EasyChair Preprint

№ 6119

Experiments of Rule Extraction from Raw Text

Olegs Verhodubs

July 20, 2021

Experiments of Rule Extraction from Raw Text

Olegs Verhodubs
oleg.verhodub@inbox.lv

Abstract. The Semantic Web technologies are being developed in order to make the Web machine-readable. These technologies are necessary because there is no equal in efficiency possibility that would allow processing raw text by machines. This research allows to decrease the lacks of the existing techniques of processing raw text by means of investigating the way of knowledge extracting from raw text. The purpose of the paper is to show the results of the program, developed for rule extraction from raw text.

Keywords: Rules, Natural Language Processing, Expert Systems

I. INTRODUCTION

Since the advent of the Web, it has been continuously saturating with information. At some stage in the development of the Web, there became so much information there that additional tools were needed to find the necessary information. At first, these were lists of websites sorted by different topics: music, sports, cars, etc. For a while this helped the users a lot, but soon it was not enough. The reason for this is that the Web has grown even more and that one-parameter categorization of each website on the Internet is ineffective. One-parameter categorization means assigning only one attribute for each website. For example, the website about Henry Ford should be categorized as biography, or car. Real life is immeasurably richer so that even a short text placed in the website is relevant to many categories. Necessity created a solution: new search engines with the possibility of assigning many attributes to one website had appeared. Search engine results have become much better: a list of hyperlinks to websites containing user-specified keywords are being produced. The user goes to every website from the list to find what he needs. The problem is that the list of hyperlinks to websites may consist of hundreds of items and sometimes it is necessary to crawl all of them, which takes a long time. It became clear that something else was needed.

A new approach is to use semantic technologies for the Web. These technologies are called the Semantic Web technologies, and they are aimed to make the Web machine-readable. It is assumed that in the future, the Semantic Web technologies will allow collecting all the information found in the Web together based on the user's request. The prospect of not browsing a website for every hyperlink from the list as it happens with modern search engines is very encouraging. Moreover, the Semantic Web technologies include the implementation of inference, which in itself is a promising innovation.

Despite the consistency of Occam’s razor in all areas, the Semantic Web technologies have been developed. Apparently, this happened because the technologies existed at that time could not provide users with a qualitative leap in the development of the Web. That is why it became necessary to develop new standards of describing information as RDF (Resource Description Framework) or OWL (Web Ontology Language). Although, it would be ideal if we could extract whatever we need from the raw text. The scientific discipline that studies raw text processing is natural language processing. There are a lot of advances in the area, but extraction of rules from raw text has not yet been explored. Howbeit, rules can be easily extracted from raw text.

Extracted rules can be used by expert systems, including Keyword Search Engine Enriched by Expert System Features [1]. The ability to reason with rules extracted from the raw text is very promising. This paper describes a computer program that extracts rules from raw text. The computer program is necessary to ensure that extracting rules from raw text is possible and sufficient in terms of the number of rules for the functioning of an expert system.

This paper consists of several sections. The next section describes the computer program and some aspects of its functioning. The third section reflects tests performed processing different texts. The conclusions are the latter.

II. COMPUTER PROGRAM

A computer program has been developed for rule extraction from raw text. That is, this program uses raw text to produce IF…THEN rules. The rules produced in this way can be used in expert systems. For example, Keyword Search Engine Enriched by Expert System Features is one of these expert systems. There are a lot of inference engines, where rules can be used, or it is possible to implement new inference engine21. So, there is no need to utilize the Semantic Web technologies to make information machine-readable in the Web, because there is a toolkit for obtaining all necessary from the Web.

The developed computer program looks like an ordinary window or form, in which there are several fields or text areas and one button. The first field is for entering the web address of the website. The second field is bigger than the first one, and it is for reflecting the refined text, which is obtained from the entered website. The third field is for showing IF..THEN rules, obtained from the refined text from the previous field (Fig.1.).

