Rule Mining from Raw Text

Olegs Verhodubs

September 20, 2021

Rule Mining from Raw Text

Olegs Verhodubs

oleg.verhodub@inbox.lv

Abstract. Expert systems require knowledge for their work. Rule-based expert systems require rules that is one of the knowledge type. It is possible to generate rules from raw ontologies, however ontologies are not very widespread in the Web at this time. The solution may be to generate ontology from raw text and then to generate rules from the ontology. Thus, ontology generation is an intermediate link, but the presence of an intermediate link always increases the difficulty and decreases efficiency everywhere. It is necessary to find the way of knowledge extraction from raw text. This paper shows how knowledge in the form of rules can be mined from raw text.

Keywords: Rules, Natural Language Processing, Expert Systems

I. Introduction

It is difficult to appreciate the importance of experience in the life of mankind. Experience empowers mankind to avoid mistakes that have occurred in the past. There are personal and collective experiences. Personal experience is an experience that is accumulated by a person during his life, and collective experience is an experience that is accumulated by a certain community of people during the existence of this community. Personal experience is accumulated in the brain of the person acquiring it; in turn, collective experience usually requires additional auxiliary means for its fixation and the possibility of repeated use. Personal experience is accumulated by the person acquiring it only that is why it is less interesting then collective experience acquiring and using by many people. Auxiliary means for fixing and repeated use of the collective experience have been constantly changing. At first it were fairy tales, myths and parables. Using them, the older generation passed on the experience to the younger through oral speech. Then writing appeared and oral folk art including fairy tales, myths and parables migrated to material objects capable of fixing them. Writing made the experience of generations more durable, because the experience transmitted orally was often lost and distorted. Over time, as the population increased, the number of written sources increased and the recorded collective experience thus became more and more extensive and varied. Over time, there were so many written sources that one person was not able not only to familiarize himself with all of them, but also to know about the existence of all written sources. The quantity has grown into a new quality. That is, the potential was formed, which in the end was embodied in the Web. The development of the Web also had its stages. Initially, the Web was a collection of information resources, mainly websites. Then themed lists of websites appeared, where you could select the category of interest and see the websites related to this category. However, the development of the Web has been continued and themed lists of websites no longer satisfied users. That is why keyword search engines appeared. At present time, they also no longer satisfy users. One of the main reasons why this happens is that the result of the keyword search engines is presented in the form of many links to websites that have to be visited to find the information you need. This means that a user has to browse, analyze and aggregate information scattered across many websites in the Web, even though such a routine work could be done by search engines, based on new principles. Here new principles are Semantic Web Technologies. Such technologies promise users to make the Web more machine-readable, which will allow them to collect information together from various sources and even infer based on the collected information. One of the systems with similar capabilities is the Semantic Web Expert System [1]. The Semantic Web Expert System and other similar systems utilize OWL (Web Ontology Language) ontologies, but currently ontologies are not widespread in the Web. The most of information in the Web is still presented in the text form. As mentioned above, texts are less machine-readable than ontologies, but it is possible to level this drawback. To this effect, the rules need to be generated from the texts. Here texts are ordinary raw texts, which the Web clogged with. Some experiments in extracting rules from raw text have already been done and have been described in [2]. The rules generated from the text are used as raw materials for intelligent systems. Being developed Keyword Search Engine Enriched by Expert System Features [3] is such an intelligent system. Technically, this computer system is being designed to endow existing keyword search engines with new properties as aggregating knowledge from the Web and inferring. Semantically, being developed intellectual system allows to manipulate the collective experience of mankind, using or disabling certain knowledge of an individual or a group of people. In theory, it will be possible to model and implement the style of reasoning of such famous people as Napoleon, Tesla or somebody else.

This paper consists of several sections. The next section discloses non-obvious possibilities for identifying rules in raw text. The third section answers the question of how to generate rules from raw text. The last section is conclusions, where ideas stated in the paper are summarized.

II. Hidden opportunities

There are many definitions for concepts such as data, information and knowledge [4]. Naturally, each definition is subjective, but a set of subjective definitions of a certain object forms true and integral idea of this object. Analyzing various definitions, it is possible to come to conclusion that data are facts or better to say derivation rules without premises, information is a set of facts with a common context, knowledge is a set of dependencies among facts obtained from information. Knowledge can be different, namely domain, tacit, common sense knowledge and so on [4]. In addition, knowledge can be contradictory and meta-knowledge helps to resolve these contradictions. By the way, meta-knowledge are knowledge about knowledge, but meta-knowledge as types of knowledge is not the topic of this research. Forms of knowledge representation are also different, but we are interested in knowledge in the form of rules. Each rule consists of premise and conclusion. It is also possible that there are several premises, or several conclusions in the rule. Usually, rules are represented as constructions in the form "IF … THEN …", where premises follow the "IF" and conclusions follow the "THEN". Writing a rule in the form of "IF...THEN..." means that the conclusions are true only if the premises are true.

