian WordNet, that is why automatic extraction of hyponyms is of great value for Russian Natural Language Processing. The concept of this work was created after the investigation of Russian syntactical structures, which contain hypernyms and hyponyms, and the review of Serelex project [14], developed for English and French [10].

The aim of this project is to extend the approach devised in the Serelex project to the Russian language. In particular, to accomplish this task we are using corpus processing tool Unitex 3.1beta [16] for automatic extraction of hypernyms and hyponyms from Russian DBpedia [12], [6]. The extracted hypernyms and hyponyms can be used to ease the creation of Russian thesauri such as RussNet [2] or YARN [3] or for query expansion in information retrieval systems.

2 Related Work

There are a lot of methods of hypernym extraction, from simple lexical patterns [7], [9], a combination of a morphological analyzer and surface syntax parsing [1], to machine learning techniques [4-5], [13] and [11]. One of the highest-coverage methods is proposed by Snow et al. [15] Firstly, they are looking for the sentences that contain two terms which are known to be in the taxonomic relations, then they parse the sentences and automatically extract patterns from the parse trees. Finally they train the hypernym classifier based on these features. Lexico-syntactic patterns are generated for each sentence relating a term to its hypernym, and a dependency parser is used to represent them.

Hearst [7] designed 6 lexico-syntactic patterns for English, which were later extended by Panchenko et al. [10] with 12 further patterns for English and French. The results of the extraction are used in Serelex, a lexico-semantic search engine. Given a query, it returns a list of related words. The system gives the opportunity to discover the meaning of words in an interactive manner, search for synonyms and more. For example, for the query “fruit” the output is “vegetable”, “mango”, “apple”, etc. [14].

3 Russian Lexico-Syntactic Patterns for Hypernym and Hyponym Extraction

We hypothesize that the hyponymic relations are specific for most notional lexico-grammatical classes, but they are better defined for nouns and verbs. In this study, we investigate only nouns.

Our method is based on our patterns deduced in the previous work, as well as the patterns made in the Serelex project. The aim was to translate the existed patterns, to interpret them for Russian, to complete them and to create new patterns.

The Extended Abstracts corpus without accents [6] of Russian DBpedia [12] was used as the material for the research. The corpus consists of 1,325,859 sentences and ~47,000,000 tokens. For the practical part of the research Unitex was used.

Unitex is a collection of programs developed for natural language analysis using linguistic resources and tools (electronic dictionaries, grammars and lexico-grammatical tables), that gives the opportunity to develop FSTs in the graphical interface for the designed patterns. It was created for French by Maurice Gross and his students at the Laboratoire d’Automatique Documentaire et Linguistique (LADL). Similar resources were developed for other languages in the context of the RELEX laboratory network.

The electronic dictionaries specify simple and compound words with their lemmas and a set of grammatical codes. The availability of these dictionaries is the main advantage for pattern searching. The information they contain can be used for searching and matching the contexts from which the lexico-semantic relations can be extracted. These dictionaries were made by teams of linguists for different languages: English, French, Greek, Italian, German, Korean, Polish, etc. [17]. We use the full version of the Russian computational morphological dictionary, developed at CIS, Munich [8].

During the research we designed 6 patterns for the hyponym and hypernym extraction from Russian texts. For every pattern we developed a finite state transducer in Unitex and applied them to the text corpus of Russian DBPedia without accents.

To reduce the probability of matching incorrect contexts special rules were desined. These rules are mostly exceptions, enclosed in the right negative contexts to the left of the probable hypernym or hyponym. Hyponyms and hypernyms are matched using the special symbols <N> (any noun) or <!DIC> (any token not found in the dictionary). To increase the probability, that <!DIC> will match a noun, additional special symbols are placed before and after this token in the pattern, e.g. a lexico-grammatical classes like <A> for adjectives, <PREP> for prepositions, etc. or lexical masks like <первый> (first) for matching any word form of this numeral.

The patterns with the examples are presented below (X – hypernym, Y – hyponym).

Pattern 1. Такие/таких/таким X, как Y[, Y] и/или Y. (Such X as Y,[Y,] and/or Y). An example of a matched context:

В Индии зародились такие {[*религии*]=HYPER} *как* {[*индуизм*]=HYPO}, {[*буддизм*]=HYPO}, {[*сикхизм*]=HYPO} *и* {[*джайнизм*]=HYPO}.