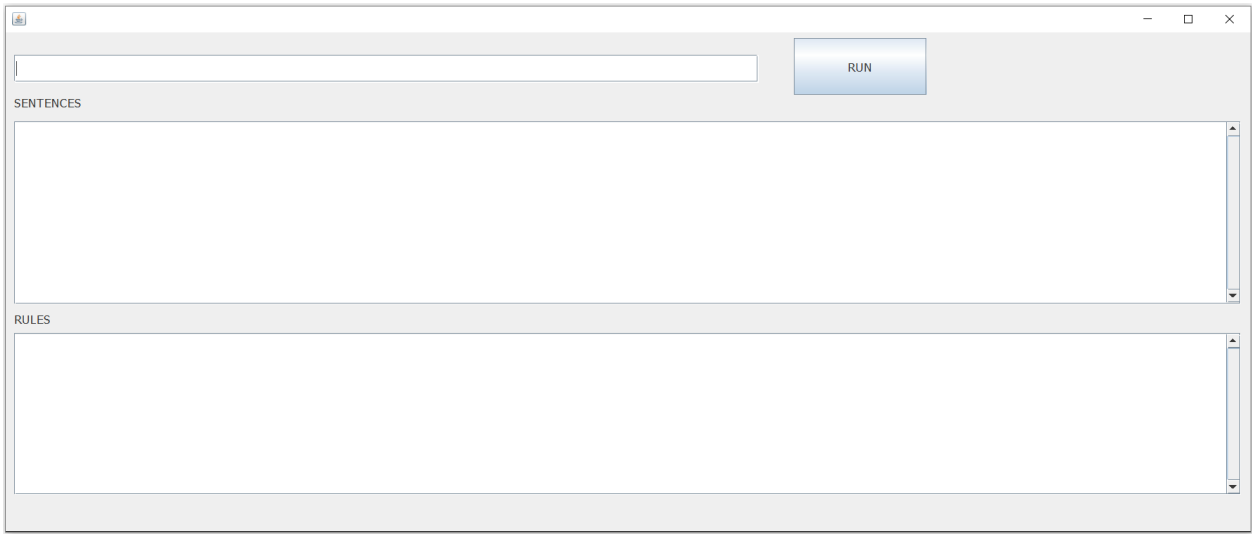

Fig.1. Main form of the program.

Here “refined text” means the text, obtained from the website, but without HTML tags. In fact, the whole text from the website is not used. The text between “p” HTML tags in the website is used. The text being divided into sentences before showing in the first text area. Apache OpenNLP library is used for dividing the text into the sentences. The same library is used for extracting rules from the sentences. The Java programming language had been chosen to implement the described program.

Extraction of four types of rules were realized in this program to demonstrate the ability of rule extraction from raw text. However there are a lot of other types that can be implemented. The only reason why other types were not implemented is that this is a demo program.

III. TESTS

Wikipedia pages were utilized for testing this program. The first web page from the Wikipedia was https://en.wikipedia.org/wiki/Sailing_ship, which was visited in May 24, 2021. The rules and sentences, from which these rules had been generated, are shown in Table I.

TABLE I. Rules and sentences from the first web page.

Rules	Sentences
IF ship THEN vessel	A sailing ship is a sea-going vessel that uses sails mounted on masts to harness the power of wind and propel the vessel.
IF was invented by Chinese. THEN compass	The compass was invented by Chinese.
IF predecessor of the galleon, THEN carrack	As the predecessor of the galleon, the carrack was one of the most influential ship designs in history; while ships became more specialized in the following centuries, the basic design remained unchanged throughout this period
IF brigantines THEN Baltimore clippers	Fast schooners and brigantines, called Baltimore clippers, were used for blockade running and as privateers in the early 1800s.

The second web page for testing the program is https://en.wikipedia.org/wiki/Car, which was visited in May 24, 2021, too. The rules and sentences, from which these rules had been generated, are shown in Table II.

TABLE II. Rules and sentences from the second web page.

Rules	Sentences
IF automobile THEN motor	A car (or automobile) is a wheeled motor vehicle used for transportation.
IF was challenged by Henry THEN patent	His patent was challenged by Henry Ford and others, and overturned in 1911.
IF was started by Ransom THEN production-line	Large-scale, production-line manufacturing of affordable cars was started by Ransom Olds in

	1901 at his Oldsmobile factory in Lansing, Michigan and based upon stationary assembly line techniques pioneered by Marc Isambard Brunel at the Portsmouth Block Mills, England, in 1802. IF was started by Ransom THEN production-lin
IF sector THEN contributor	The transport sector is a major contributor to air pollution, noise pollution and climate change.

In general, two tests are not sufficient for a complete program check, but these tests allow to make a preliminary assessment of the program. First, the experiments of rule extraction from raw text are considered satisfactory. Rules can really be extracted from raw text. Second, there are some complexities during the task have been performed. This is due to the fuzziness of human speech. For example, the first rule from the Table I is "IF ship THEN vessel". This rule is useful, but the sentence gives more defining information for rule. So that the rule could be the following "IF sailing ship THEN sea-going vessel". Or more expressive example from the Table II: "IF sector THEN contributor". This rule does not reflect the idea of the sentence. The better rule could be "IF transport sector THEN contributor to climate change".

V. CONCLUSION

Thus, despite the ability of extracting knowledge (rules are knowledge) from raw text, technical realization of this mechanism could be better. Here are several ways. The first one is improving the realized algorithm. This way is fully automatic and does not require human involvement. On the contrary, the second way requires human involvement. This way uses completely different principles in contrast to the first way.

Extracted rules can be utilized in question-answer system. However, utilizing the rules in expert systems is more interesting. Expert systems imply an inference engine. So, it is necessary either to implement own inference engine, or to use an existing one. The use of existing inference engine is preferable because of lower costs in time and money. One of Semantic Web inference engines could be such an inference engine. But there is a problem of rule transformation to OWL ontology. Fortunately, this problem has been partially resolved [2].

ACKNOWLEDGMENTS

This work has been supported by my family. This work and all previous works have been performed using the computer of the author.

REFERENCES

[1] Verhodubs O., Keyword Search Engine Enriched by Expert System Features, 2020.

[2] Verhodubs O., Mutual transformation of information and knowledge, 2016.