Proceedings of the 2007 IEEE Symposium on
Computational Intelligence and Data Mining (CIDM 2007)

A Constraint-based Method for Semantic Mapping from Natural Language Questions to OWL

Mingxia Gao*, Jiming Liu*†, Ning Zhong*‡, Furong Chen§
*The International WIC Institute, Beijing University of Technology, 100022, Beijing
Email: gaomx@emails.bjut.edu.cn
†School of Computer Science, University of Windsor, Windsor, Ontario, Canada
Email: Jiming@uwindsor.ca
‡Maebashi Institute of Technology, Maebashi, Japan
Email: zhong@maebashi-it.ac.jp
§R&D Center TravelSky, Technology Limited
Email: bjcfr@163.com

Abstract— **The goal of an on-line ontology-based question-answering system is to automatically derive answers from ontology knowledge bases without demanding additional information or intervention from users. This paper focuses on the problem of automatically mapping the tokens of a question into OWL elements, as an important step towards the further construction of answers. This problem can be essentially viewed as that of question understanding. The basic ideas underlying our method can be stated as follows: first we translate the tokens of a question as well as their syntactical and semantic relations (as in NLP) into constrained question variables and functions, and thereafter, we utilize an optimization-based assigning mechanism to substitute the question variables with the corresponding constructs in OWL knowledge bases. In the paper, we will discuss our preliminary studies using the questions collected from, and the knowledge base built at, the International WIC Institute (WICI).**

I. Introduction

In developing a question-answering system, one needs to address the issue of how to derive appropriate answers without requiring additional inputs or intervention from users. Ontology-based question-answering makes use of some pre-constructed, potentially useful ontology knowledge sources to construct answers. One of key problems in building an ontology-based question-answering is to understand a question at hand through some candidate knowledge bases, that is, to map a natural language question into logical queries for inferring answers from the candidate knowledge bases given certain semantic constraints.

In our work, since we are interested in on-line applications, the candidate knowledge bases that we deal with will be represented in OWL [1], [2], a standard Web Ontology Language from W3C. Besides the candidate knowledge bases, the process of mapping a question can also be affected by many syntactical, semantic, and contextual factors, such as results of in-depth linguistic analysis, profiles of users who ask in the question, and the underlying common sense that is involved. The paper will consider these factors and in particular focus on how to formulate and translate the tokens of a question into the OWL constructs subject to the above-mentioned determining factors. On the one hand, different tokens in a question are formulated into a set of variables, which are constrained by factors originating from associated syntactical, semantic, and contextual knowledge. On the other hand, each OWL knowledge base is indexed like a ‘dictionary’ that is useful for constructing the domains of question variables. Finally, our method composes an optimization-based objective function for matching question variables with sound OWL elements, e.g., for composing RDF triples as queries.

In our work, we have performed some preliminary evaluations using the questions and knowledge base available from the International WIC Institute (WICI), in order to validate whether or not the proposed method given in the paper can perform better than the keyword-based matching method with lemmatization and normalization, a non-trivial benchmark as used in many information retrieval studies.

The paper is organized as follows. Section 2 surveys related work. Section 3 presents the constraint-based mapping method in detail. We provide our experimental results in Section 4. Section 5 concludes the paper and presents future work.

mds
November 18, 2002

A. Related Work

As reported in the literature, different QA systems have used different methods of question understanding. Generally speaking, the goal of question analysis in database-based QA is to map natural language questions into well-formed SQL queries. Decomposing questions for Web-based QA focuses on rewriting natural language questions into search engine queries. The goal of question understanding in ontology-based QA is to interpret natural language questions with an available ontology query language.

Much of the early work has been centered around the methods of natural-language question mapping based on databases [3], [4], [5], [6], [7], [8]. Early methods [3], [4] have used predicate logic as the representation language to manually construct a concept map that captures the concepts and roles involved in a question. PRECISE NLI [5], [6] parsed questions to the corresponding SQL queries using a statistical parser as

1-4244-0705-2/07/$20.00 ©2007 IEEE

a "plug in", lexicon, and a maxflow algorithm. Others [7], [9] have explored a learning-based approach that combines different learning methods in inductive logic programming (ILP) to allow learners to produce more expressive hypotheses than that of an individual learner and to build a predicate lexicon with different learning methods.

