Study and Development of Question Answering System based on Ontology Query

Xiaoqiang Liu, Zhenbo Guo
Department of Information Engineering College
Qingdao University
Qingdao, China
xiaoq95a@163.com, gzb@qdu.edu.cn

Kaixi Wang, Wenxu Jiang
Department of Information Engineering College
Qingdao University
Qingdao, China
kxwang@qdu.edu.cn, xuwenjianghlg@163.com

***Abstract*—With respect to the ontology knowledge representation of refrigerator, a question answering system was designed based on ontology triple structure query. The ontology knowledge representation of refrigerator is built with representing the refrigerators' instance, attributes and values with the subject-predicate-object triples format. We analyze the natural language question, extract key words, analyze the dependencies between keywords, and build the query for ontology triples. The question answering system is implemented based on the Jena API and supports natural language querying OWL ontology. The results showed that, compared with Yang Tianqi's method, this method improve recall ratio and precision by 4.96% and 3.81% respectively.**

Keywords—domain ontology; natural language question; semantics understanding; question and answering (QA)

I. Introduction

Ontology can clearly express the relationship between things and their relationship, plays an important role in the semantic analysis of the automatic question answering system. Ontology in question answering system is mainly used to calculate the distance of concepts, methods WordNet meaning analysis of keyword similarity calculation[1], to extend keywords. This approach extends the semantic information of a single keyword, but ignored the relationship between keywords, resulting in omission user statement semantic information.

In recent years, with the rapid development of Linked Open Data, more and more companies will put these data in the Semantic Web with URI and RDF format, building the association with other data sources. The standard approach to query RDF is using structured queries in triple-pattern-based language like SPARQL, but only professional programmers are able to precisely build a structured query triples. as for ordinary users, the only choice to query knowledge base is by keywords or natural language sentences.

Integrated the above two points, we design and implementation a method of querying RDF format data by natural language, thus implement the question answering system. We design ontology to express formula information, using Stanford Dependency analyzing the user's natural language question, build a triple query, returned to the user a precise answer.

II. PREVIOUS WORK

At home and abroad in recent years, many natural language query ontology question answering system were designed. Pythia[2] and ORAKEL[3] based on formal grammar, assign each lexical unit a syntactic and semantic expression, the expression of vocabulary has been consistent with the ontology of vocabulary, through calculation of the grammar portion of combination of semantic questions the semantic information of the whole. Aqualog and PowerAqua[4] system map the language structure to a compatible with ontology semantic structure, the subject of natural language matches the knowledge base of the elements in question and the predicate object. FREyA[5] name after feedback, refinement and Extended vocabulary Aggregation. FREyA and Querix[6] both interact with the user to further confirm the user's query information, carries on the deep syntactic analysis, in order to eliminate ambiguity. Yahya etal.[7] based on an integer linear program, mapping the phrase of natural language question to the class, individual, and properties of the knowledge base , constructs the SPARQL structured query, in this process to achieve entity disambiguation and predicate disambiguation jointly. In the domestic, Zhang Zongren and Yang Tianqi[8] proposed a method to translate natural language into SPARQL query, using the Stanford parser do the syntactic analysis of user's natural language question, construct three tuple SPARQL query. Xu Kun and Feng Yansong[9] put forward the concept of semantic query graph, transform the natural language into SPARQL queries.

