

A Text Instantiation Method for Knowledge Extraction from Scientific Research Texts

Shiliang Wan^{1, a}

¹ The Chinese People's Liberation Army Unit 91550, Dalian, 116023, China

^aemail: wanshilian_DL@163.com

Keywords: Scientific research text; Instantiation; Knowledge; Extraction

Abstract. Text instantiation is the key step in the research of text knowledge extraction. This paper has compared the rule method with the statistical method. Based on the method of rules, the basic framework of processes instructing research text instantiation is established, the problems that are needed to be solved in the four stages in the process are analyzed, and the corresponding solutions for each stage are proposed.

Introduction

Text instantiation is a key step in the knowledge extraction of scientific research text, which chooses the pre-processed XML intermediate text as research subject [1]. At present, the text instantiation method includes based statistics, based rules. This paper analyzes the advantage of model matching method based on rule instruction over that based on statistical guidance so as to build a semantic sequence extraction model which could provide a feasible basis. Building on the method of rules, we establish a scientific text instantiation of the basic framework of the process, analyze the process in vocabulary sequence boundary demarcation, in semantic acquisition and screening, and the problems in the four stages of text instantiation. We also design the corresponding solutions for each stage and give the examples.

Method comparison between statistical method and rule method

The statistical principle is used mainly in the statistical method to obtain the guidance resources through the training of the term characteristics of large scale text, which is used as a reference for knowledge extraction in the new text [2]. This method need not worry about difficult situation caused by the complexity of the grammar, vocabulary, semantic of the text. But because of the expansion of the text size, we gain access to more real term characteristic of statistical rules [3]. But its disadvantage is that the absolute accuracy can not be guaranteed. The knowledge obtained by the guiding method to extract knowledge, from a logical point of view, is based on probabilistic conclusion pertaining to the principle of induction. These fields demand high accuracy for knowledge extraction such as scientific research, military, security and etc. However we should take cautions to the actual state of this knowledge extraction.

Rule method for the extraction of text knowledge and knowledge is of a high degree of accuracy, which can not be rivaled by the statistical method group so far [4]. As long as the objects in the text conform to the rules, it is absolutely correct to extract the knowledge we need without having to worry about the possibility of the error of knowledge [5]. For example, if in pattern matching method, the text of the object matches with a model, it can, based on model set position, obtain concept of knowledge from the object. According to the structure characteristics of the model, we could extract the relationship between knowledge. The major problems in the rule method are its low matching level when the rule methods are used to process complex grammar, free word, semantic ambiguity of text. So the rule method fails to perform effectively in knowledge extraction in common text. However, under the circumstance of small scale, professional vocabulary and grammar, the rule approach can be very good for the extraction of text knowledge. As a result, the rule method is superior to the statistical method in the research of the knowledge extraction of

scientific research texts.

Lexical sequence boundary division

A large number of statements are described by the vocabulary of scientific research texts. To complete the ontology of the text, the text must first be divided into the sequence of words, to ensure that each sequence only describes an effective statement [6].

The grammar of the text is strictly regulated, and the boundaries of each statement are classified by means of special markers. In this paper, we use the method based on the extensible symbol dictionary to divide the boundary and the level. Through the establishment of the boundary symbol dictionary, the boundary of the lexical sequence is defined, and the representation of the sequence is determined by the hierarchical structure of the ontology knowledge base, which ensures the integrity of the ontology of the statement. In areas such as equipment index boundary level symbol dictionary of symbol boundary ★、◆、◇、▽ are set before and after the boundary specification and the level, which makes clear partition boundary of text sequences of words, and clear hierarchy.

The following is the research contents of the text:

★Tactical index

◆Operating S wave band

◆Detecting range

◇Instrument measuring range:

▽Maximum instrument measuring range is equal to 300 km

▽Minimum instrument range =0 km

◇Antenna speed is 6 RPM / min

◇False alarm probability Pf=60%

With the support of the boundary symbol dictionary, the word sequence set is as follows:

Word sequence1: ★Tactical index //the first level

Word sequence2: ◆Operating S wave band //the second level

Word sequence3: ◆Detecting range //the second level

Word sequence4: ◇Instrument measuring range: //the third level

Word sequence5: ▽Maximum instrument measuring range is equal to 300 km //the fourth level

Word sequence6: ▽Minimum instrument range =0 km //the fourth level

Word sequence7: ◇Antenna speed is 6 RPM / min //the third level

Word sequence8: ◇False alarm probability Pf=60% //the third level

Semantic access and screening

The acquisition of lexical semantic meaning in scientific research texts is the key to the mapping of ontology and text in the process of text knowledge acquisition. According to the corresponding characteristics of the multi word meaning of scientific research text, the mapping of lexical and semantic can be realized by using the word semantic mapping table based on the model of synonyms. Now we take the equipment index field as an example to construct lexical semantic mapping table. By dictionaries on the equipment standard and military standard, we could determine the scope of the vocabulary. Under the guidance of experts in this field, we could compare the related vocabulary with ontology domain in semantic field. Table 1 is part of the mapping table.