Most of the recent work is concerned with Web-based QA, where the issues of question analysis have been treated as those of question classification [16], [17], [18], [19], [20] and question rewriting [9], [11]. Studies on question classification classify questions according to different methods and criteria used, whereas question rewriting focuses on matching questions to query phrases based on either some simple, manually constructed rules [9] or automated learning [11].

The previous studies that are related to our work include: MOSES [14], Aqualog [15], as well as others [13]. However, MOSES can deal only with questions in Denish and Italian. The process of parsing questions involved natural language processing and domain ontology modeling. AquaLog dealt specifically with English questions by using customized triples as the intermediate representation language. It required users to manually solve the ambiguity problem in semantic understanding. Another difference between this work and ours is that it involved most of the "who" and "what" questions, but not the "when" and "where". Strictly speaking, the work presented in [13] was not a real ontology-based QA. Solvable questions of the system were a subset of natural English (controlled English). In this system, each query was translated into a discourse representation structure by a parser. To summarize, our work differs from the above-mentioned work in the following ways:

1) We attempt to incorporate associated knowledge that is learned or collected from natural language processing, machine learning, and behavioral learning.
2) Our work formalizes the associated knowledge into quantitative functions in order to constrain the mapping of question variables.
3) Our work applies different levels of natural language processing in decomposing questions.

B. The Constraint-based Mapping Method

1) A General Description: Understanding questions using candidate OWL knowledge sources is in essence to understand the meanings of different components in a question based on the elements or assertions given in OWL under some conditions. In order to acquire a sound interpretation, two subproblems must be solved. One is concerned with the unit of interpretation; another is related to the factors that contribute to the understanding. With respect to the OWL knowledge, elements such as Class or Property are the basic units. From the point of view of natural-language QA, a word or a phrase is an indivisible unit. Thus, in our work, the preferred granules in decomposing a question will be the tokens such as words or phrases, which can in turn match the elements of OWL.

In our work, the contributing factors include the domains of questions, user profiles, and common sense, etc.. Different types of associated knowledge may be derived using different learning methods, and can have different roles in the question mapping. In order to utilize them in a systematic way, we will in this adopt a constraint-satisfaction-problem (CSP) formulation.

The CSP is a fundamental problem in Artificial Intelligence, which has been extensively studied by researchers with many interesting real-world applications, such as knowledge representation, scheduling, and resource allocation. Generally speaking, the constraints in a CSP can take different forms, such as logical, polynomial, fuzzy set based constraints. Addressing the complexity of decomposing a question and the diversity of constraint factors, in our work we will provide a CSP-based mapping method for interpreting a question in terms of the elements in OWL.

The basic ideas of our method can be summarized as follows. First, it builds a knowledge dictionary consisting of the elements of OWL by using existing OWL parsers, such as jena or OWL API. This dictionary is used for composing the domains of question variables. Second, it decomposes a question into a set of variables through natural language processing. Finally, it incorporates and represents the associated knowledge of the question into constraints, and thereafter represents them as an optimization problem.

Details on constructing a dictionary will not be discussed in this paper. This section focuses primarily on decomposing a question and formalizing related constraints.

2) Definitions: In this subsection, we will provide some formal definitions that will be useful for our further descriptions.

Definition 1: Let $OntoElement :=< Type, Name, Relation >$ denote the elements in OWL, where $Type$ is the element type, including: $class$, $individual$, $DatatypeProperty$, $ObjectProperty$, and $value$; $Name := \{Token_i\}_{i=1}^{r}$ corresponds to the names of the elements in OWL, where $Token$ composed of the elements in OWL are words with lemmatization and normalization. Furthermore, $Relation := \{\{< property, subject, object >_j\}_{j=1}^{s} | Name \subseteq (property \cup subject \cup object)\}$.