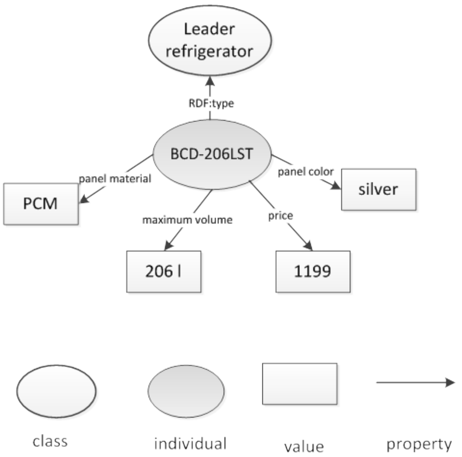
Our method has two different with previous work. Firstly, the study of the question answering system in foreign countries is focus on English text. The spaces between words simplify the segmentation process, while in Chinese sentence words are closely linked. There are many English corpora, such as WordNet, providing support for the analysis of the semantic; by comparison, Chinese corpora like TongYiCi CiLin and HowNet support for the analysis of the semantic. In English grammar has rules of the changes in the word forms, it is easy to discern the verb as the predicate, and a noun as subject or object; while in Chinese a same word can played as a noun, a verb or adjective, without any inflexion. Previous work query the knowledge base of the open field, they study a general method to transform a natural language question to a structural query; in our study, we design and implementation a method of querying RDF format data by natural language, thus implement the question answering system. We design ontology to express formula information, using Stanford Dependency analyzing the user's natural language question, build a triple query, returned to the user a precise answer.

III. APPROACH OVERVIEW

A. The design of ontology data

Ontology can clearly express the domain knowledge. RDF(Resource Description Framework) is a standard model for ontology data interchange on the Web. A collection of RDF triples constitute a RDF graph, each triple contains a subject and an object connected by a relation. We design an ontology knowledge representation of refrigerator, classify the refrigerator, the attribute of refrigerator as property, value of the attribute as object, thus with a triple format express refrigerator information.

Part of the ontology data as follows:



graph1 part of the ontology data

“BCD-206LST panel color silver”, “panel color” as the properties of refrigerator and at the same time as triples predicate connect a resource and a literal.

B. The construction of the triple ontology query

Out Of Vocabulary (OOV) means words that were not included in the tokenizer word but as a word in domain knowledge. In our paper, we mainly handle product model such as “BCD-206LST”, “BCD-196TMPI”. To identify the OOV, we add the OOV into segmentation tools AnsjSeg user dictionary. For user input errors may exist, we take the corrective action of OOV. We take the Levenshtein[10] algorithm match OOV with the user questions, to judge whether a user question containing the OOV by set a given threshold value. “What’s the color of BCD-196TMP?” would be corrected as “What’s the color of BCD-196TMPI?”

At the start of our approach we add the class, instance and property names in the ontology to segmentation tools AnsjSeg user dictionary, Among them, the property include object property and data property (an object property connect two individuals, a data property connect an individual and a literal), respectively, set its nature as “oc”, “ind”, “dp” and “op”, providing support for the analysis of the user’s question.

Table 1 part of word nature

word nature	meaning
n	noun
nr	person name
ns	place name
v	verb
oc	ontology class
dp	data property
op	object property
ind	individual

After word segmentation, the user's natural language questions translate into Qnl = {<word, nature>| word ∈ userDic or word ∈ systemDic, nature ∈ Nature }. The word segmentation result of “what is the panel color of BCD-196TMPI?” can correctly identify “BCD-196TMPI” as an individual, “panel color” as a data property in our ontology knowledge base. According to the word nature determine the subject-property-predicate of sentence construction. Match words which nature as verb to ontology property, match words which nature as noun to ontology instance. According to the above analysis, build query which “BCD - 196 TMPI” as subjects, “panel color” as a predicate.

Using the Stanford Dependency analyze dependencies between the keywords. Dependency analysis result is a set of triples of words, get set Raa={<reln,arg1,arg2>|reln∈Relation, arg1,arg2 ∈ Qnl.word}, Relation are defined in Stanford Dependency.

Table 2 part of dependency relation

relation	meaning
root	root
top	topic
assmod	associative modifier
assm	associative marker
attr	attributive
appos	appositional modifier
dep	dependent
dobj	direct object

Each triple like < arg1,reln,arg2> is the seed of a SPARQL query, in the question ”What is BCD-206LST’s maximum volume?”, Stanford Dependency analyze a triple < assmod, BCD-206LST, maximum volume >, thus build query which “BCD-206LST” as subjects, “maximum volume” as a predicate.