Semantic sequence not only describes a complete statement, construction of key semantic ontology instances, but also some pure tone word semantic. These auxiliary class semantics can not affect text instantiation when their function is to avoid context ambiguity so that the semantics of the text should not be distorted. So it is necessary to do a screening work on the semantic sequence after the boundary division, so as to lay the foundation for the next step to extract semantic and semantic association based on the semantic sequence extraction model.

Tab.1. Semantic mapping example on equipment's indicatory domain

vocabulary	sematics
Level side lobe level	Horizontal antenna side lobe level
Vertical side lobe level	Vertical antenna side lobe level
Average emission side lobe level	Average transmit antenna side lobe level
air transport	Transport aviation
Railway	Railway transportation
Highway	Road transportation
Water transportation	Waterway transportation
Transport method	Type of transportation
Scanning method	Scanning mode
Maximum bending diameter	Maximum turning radius
FCR	BITE fault coverage
FDR	BITE fault detection rate
FIR	BITE fault isolation rate
Bite fault coverage	BITE fault coverage
Bite fault detection rate	BITE fault detection rate
Bite fault isolation rate	BITE fault isolation rate

By preprocessing the resources of the part of speech, we could tag each word. And by applying the existing "class" concept in the ontology library, we can select the semantic and retain the necessary semantic resources. The screening method for semantic is described as follows:

The first step, search the semantic search object in the ontology library. If the ontology library contains the label, then retain the key semantic, if not, then enter into the second step.

The second step, filter the semantic parts of speech. If filter conditions are met, then the semantic parts are to be retained, if not, remove them.

The main function of part of speech filter is to set up the filtering condition of the semantic aspect. According to different needs of different areas, the user can set the parameters of part of speech filter. For example, we could use quantitative words as the qualified words, adjective words and auxiliary words "as necessary words to eliminate the part of speech.

Ontology instantiation

After the division of boundary of sequences of words, semantic acquisition and screening, we construct semantic sequence extraction model in order to extract the key concepts and their conceptual relations. By this way, we obtain specific content for the instantiation of ontology.

When text grammar rules are simple and syntactic structure is logic, it provides a convenient way to extract the relationship between the concept and the concept of the syntactic structure model of scientific research in the specific domain.

To achieve the sequence of semantic ontology instantiation, one should realize the first division of the binary sequence boundary. Then the three elements in the sequence of semantic unit are converted into the elements of corresponding elements in series model. Then the model of three sequences is matched. After the success of matching, the model is marked with the location of the

extracted semantic element sequences in the corresponding position in the concept of knowledge according to the sequence elements. According to structural characteristic of sequence element model, we obtain the relationship of semantic element knowledge. Finally, we will make the knowledge as specific content related to ontology corresponding instantiated knowledge. We use appropriate structure to describe the relationships between instances, based on the semantic sequence model in three sequence element composition.

Based on the relative position of the three elements and the semantic set of the boundary feature of the meta sequence, the relative position of the element sequence helps to determine the selection order of the feature set of the boundary feature, and the semantic set of the boundary feature helps to determine the specific boundary location. The semantic focus of the boundary feature is mainly related to the feature semantics, which predicates word meaning, and the three kinds of attribute semantics. All of these are labeled with the front and back boundary markers. The boundary elements in the semantic sequence will correspond to these three kinds of semantic types. As long as the judgment is read and one of the semantic matches, it can be based on the initial set to determine if the element should bear the boundaries of responsibility.

The relative position of the element sequence is determined by the finite state machine according to the model of the semantic sequence. As shown in figure 1.