(In India such {[*religions*]=HYPER} *as* {[*Hinduism*]=HYPO}, {[*Buddhism*]=HYPO}, {[*Sikhism*]=HYPO}, *and* {[*Jainism*]=HYPO} *were born.)*

Pattern 2. X, такие/таких/таким как Y[, Y][и/или Y]. (X, such as Y,[Y,][and/or Y]). An example of a matched context:

...{*систем* [*верований*]=HYPER}*, таких как* {[*шаманизм*]=HYPO}, {[*политеизм*]=HYPO}, {[*пантеизм*]=HYPO}, {[*анимизм*]=HYPO}.

(...{*systems of* [*faith*]=HYPER}*, such as* {[*Shamanism*]=HYPO}, {[*Polytheism*]=HYPO}, {[*Pantheism*]=HYPO}, {[*Animism*]=HYPO}*.)*

Pattern 3. X: Y[, Y] и/или Y. (X: Y,[Y,] and/or Y). A matched context:

...мир, передаваемый человеку через {*его* [*ощущения*]=HYPER}*:* {[*зрение*]=HYPO}, {[*слух*]=HYPO}, {[*обоняние*]=HYPO}, {[*осязание*]=HYPO} *и другие.*

(...the world, transferred to a human through {*his* [*senses*]=HYPER}*:* {[*vision*]=HYPO}, {[*hearing*]=HYPO}, {[*smelling*]=HYPO}, {[*feeling*]=HYPO}*, etc.)*

Pattern 4. Y[, Y][(, а также)/(также как и)/и/или] другие/другим/других/о других X. (Y,[Y,] [(as well as)/and/or] other X). An example of a matched context:

Распространение ВИЧ-инфекции связано, главным образом, с незащищенными половыми контактами, использованием зараженных вирусом {[*шприцев*]=HYPO}, {[*игл*]=HYPO} *и* {*других* {*медицинских и парамедицинских* [*инструментов*]=HYPER}}*...*

(The major modes of HIV transition are sexual intercourse, unsterile reuse of single use {[*syringes*]=HYPO}, {[*needles*]=HYPO} *and* {*other* {*medical and paramedical* [*instruments*]=HYPER}}*...)*

Pattern 5. Виды/типы/формы/разновидности/сорта X, как Y[, Y] и/или Y. (Kinds/types/forms/sorts of X, such as Y,[Y,] and/or Y). A matched context:

Такие виды {[*оружия*]=HYPER} *как* {[*шпага*]=HYPO} *и* {[*рапира*]=HYPO} *тоже причисляют к мечам, что не совсем верно.*

(Such kinds of {[*weapon*]=HYPER} *as* {[*épée*]=HYPO} *and* {[*rapier*]=HYPO} *are classified as swords, that is not totally correct.)*

Pattern 6. Y — вид/тип/форма/разновидность/сорт X. (Y is a kind/type/form/sort of X). An example of a matched context:

{[*Хобби*]=HYPO} *— вид* {*человеческой* [*деятельности*]}*, некое занятие …*
({[*Hobby*]=HYPO} *is a kind of* {*human* [*activity*]}*, some engagement, interest…)*

4 Results

We ran 6 finite state transducers for the corresponding 6 patterns on a test corpus of the first 85,071 sentences of the full corpus [6]. It contains 3,058,878 tokens. We manually verified the results, and found that 79.5% of the units were extracted correctly (see Table 1).

Table 1. The number of extracted units from the test and the full version of the corpus

Pattern	Extracted contexts from the test corpus	Extracted hypernyms (errors)	Extracted hyponyms (errors)	Errors, %	Extracted contexts from the full corpus
1	36	36 (5)	110 (12)	11.6	364
2	51	51 (13)	113 (24)	22.6	653
3	137	148 (37)	560 (120)	22.2	1402
4	97	99 (15)	284 (51)	17.2	761
5	48	48 (12)	110 (19)	19.6	395
6	59	59 (18)	59 (18)	30.5	1279
Total:	428	441 (100)	1236 (244)	20.5	4854

The second column contains the number of matched contexts with extracted hypernyms (the third column) and hyponyms (the fourth column). We also applied these 6 developed FSTs on the full corpus [6]. This yielded 4,854 extracted contexts, in which approximately 3,859 hypernyms and 11,144 hyponyms were extracted correctly.