Definition 2: Let $QuesBase :=< QV, QC >$ be a formal representation of a question in a given context, where $QV := \{qv_i\}_{i=1}^{n}$ denotes a set of question variables, and each question variable is written as $qv_i :=< ID, Term, Attribute >$, where ID is an identifier, $Term := \{Token_j\}_{j=1}^{r}$ denotes question elements composed of words with lemmatization and normalization in a question, and Attribute corresponds to the properties of a variable. Examples of attributes include: $\{LEMMA$, SYN, POS, $PHTYPE$, $PHHEAD$, and $NETYPE\}$. $QC := \{qc_k\}_{k=1}^{m}$ is a set of constraints related to the question, and its element is a question constraint $qc_k :=< S, R >$, where $S \subseteq QV$ is a set of variables, called the constraint scope, and $R : S \longrightarrow D$ is a function from S to D, called the constraint relation.

Definition 3: Let $D := \{\{D_i\}_{i=1}^{n} | D_i = \{OntoElement\}_{j=1}^{s}\}$ denote a set of candidate OWL elements, called the domain of a question variable.

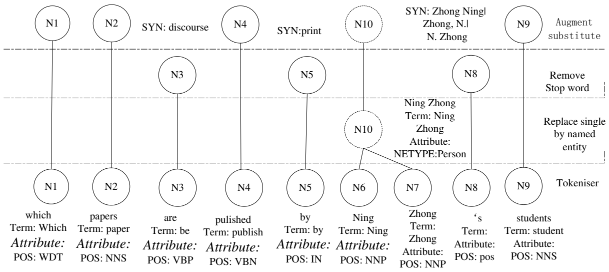
N1
N2
SYN: discourse
N4
SYN:print
N10
SYN: Zhong Ning\| Zhong, N.\| N. Zhong
N9
Augment substitute
N3
N5
N8
Remove Stop word
Ning Zhong Term: Ning Zhong Attribute: NETYPE:Person
N10
Replace single by named entity
N1
N2
N3
N4
N5
N6
N7
N8
N9
Tokeniser
which Term: Which Attribute: POS: WDT
papers Term: paper Attribute: POS: NNS
are Term: be Attribute: POS: VBP
pulished Term: publish Attribute: POS: VBN
by Term: by Attribute: POS: IN
Ning Term: Ning Attribute: POS: NNP
Zhong Term: Zhong Attribute: POS: NNP
's Term: Attribute: POS: pos
students Term: student Attribute: POS: NNS

Fig. 1. A schematic illustration of question decomposition.

3) Question Decomposition: The ultimate goal of question decomposition is to identify the tokens in a question (called question variables) that can be interpreted as the elements of OWL. In doing so, various basic techniques of natural language processing will be involved, including tokenization, identification of named entities (NE) and synonymies. Furthermore, the identified tokens or question variables must be in agreement with syntactical, semantic, and contextual constraints.

Figure 1 provides a schematic illustration of question decomposition, along with a sketch of question variables. As shown in the figure, an original question is first decomposed through tokenization into candidate question variables containing attributes and terms. Next, some parts of the variables are replaced with named entities, as can be learned from related texts, and at the same time, stop words are removed. Finally, we supplement attributes of the variables by synonyms or abbreviation.

4) Formalizing Basic Constraints: Generally speaking, there are two important aspects to be taken into account in understanding a question: question-based basic syntactical and semantic constraints, and user-based external constraints. The basic constraints are concerned with the degrees of similarity between the texts corresponding to question variables, and the associated relations as identified from syntactical and semantic analysis. The external constraints include the common sense knowledge related to the question domain, and in addition, the background or profile of users. Both constraints, with different representations, can be acquired by applying various techniques, such as statistical analysis, machine learning, and behavioral modeling. They can be formally represented using quantitative functions that can, in turn, be utilized to optimize question variables in OWL. At the present stage of our work, we consider only the basic constraints.