It turns out that rules can be generated not only from OWL ontologies [5], [6], but also from raw text. A series of experiments on extracting rules from raw text had been carried out and partially described in [2]. The Java programming language and also Apache OpenNLP library was used in order to generate rules from raw text. The programmed algorithm of rule generation from raw text performs the task following the human logic, that is, it analyzes parts of the text for the correspondence of premises and conclusions. Pieces of text and the rules guessed from them are discussed below. First, simple or obvious cases of extracting rules from raw text are being considered. More complex cases are being considered further.

Table 1. Rules from text pieces.

Nr. p.k.	**Part of Text**	**Rule**
1	If there is no God, everything is allowed.	IF no God THEN everything is allowed
2	The holiday is the unity of the community.	IF holiday THEN unity of the community
3	At the heart of every challenge is opportunity.	IF challenge THEN opportunity (IF there is challenge THEN there is an opportunity)
4	I don't believe in action without a	IF action THEN purpose

	purpose.	(IF there is action THEN there is a purpose of this action)
5	Anyone who cannot speak will not make a career.	IF can speak THEN will make a career
6	After a breakthrough, there should always be time to digest the situation, to adapt.	IF breakthrough THEN should be time to digest the situation, to adapt
7	Love is the only force capable of transforming an enemy to a friend.	IF love THEN the only force capable of transforming an enemy to a friend
8	He did not win a single battle, but the way in which he retreated, inflicting maximum damage on the enemy, earned him respect among the British, and even Sir Arthur Wellesley himself said that he had no more worthy opponent.	IF did not win a single battle, but retreated, inflicting maximum damage on the enemy THEN earned respect

Various quotes from famous people are listed in the table in the "part of the text" section. It makes no sense to name the authors of these quotes, because our task is to understand how rules can be generated from (raw) plain text. Many more examples of quotes could be cited from which it is possible to generate rules, but this is not necessary. It is important to understand that there are two main ways to generate rules from raw text:

1) using meanings in the piece of text,

2) using the internal structure of the piece of text.

Among the mentioned quotes, the last one is most suitable for the first way of rule generation. This would be especially true if instead of “earned respect” the quote would contain something like “earned awe”, but generated rule would still contain “earned respect”. Here meanings and their transitions are followed in order to generate rules. This is a difficult way, or in other words, this way is more difficult than the second one. The second way can be called mechanical. This is so, because this way gives an opportunity to generate rules from raw text without thinking about the meanings of its constituent parts. We are talking about a certain scheme in a sentence that can automatically be converted into a rule. For example, the first quote is a ready-made rule. One more typical example of rule generation is observed in the second and in the third quote. Both cases contain the construction “is a”. Rules from this construction can be generated mechanically or automatically. The same applies to the case described in [2]. That is, the sentence

“The compass was invented by Chinese” (I)

can produce the following rule:

“IF was invented by Chinese THEN compass”. (II)

Of course, the words “invent” and “by Chinese” are not important for rule generation here. The passive voice is much more important. For example, such a sentence with passive voice can provide us with a rule, too:

“Chess was created by Indians”. (III)

This piece of text can give us this rule:

“IF was created by Indians THEN chess” (IV)

One more scheme for rule generation is the construction of the sentence, where is used the word “called”. For example, the sentence

Fast brigantines, called Baltimore clippers, were used for blockade running (V)

can be transformed to the following rule:

IF Baltimore clippers THEN fast brigantines (VI)

Some sentence structures are suitable for generating identifying rules. Here identifying rules are such rules, which identify an object or subject by its properties. For example, this sentence

A green apple is on the table (VII)

is suitable for generation such a rule:

IF green THEN apple (VIII)

It is possible to generate an identifying rule from the sentence, where are several adjectives. For instance, the sentence

An oak is tall and mighty (IX)

can provide us with the rule like this:

IF tall and mighty THEN oak (X)

Identifying rules can be generated from various syntax expressions in the sentence. One of such expression is “consist of”. For example, the sentence

A plane consists of wings and engine (XI)

can serve as material for generation this rule:

IF wings and engine THEN plane (XII)

There are other syntax construction with the similar possibilities. For instance, the “part of” construction fits for generation identifying rule, too. Eg, the sentence