5 Conclusion

During the research we designed 6 lexico-syntactic patterns and verified them on a large corpus. We developed 6 finite state transducers corresponding to these patterns in Unitex. These FSTs matched 428 contexts on the test corpus and 4,854 contexts on the full corpus, 79.5% of the extracted units from the test corpus were correct.

Acknowledgements. This work is partially supported by the RFH grant #13-04-12020 “New open electronic thesaurus for Russian”.

References

1. Agirre, E., Olatz. A., Arregi, X., Artola, X., Diaz de Ilarraza, A., Lersundi, M., Martinez, D., Sarasola, K., Urizar, R. Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar. In Proceedings of Euralex (2000)
2. Azarowa, I. V. RussNet as a computer lexicon for Russian (2008)
3. Braslavsky, P., Mukkin, M., Lyashevskaya, O., Bonch-Osmolovskaya, A., Krzhizhanovsky, A., Egorov, P. YARN: the beginning. In Computer Linguistics and Intelligent Technologies 2013. V. 12(19). Part 3. (2013) - Браславский, П., Мухин, М., Ляшевская, О. Н., Бонч-Осмоловская, А. А., Кржижановский, А., Егоров, П. YARN: начало // Компьютерная лингвистика и интеллектуальные технологии 2013. Т. 12(19). Ч. 3. 2013.
4. Caraballo, S. Automatic constraction of a hypernym-labeled noun hierarchy from text. In Proceedings of the 37th Annual Meeting of the ACL, pp. 120-126 (1999)
5. Dolan, W., Vanderwende, L., Richardson, S. Automatically deriving structured knowledge bases from on-line dictionaries. In Proceedings of the First Conference of the Pacific ACL, pp. 5-14 (1993)
6. Extended Abstracts Corpus without accents of Russian DBpedia: `http://cental.fltr.ucl.ac.be/team/~panchenko/data/serelex/corpus-ru-dbpedia-short-dea.csv`
7. Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. In ACL, pp. 539–545 (1992)
8. Nagel, S. Formenbildung im Russischen. Formale Beschreibung und Automatisierung für das CISLEX-Wörterbuchsystem (2002)
9. Oakes, M.P. Using hearst’s rules for the automatic acquisition of hyponyms for mining a Pharmaceutical corpus. In Proceedings of the RANLP Workshop, pp. 63-67 (2005)
10. Panchenko, A., Morozova, O., Naets, H.: A Semantic Similarity Measure Based on Lexico-Syntactic Patterns. In Proceedings of KONVENS 2012 (Main track: poster presentations), pp. 174-178 (2012)
11. Ritter, A., Soderland, S., Etzioni, O. What is this, anyway: Automatic hypernym discovery. In Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read, pp. 88-93 (2009)
12. Russian DBpedia: `http://wiki.dbpedia.org/Downloads39`
13. Sanfilippo, A., Poznański, V. The acquisition of lexical knowledge from combined machine-readable dictionary sources. In Proceedings of the third Conference on Applied Natural Language Processing, pp. 80-87 (1992)
14. Serelex, `http://serelex.cental.be/`
15. Snow, R., Jurafsky, D., Ng, A. Learning syntactic patterns for automatic hypernym discovery. In Proceedings of Advanced in Neural Information Processing systems, pp. 1297-1304 (2004)
16. Unitex 3.1beta. Available under LGPL license: `http://www-igm.univ-mlv.fr/~unitex/` (2013)
17. Unitex 3.1beta Manual. `http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf`

Автоматическое извлечение гиперонимов и гипонимов из русскоязычных текстов

Кристина Сабирова, Артём Луканин

Южно-Уральский государственный университет, Челябинск, Россия
{bezadresa.net,artyom.lukanin}@gmail.com

Аннотация. Описанный в статье подход по извлечению гиперонимов и гипонимов из русскоязычных текстов основан на использовании правил. Правила описаны с помощью конечных преобразователей. Мы разработали 6 конечных преобразователей, кодирующих 6 лексико-синтаксических шаблонов. Данный подход показывает достаточно высокую точность на корпусе русскоязычной DBPedia: из 79.5% найденных контекстов правильно извлечены слова, находящиеся в гиперонимических отношениях.

Ключевые слова. Анализ текста, ворднет, гипероним, гипоним, существительное.