Here, question variables refer to the words or phrases in a question. Their corresponding candidate values are the elements in OWL as acquired through surface text mapping. The task of surface text mapping is to determine whether or not the tokens of variables are the same as the OWL elements, and if not, how similar the variables are to candidate elements. Since each candidate element may be composed of a certain number of tokens, a variable has different similarity comparing with different candidate value. We use Eq. 1 to calculate the degree of similarity, where $a \in [0,1]$ is a decreasing coefficient corresponding to the different expressions of the same meaning, e.g., using synonyms or abbreviations. $\mu = \{0,1\}$ is used to activate the decreasing coefficient.

TABLE I

THE RELATIONS BETWEEN POS OF A QUESTION VARIABLE AND THE ELEMENT TYPE IN OWL

POS	Element type in OWL				
	Class	Individual	OP	DP	rang
NNP?	0	1	0	0	1
VB?	0	0	1	1	0
CD	0	0	0	0	1
FW	0	1	0	0	1

$$f_{sim}(x|_{x \in D_i}) = a^{\mu} \frac{|x.Name \cap qv_i.Term|}{x.Name}. \qquad (1)$$

Although natural language expressions and OWL are two types of knowledge representations, there are some corresponding relations between them. In particular, an element type in OWL implies that an element with given a type is more appropriate to be associated with a question variable with a certain attribute (e.g., POS). For example: a proper noun is a noun that names a specific person, place, or thing, whereas a specific person, place or thing usually is asserted by individual or $DatatypeProperty : rang$ in OWL. Thus, a variable with a proper noun in a question can be generally interpreted as an element with an individual in OWL. In order to formalize such qualitative corresponding relations, we have defined a quantitative function, Eq. 2, which explains the possibility of a question variable with a given attribute POS being interpreted as an element type in OWL:

$$f_{POS} : \prod_{x \in S_i} D_i \longrightarrow [0,1] \qquad (2)$$

where 1 and 0 correspond to the full possibility and impossibility, respectively. For other non-measurable methods, we use 0.5 as an output. Table I presents the relations between POS (POS tag is from Penn Treebank II Tags) and the element types in OWL.

Note that members within the same phrases, modifiers, and predicates constitute the dependent relations of question variables, as identified from syntactical and semantic analysis. The variables with the dependent relations are easy to be mapped into OWL, as exhibited by RDF triples. In our present work, we introduce a binary function, Eq. 3, to indicate whether or not the existing OWL knowledge supports these dependent relations:

$$f_{(phr,mod,pre)}(x) = \{ \begin{array}{l} 1, x \in D_i \text{and} |\exists x.Relation \\ \quad \cap \{R_{phr}, R_{mod}, R_{pre}\}| \geq 2 \\ 0, other \end{array} \qquad (3)$$

where R_{phr}, R_{mod}, and R_{pre} denote a phase relation, a modifier relation, and a predicate relation, respectively.

Once the existing qualitative constraints are formalized into quantitative functions, finding a sound assignment to the

TABLE II

THE ELEMENTS IN INSTITUTION.OWL

Element type	Class	Individual	OP	DP
Number	83	90	37	20

question variables can be treated as solving an optimization problem. The specific objective function are given in Eq. 4, where f_k refers to different constraint functions, f_{sim}, f_{POS}, and f_{phr}. w_k is the weight of a corresponding constraint, which is used to express the importance of each constraint.

$$f(q) = \sum_{qv_i \in QV} \max_{x \in D_i} \{ \sum_{k=1}^{m} w_k \times f_k(x) \}. \qquad (4)$$

C. Experiments

In our preliminary experiments, we set the same weights for all constraints. We compare the performance of the constraint-based method (called CBM) with that of a lexical-level, keyword-based matching method with lemmatization and normalization (called match). Note that the match method we will compare with is a non-trivial method which has been previously used in other information retrieval tasks.

