A Design of Knowledge Extraction and Consistency Checking System for Scientific Research Texts

Feng Gao[1, a]

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

[a]email: gaofeng_DL@163.com

Keywords: Scientific Research text; Knowledge; Extraction; Consistency; Checking

Abstract. This paper addresses the practical problems involved in the process of knowledge extraction and consistency checking system for scientific research texts. Based on the theoretical research on knowledge extraction and reasoning for scientific research texts, this paper presents a design of knowledge extraction and consistency checking system. This system can perform the knowledge extraction of scientific research text and represent the four tuple form of ontology instance. It can run the consistency checking of the corresponding knowledge in any two texts as well.

Introduction

In most of the specific areas of knowledge of the scientific research, each subject in scientific research is a specific description of a particular field of knowledge [1]. The relationship between subjects reflects the relationship between knowledge in the field. The organization of these knowledge together constitute the complete domain knowledge system structure [2, 3]. In the real work, we often need to check the consistency of the content of the whole or that of different texts to ensure that there will not be conflicting ideas. A complete and consistent knowledge system is a collection of efficient and useful knowledge. If knowledge set contains the problems of the content inconsistency, it must negatively affect the related field work, which could lead to errors and even influence knowledge expansion and derivation [4]. In the past, the checking of the text is realized by the way of manual comparison, which is of low efficiency, long time and high cost. Later, the checking work of knowledge is improved by using the knowledge extraction and comparison of the text based on the key words [5]. However, the relationship between knowledge could be not simply revealed by the key words matching on vocabulary level. For example, when the description of equipment index one: "maximum detection distance is greater than 100 km" is compared with that of index two: "maximum detection range is 130 km", the index two description expression semantics qualifies the logical constraints in the semantic description of index one, which results in content consistency without any contradiction. However, if we only use the keyword matching method, it is easy to draw the conclusion that the two descriptions are inconsistent. Thus it can be seen that the potential relationship of knowledge in terms of classification and the inherent correlation between knowledge based on first-order logic are not well utilized [6]. Therefore, the efficiency of knowledge checking needs to be improved [7].

Based on the research on the theory of knowledge extraction and reasoning, this paper designs a knowledge extraction and consistency checking system. According to the system, it can realize the knowledge extraction of scientific research text and the represent the four tuple form of ontology instance. It can also realize the consistency checking of the corresponding knowledge in any two texts.

System description

Procedures in the operation of the system:

1. The system is able to import the domain ontology library, which is constructed by hand, to realize the knowledge extraction and materializing of ontology instance of the scientific research

text through the support of the domain ontology.

2. IRLAS system will be used to pre-process the generated XML intermediate text as an application object. The design of the interface between the two intermediate text objects will select one as the main text and the other as a text from the text.

Although there is no difference in methods and goals between knowledge extraction and structuralization of knowledge, we should target the main text examples of the four tuples as the main object of verification in the knowledge consistency check. Because the different positions of the logical inference rule base in the implementation of the corresponding rule reasoning will produce different results. If the comparison of the main text with the text qualifies this rule, then consistency checking is completed. But once the two positions were reversed, it obviously does not meet the rule. Then the result is quite the opposite.

3. This system has designed the semantic element sequence model element in the field of equipment index and concept mapping table of the domain ontology. The semantic sequence model, element sequence model matching algorithm comprise the basis of the text knowledge extraction and representation. Through the combination of ontology technology, the system can successfully extract and represent the knowledge from the text.

4. This system contains the functions of consistency checking to the main text and sub-text. Through the example based on the logical rules of the reasoning mechanism, it could successfully implement the checking of the knowledge of scientific research.

Knowledge extraction function design

The realization of the function of knowledge extraction, the system design process is as follows:

1. Select the pre-processed text of IRLAS system as the initial object.

2. By boundary symbol library, the lexical sequences are divided into boundary and the lexical sequences are categorized in terms of vocabulary level in the text.