C. Entity disambiguation and predicate disambiguation

Actually user’s questions usually don't use the same words in our knowledge base, How to match a given a word to a class, individual, property or literal value in knowledge base, essentially is a disambiguation process.

In our method, we rely on the ontology and segmentation recognizing a word if a class, individual or property, as for literal value, we build a table to identify a word if a color or a place, with a complete and detailed ontology could do better in entity disambiguation and predicate disambiguation.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The study of the Chinese question answering system is not yet mature, without a recognized international Chinese question-answering system test set and evaluation method. For this reason, we select 18 typical commodities consulting question from the Haier's official website of as the test set of this system, artificial judgment system return answer is correct. 5 questions are due to the incompleteness of the ontology knowledge base, 3 questions are caused by predicate disambiguation, 1 question is caused by inaccurate object recognition, 9 questions get the exact answer. Recall Ratio is 84.6%， Precision Ratio is 85.15%. Compared to Yang[8] method, this method combined dependence analysis with word’ s nature build ontology triples query, makes the recall ratio and precision improved by 4.96% and 3.81% respectively; Compared to Xu Kun and Feng Yansong[9] method, precision improved by 40.15%, this is mainly due to their method study for open field, the predicate disambiguation is difficult, can't eliminate ambiguity targeted to domain ontology.

V. SUMMARY

This paper presents a design and implementation framework of question answering system based on natural language understanding, constructs the SPARQL structured query. The experimental results show that the proposed framework can effectively convert Chinese natural language questions to structured query, providing a solution for the intelligent question answering system.

Through error analysis, we found that about 75% of the error caused by predicate disambiguation, so in the future work the predicate disambiguation still be the key research, and building a complete and detailed ontology to support natural language understanding.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Special for Scientific and Technological development of the Construction Cause of Qingdao (No. JK2014-34).

REFERENCES

[1] Li Rong, Yang Dong, Liu Lei,. Based on the concept of ontology similarity calculation method research [J]. Journal of computer research and development,2011,48(z2):690-695.

[2] Cimiano, P., Haase, P., Heizmann, J., Mantel, M., Studer, R.: Towards portable natural language interfaces to knowledge bases – the case of the ORAKEL system. Data & Knowledge Engineering 65(2), 2008:325–354

[3] Unger, C., Cimiano, P.: Pythia: Compositional meaning construction for ontologybased question answering on the semantic web. In:Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 153–160. Springer, Heidelberg

[4] Lopez V, Fern, Ndez M, Motta E, Stieler [1] N. PowerAqua: Supporting users in querying and exploring the Semantic Web. Semantic Web. 2011;3(3):249-65.

[5] Damljanovic D, Agatonovic M, Cunningham H. FREyA: an interactive way of querying linked data using natural language // QALD-1, ESWC. Crete, Greece,2011:115–120

[6] Kaufmann, E., Bernstein, A., Fischer, L.: NLP-Reduce: A Naive but Domainindependent Natural Language Interface for Querying Ontologies. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 1–2. Springer, Heidelberg

[7] Yahya M, Berberich K, Elbassuoni S, et al. Natural language questions for the web of data // Natural Language Processing and Natural Language Learning. Sydney, Australia,2012:379–390

[8] Zhang Zongren, Yang Tianqi. SPARQL ontology query based on natural language understanding. Computer applications. 2010;30(12):3397-400.

[9] Xu Kun, Feng Yansong, Zhao Dongyan, Chen Liwei, Zou Lei. Chinese natural language semantic understanding of question based on knowledge base. Journal of Beijing university (natural science edition). 2014;50(1):85-92.

[10] KOS Victor . The Levenshtein Distance. Debreceni Műszaki Közlemények. 2012, 2011(3)