A house is part of town (XIII)

is transformable to the rule like this:

IF house THEN town (XIV)

Another syntax construction does not fit to generate identifying rule. Rather, this syntactic construction fixes the cause-and-effect relationship in a visual way. This is about constructions, where “because” word is present. For instance, the sentence with “because”

He is clever because reads a lot (XV)

is transforable to such a rule:

IF reads a lot THEN clever (XVI)

Interestingly, there are cases, when one sentence is enough to generate several rules at once. For example, the sentence (V) in addition to the above-mentioned rule may be served as a material for generation one more rule:

IF were used for blockade running THEN Baltimore clippers (XVII)

So, the idea of rule generation from raw text was examined in detail. Certainly, there are plenty of other cases, when raw text is transformable to “IF....THEN” rules, but it is not necessary. It is more important to understand how the generation of rules from raw text can be implemented.

III. Ways of implementation

There are several ways to generate rules from raw text. The first way has already been partially implemented and described in [2]. This way is based on the use of Apache OpenNLP library [7] for Java programming language. More precisely, only some of the possibilities of mentioned programming library are used to generate rules from raw text. The main of used possibilities is splitting a piece of raw text into parts of speech. Splitting a piece of raw text into parts of speech is the first step of the first way of rule generation from raw text. Finding a keyword in a piece of text is the second step in generating a rule. Certainly, keywords here are different for different kinds of rules. For example, for the sentence (XV) we would look for a keyword "because". In turn, at the last step the environment of a keyword is being analyzed and a rule is being assembled in final form. This way of rule generation is mechanistic in the negative sense of this word. The algorithm for each specific type of rule should take into consideration plenty of possible variations for approximately the same semantic content. For instance, the sentence (XV) may have a slightly different form:

He is clever because reads a lot of books (XVIII)

Or something like this:

He is clever because he reads a lot (XIX)

It is possible to invent many other variations of this sentence, and this happens with every type of rule that is generated. There are a lot of variations and every feature has to be programmed. It is not always effective that is why another way of rule generation is necessary.

Probably, there are still unexplored technologies of universal classification, but only artificial neural networks are available to us. Presumably, it is neural networks that can help us to generate rules from raw text. This is where a large field of research begins. However, some assumptions about the direction of research can be made already now. For example, it is possible to assume that parts of speech and also the distance from the keyword to other elements of the being generated rule are main criteria that have to be used in rule generation by means of artificial neural networks. Besides, an expert has to train an artificial neural network including different sentences or other fragments of raw text. Different sentences means sentences with different size, different structure and different keywords. In turn, keywords are such words as "because", "part of", "consist of" and many others, that is, such words, which are determinant for choosing one or another rule generation algorithm. Thus, this way is a realization of supervised learning.

IV. Conclusion

This paper clarifies the idea of rule generation from raw text. The examples of text fragments and possible rules from them are showed in the paper. Two different ways of rule generation from raw text are mentioned. One of these two ways has already been partially implemented earlier. The second way was briefly described, however this description is rather sketchy, and it is clear that there are a lot of work here.

Rules generated from raw text are necessary for using in keyword search engines with expert system functions. Functions of expert systems are not only answers to questions, but also inferring based on user answers and rules. One of such systems is a system described in [3].

The predecessor of the Keyword Search Engine Enriched by Expert System Features is a Semantic Web Expert Systems, which is described in [1]. The difference between these systems is that the Semantic Web Expert System generates rules from OWL ontologies, but Keyword Search Engine Enriched by Expert System Features generates rules from raw text. The necessity of Keyword Search Engine Enriched by Expert System Features occured, because nowadays ontologies are not widespread in the Web.

Though there are still a lot of work in the development of the system, main traits of Keyword Search Engine Enriched by Expert System Features are clear and all of them can be implemented within the commercial or community project, but for this purpose funding is necessary.

Acknowledgments

This work has been supported by my family.

References

[1] Verhodubs O., Grundspenkis J.: Towards the Semantic Web Expert System (2011).

[2] Verhodubs O.: Experiments of Rule Extraction from Raw Text (2021)

[3] Verhodubs O.: Keyword Search Engine Enriched by Expert System Features (2020)

[4] Verhodubs O.: Mutual transformation of information and knowledge (2016)

[5] Verhodubs O.: Grundspenkis J. Evolution of ontology potential for generation of rules (2012)

[6] Verhodubs O.: Ontology as a Source for Rule Generation (2014)

[7] Apache OpenNLP: https://opennlp.apache.org/, accessed 23.07.2021