1) Experimental Data: Although there exist various OWL knowledge bases on the Internet, they, generally speaking, only define some taxonomies in some domains but lack real-world instance knowledge. In our present work, we are more interested in an OWL knowledge base pertaining to specific individuals and their properties, as related to the instances of a university and the International WIC Institute (WICI), the size of which is summarized in Table II.

We have used two types of natural language questions in our experiments: One type of questions is related to the WIC Portal [22], and another is from simulated questions regarding the instances of the International WIC Institute (WICI) based on the question set from Webclopedia [21]. The questions with different types of answers and different degrees of complexity are classified according to the question markers, including who, which, and what etc., as shown in Table III. In the questions, subjective questions, such as why and how, are not included; an example of "other" questions is "Name the person that has same advisor as Su Yila."

2) Pre-processing: In the pre-processing stage, the questions are decomposed based on semantic entailment [23], [24]. In order to eliminate the errors as introduced from the question pre-processing, we need to manually check the question set, and replace those that are falsely decomposed or falsely labeled with some equivalence expressions through manual or automatic semantic entailment.

The above-mentioned pre-processing based on semantic entailment can also reduce the question variables that do not contains candidate values.

3) Experimental Results: Tables III and IV show our experimental results. From the results, we can note that the constraint-based method performs reasonably well some of the cases. On average, this method presents a significant improvement over the match method; the precisions across all classes of questions have been improved by over 15%.

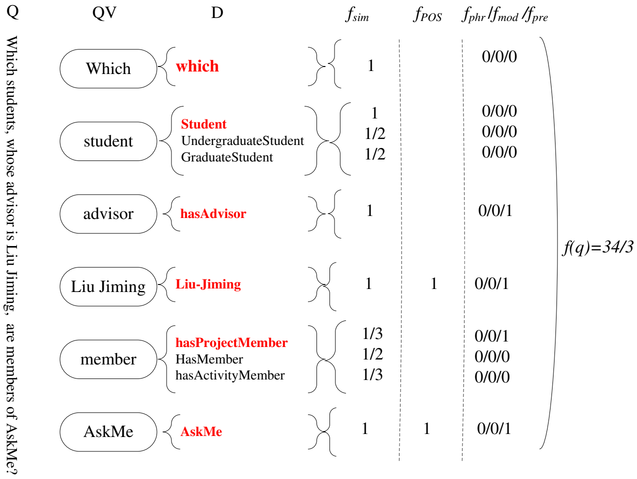

Fig. 2. A walk-through example of the constraint-based method

A walk-through example, as given in Figure 2, shows a mapping process between the question variables and the elements in OWL.

II. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed the basic ideas and formulation of a constraint-based method for semantic mapping from a natural language question to the elements in OWL. In this method, we first decompose questions into a set of variables by means of syntactical and semantic analysis, and then formulate their underlying constraints, e.g., associated knowledge, into different quantitative functions. Thereafter, we can make use of an optimization-based objective function to find sound substitutes in the OWL knowledge representation for the question variables. Our preliminary experiments using the WICI knowledge base and question sets have indicated that the proposed method is promising for further development.

In our future work, we will investigate in-depth how to represent associated knowledge and how to systematically derive and formulate the corresponding constraints for the purposes of question understanding and semantic mapping.

ACKNOWLEDGMENT

The work is partially supported by the NSFC major research program: "Basic Theory and Core Techniques of Non-Canonical Knowledge" (60496322) and Open Foundation of Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology.

REFERENCES

[1] http://www.w3.org/TR/owl-features/.
[2] http://www.w3.org/TR/owl-guide/.
[3] I. Androutsopoulos, G. Ritchie, P. Thanisch: Natural Language Interfaces to Databases—An Introduction, Journal of Natural Language Engineering 1(1), 29-81, 1995.

