Extraction of Semantic Relations from Wikipedia Text Corpus

Olexandr Shanidze and Svitlana Petrasova[0000-0001-6011-135X]

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

s.alexandr21@gmail.com, svetapetrasova@gmail.com

Abstract. This paper proposes the algorithm for automatic extraction of semantic relations using the rule-based approach. The authors suggest identifying certain verbs (predicates) between a subject and an object of expressions to obtain a sequence of semantic relations in the designed text corpus of Wikipedia articles. The synsets from WordNet are applied to extract semantic relations between concepts and their synonyms from the text corpus.

Keywords: semantic relations, rule-based approach, Wikipedia, text corpus, synsets, WordNet.

Due to the growing volume of information, it requires systematization and processing. For example, the increasing number of natural-language texts greatly complicated the process of retrievalof necessary information. Therefore, developing data storage tools and mechanisms for their rapid and efficient processing is an urgent task of NLP. Attempts to cope with this problem led to the development of Information Extraction (IE). According to the extracted information, IE includes the following issues: named entities recognition; attributes/relations extraction; facts/events extraction.

The most challenging task is to get information about semantic relations between objects. A semantic relationis established between lexical units (words, collocations) within the certain semantic field that may be a class, meronymy/holonymy, synonymy, antonymy, and others:

– ISA relation (relation of classification):Object (Member of Class) *–>is a–>* Subject (Class);
– hypernymy: Subject *–>group of –>* Object;
– hyponymy: Object *–>variant of –>* Subject;
– meronymy: Object *–>component of –>* Subject [1].

For extracting information, semantic relations in particular,rule-based methods (using patterns) and machine learning methods (naive Bayes classifier, decision trees, support vector machine (SVM), Hidden Markov Models (HMM), etc.) are applied [2].

The paper proposes the algorithm for automatic semantic relations extraction using the rule-based approach.

Step 1. Preprocessing of the developed text corpus of 200 Wikipedia articles. The Internet encyclopedia presents a system for categorizing pages in the form of a category tree which shows the representativeness of the corpus in a certain category. For our research, we chose articles of Information Technologies category [3].

Step 2. Identifying certain verbs between a subject and an object of expressions in the texts that are assumed as semantic relations, e.g. Subject *->include*, *consist of*, *contain* ->Object.

Step 3. Extracting semantic relations (unidentified at the previous stage):

1. search the subjects and objects of predicates (verbs that represent semantic relations identified at the previous stage);

2. obtain synonyms for defined subjects and objects (concepts) from WordNet [4];

3. extract semantic relations between the concepts and their synonyms.

Table 1 shows the semantic relations extracted from the designed text corpus.

Table 3. Semantic Relations Extractedfrom Wikipedia Articles

No	Semantic relation	No	Semantic relation	No	Semantic relation
1	include	11	ability of	21	body of
2	contain	12	aspect of	22	component of
3	consist of	13	member of	23	control of
4	branch of	14	method of	24	mode of
5	class of	15	version of	25	subset of
6	block of	16	part of	26	group of
7	collection of	17	property of	27	quality of
8	description of	18	set of	28	variant of
9	form of	19	type of	29	characteristic of
10	list of	20	use of	30	section of

Consequently, we get the semantic information (the semantic network) of words from the text corpus, i.e. semantic relations between concepts (subjects and objects).

The use of technologies of extractionof semantic relationsfrom texts serves as the basis for developing text analysis tools that operate at a higher level, e.g.text mining. The result of automaticextraction of semantic relations can be used in search engines to extend queries, to construct ontologies, to expand existing and create new thesauri.

References

1. Petrasova, S.V., Khairova, N.F.: Automated semantic network construction based on the glossary. In: Horizons of Applied Linguistics and Linguistic Technologies: International Scientific Conference Megaling–2013,http://megaling.ulif.org.ua/tezi-2013-rik/, last accessed 2019/02/07.
2. Bolshakova, Ye.I., Vorontsov, K.V., Yefremova, N.E.: Automatic natural language texts processing and data analysis. Moscow, Higher School of Economics National Research University, 269 (2017)
3. Wikipedia. Information Technologies Category,https://en.wikipedia.org/wiki/Category:Information_technology, last accessed 2019/02/07.
4. WordNet,https://wordnet.princeton.edu, last accessed 2019/02/07.