3. Select semantic content through synonyms thesaurus.

4. Filtering semantics by part of speech filter and ontology concept library.

5. Through the construction of semantic sequence model base, we divide it into three element sequence boundary by finite state machine algorithm and then make the three element sequences as the partition boundary of sequence. Semantic database of the three element sequences boundary are used as the basis of the judgment of semantic boundary in accordance with the order of boundary division.

6. Based on the concept of knowledge and the extraction of association relation, the concept of matching in ontology library is used to search for the concept of matching.

7. Through the examples of the four tuples represent knowledge from semantic sequence extracted instantiated, we respectively put three sequences extracted knowledge into corresponding elements in. The position of these elements in also reflects the non-taxonomic relationships in knowledge.

The flow chart of extraction function design is shown in Figure 1.

S0 state: pre-processed XML text.

S1 state: the text is divided into the sequence of lexical sequences.

S2 state: get the initial semantic sequence of lexical semantics in the synonym mode.

S3 state: through the filter to get the semantic sequence.

S4 state: the meta sequence is obtained by the semantic sequence model.

S5 state: the ontology concept is based on the meta-sequence specific domain model.

S6 state: using the four tuple structure instance set to realize the structured representation of knowledge.

Knowledge consistency checking function design

In order to realize the knowledge consistency checking function, the system constructs the logic rules algorithm library, logic rules interface library. A reasoning process of access object inference

and reasoning process of reasoning objects were given to the various elements of the four tuple structure of case knowledge according to their positions.

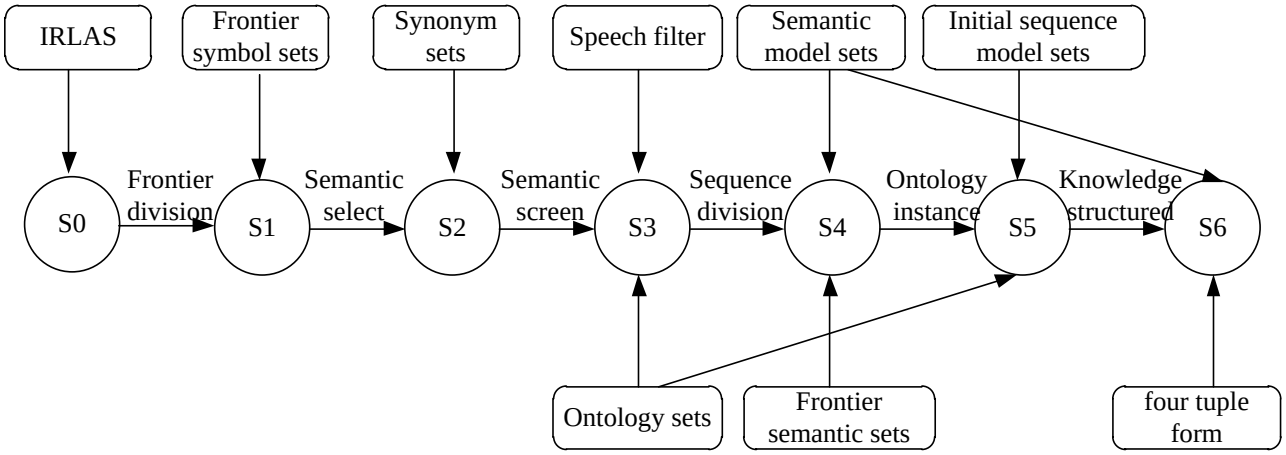

Fig.1. Flow of the function of Knowledge extracting

The main procedures of knowledge consistency checking are as follows:

1. Choose the main text and examples from the text, these examples have been achieved OWL format storage.

2. Based on the location of the instance in the four tuple structure, the inference object is extracted.

3. The selection of the access inference rule, based on the reasoning object in the logical rules of the interface library to obtain the rules of the portal address, in the logical rules algorithm library to get the rules algorithm.

4. Access inference.

5. Reasoning objects in the reasoning process based on the location of the instances in the four tuple structure.

6. The inference rule is selected, and the logical rule algorithm is obtained and the consistency check is obtained in the logical rule algorithm library according to the inference object in the logical rule interface database.