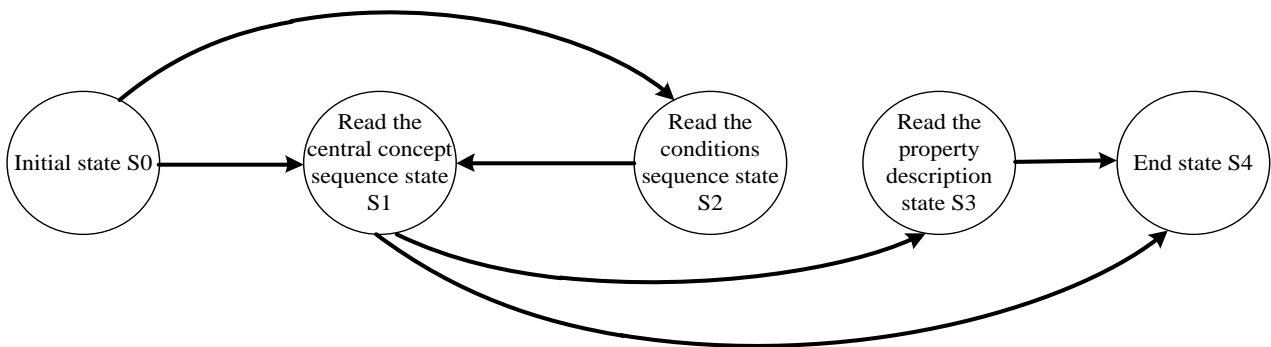


Fig.1. Semantic sequence's finite status machine

For example, the equipment index in the field of scientific research in the text of a semantic sequence "in / for / plane // discovery probability / equal to /50%/ / / /, / maximum detection range / more than or equal to /210/ km /, element sequence boundary is divided. First, based on the finite state machine over a sequence of read, we should locate the existence of semantic unit in the constraint sequence in the boundary feature semantic database, and we can find that the "in" and "situation" is in line with each other, which can be used to identify whether the "in / for / plane // discovery probability / equal to /50%/ / situation" is the constraint sequence; then we should judge the position of the sequence of central concepts, due to the "maximum detection range / more than or equal to /210/ km /." "Greater than or equal to" belong to the predicate word meanings, which is present in the boundary feature semantic database, center concept sequence boundary has also been confirmed, the sequence of central concepts for maximum detection distance; then attribute description sequence means apparently "is more than or equal to /210/ km".

In different fields, the specific semantic content of each element in the meta sequence model is different. When the sequence of semantic elements is matched with these models, the correspondence between the semantic units and the model elements depends on the mapping of the classes or attributes in each domain ontology.

In order to enable semantic element sequence and sequence element model to match, we must first establish a mapping table and various fields of ontology and a collection of sequence model elements. Then we proceed with knowledge extraction in the text of the corresponding areas to choose the appropriate mapping table, and the element sequence of each semantic unit will be mapped to the elements of the model, resulting in the achievement of the model matching.

In order to read the semantic unit, the matching process uses the finite state machine to judge the semantic sequence and the model matching, and obtains the related concept and the correlation relation from the sequence. The corresponding "classes" and "attributes" are also established by

using these concepts and related relationships in the ontology library, and construct the hierarchical relationships between them.

Conclusion

Text instantiation holds a key position in the research of knowledge extraction from scientific research texts. The two current mainstream in text instantiation method, matching method based on rules to guide model and statistical guidance method, are compared with each other on the respect of their respective advantages. The paper has pointed out that the method based on rules is more suitable for scientific text knowledge extraction, so it offers a theoretical and feasible basis for building a semantic sequence extraction model. The paper talks about the establishment of the basic framework of a scientific text instantiation, analyzes the process in vocabulary sequence boundary demarcation, in semantic acquisition and screening, raises the problems that is requested to be solved in three stages of ontology instantiation, and designs the corresponding solutions for each stage and gives the examples.

References

- [1] Witte R, Li Q, Zhang Y. Text mining and software engineering: an integrated source code and document analysis approach [J]. The Institution of Engineering and Technology, 2008, 2(1):3-16.
- [2] Xu Lin. A review of the recently funded project of the National Natural Science Foundation of China in the field of Natural Language Processing [J]. Journal of software, 2005, 16 (10):1853-1858.
- [3] Ma Jin. Analysis and study of Chinese dependency parsing based on statistical methods [D]. Harbin: Harbin Institute of Technology, 2008.
- [4] Uren v s, Cimiano p, Iria j. Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art [J]. Journal of Web Semantics, 2006, 4:14-28.
- [5] Xiong Yunbo. Research on some key technologies of text information processing [D]. Shanghai: Fudan University, 2006.
- [6] Nadeau d, Sekine s. A Survey of Name Entity Recognition and Classification [J]. Lingvisticae Investigationes, 2003, 30: 1-20.