7. Inference of participating objects based on logical rule algorithm.

Figure 2 Knowledge reasoning flow chart:

S0 state: an example of text owl format generated by API module.

S1 status: Based on the four tuple structure the corresponding examples and examples of attributes will be access to reasoning objects.

S2 state: access inference rules are obtained by the support of logical rule interface library.

S3 state: access reasoning leads to the results of reasoning and true results continue to S4 state.

S4 state: according to the four tuple structure, the corresponding instance and the instance attribute are used as the consistency reasoning object.

S5 state: by the aid the logical rule interface library to obtain the consistency of the rules of inference.

S6 state: the inference results obtained by consistent inference.

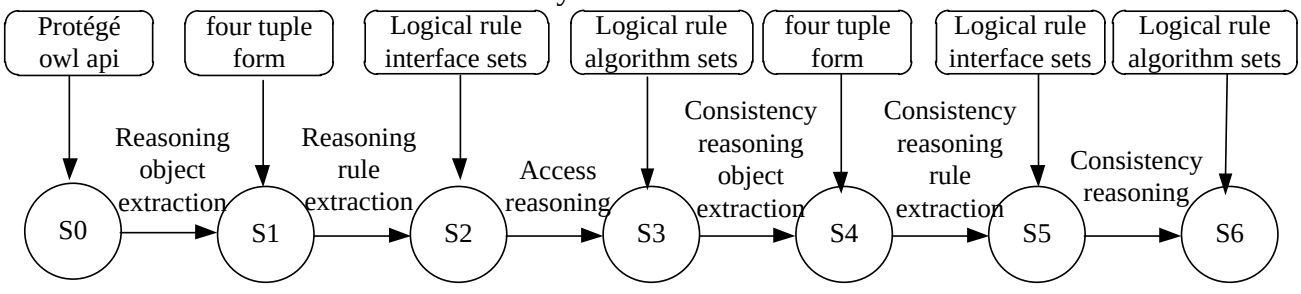

Fig.2. Knowledge reasoning flow chart l

Conclusion

In most of the specific areas of knowledge of the scientific research, each subject in scientific research is a specific description of a particular field of knowledge. The relationship between subjects reflects the relationship between knowledge in the field. The organization of these knowledge together constitute the complete domain knowledge system structure. Thus it can be seen that the potential relationship of knowledge in terms of classification and the inherent correlation between knowledge based on first-order logic are not well utilized. Therefore, the efficiency of knowledge checking needs to be improved. This paper addresses the practical problems involved in the process of knowledge extraction and consistency checking system for scientific research texts. Based on the theoretical research on knowledge extraction and reasoning for scientific research texts, this paper presents a design of knowledge extraction and consistency checking system. This system can perform the knowledge extraction of scientific research text and represent the four tuple form of ontology instance. It can run the consistency checking of the corresponding knowledge in any two texts as well.

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] Popov b, Kiryakov a, et al. KIM-Semantic Annotation Platform[C]. Proceedings of the 2nd International Semantic Web Conference (ISWC 2003). Berlin: Springer, 2003: 484-499.

[3] Zhang Huipeng. Research and implementation of Chinese lexical analysis technology [D]. Harbin: Harbin Institute of Technology, 2006.

[4] Vargas vera m, Motta e. A Tool for Automatic Support on Semantic Markup. KMI Technical Report, TR Number133, 2003.

[5] Huang rh, Riloff e. Inducing Domain-specific Semantic Class Taggers from Nothing[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, USA: Association for Computational Linguistics, 2010: 275-285.

[6] Liu Dongxu. Chinese word segmentation and part of speech tagging in natural Chinese [D]. Chengdu: University of Electronic Science and technology, 2003.

[7] Li s, Xia r, Zong cq, et al. A Framework of Feature Selection Methods for Text Categorization [C]. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. Singapore: ACL and AFNLP, 2009: 692-700.