TABLE III
THE QUESTION SET AS USED

	Who	Which	What	How mang	When	Where	Yse/no	Other	Overall
Simulative questions	4	5	9		3	4	7	3	35
Real questions	7	7	9	3	2	3	5	4	40

TABLE IV
THE PRECISION COMPARISONS BETWEEN THE MATCH AND CBM METHODS IN ANALYZING REAL QUESTIONS

	Who(%)	Which(%)	What(%)	How mang(%)	When(%)	Where(%)	Yse/no(%)	Other(%)	Overall(%)
Match	42.9	14.3	44.4	66.7	50.0	33.3	20.0	25.0	35.0
CBM	42.9	42.9	66.7	66.7	100.0	66.7	60.0	50.0	57.5

TABLE V
THE PRECISION COMPARISONS BETWEEN THE MATCH AND CBM METHODS IN ANALYZING SIMULATED QUESTIONS

	Who(%)	Which(%)	What(%)	When(%)	Where(%)	Yse/no(%)	Other(%)	Overall(%)
Match	75.0	100.0	22.2	33.3	75.0	14.3	0	42.9
CBM	75.0	100.0	44.4	33.3	75.0	57.1	0	57.1

[4] I. Androutsopoulos, G. Ritchie, P. Thanisch: MASQUE/SQL-An Efficient and Portable Natural Language Query Interface for Relational Database, Proceedings of the 6th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert System, Edinburgh, 1993.

[5] A. Popescu, O. Etzioni, H. Kautz: Towards A Theory of Natural Language Interfaces to Databases, Proceedings of the 8th international conference on Intelligent user interfaces, Miami, Florida, USA, 2003.

[6] A. Popescu, A. Armanasu, O. Etzioni, et al: Modern Natural Language Interfaces to Databases: Composing Statistical Parsing with Semantic Tractability, Proceedings of the 20th International Conference on Computational Linguistics, Geneva, August 2004.

[7] C. Thompson, R. Mooney: Automatic Construction of Semantic Lexicons for Learning Natural Language Interfaces, Proceedings of the 16th National Conference on Artificial Intelligence, 487-493, Orlando, FL, July 1999.

[8] L. Tang, R. Mooney: Using Multiple Clause Constructors in Inductive Logic Programming for Semantic Parsing, Proceedings of the 12th European Conference on Machine Learning, 466-477, 2001.

[9] E. Brill, S. Dumais, M. Banko: An Analysis of the AskMSR Question-Answering System, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 2002.

[10] E. Agichtein, S. Lawrence, L. Gravano: Learning Search Engine Specific Query Transformations for Question-Answering, Proceedings of the 10th World Wide Web Conference, 169-178, Hong Kong, 2001.

[11] D. Radev, H. Qi, Z. Zheng, et al: Mining the Web for Answers to Natural Language Questions, Proceedings of ACM CJKM 2001, Atlanta, Georgia, 2001.

[12] D. Azari, E. Horvitz, S. Dumais, et al: Action, Answers, and Uncertainty: A Decision-Making Perspective on Web-based Question-Answering, Information Processing and Management 40(2004), 849-868, 2004.

[13] A. Bernstein, E. Kaufmann, N. Fuchs, et al: Talking to the Semantic Web—a Controlled English Query Interface for Ontologies, AIS SIGSEMIS Bulletin 2(1), 42-47, 2005.

[14] P. Paggio, D. Hansen: Ontology-based Question Analysis in A Multilingual Environment: The MOSES Case Study, Proceedings of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, 1-8, May 2004.

[15] V. Lopez, M. Pasin, E. Motta: AquaLog: An Ontology-Portable Question-Answering System for the Semantic Web, Proceedings of the 2nd Annual European Semantic Web Conference, LNCS 3532, 546-562, 2005.

[16] X. Li, D. Roth: Learning Question Classifoers, Proceedings of the 19th International Conference on Computational Linguistics, 556-562, 2002.

[17] D. Metzler, W. Croft: Analysis of Statistical Question Classification for Fact-based Questions, Information Retrieval 8, 481-504, 2005.

[18] D. Zhang, W. Lee: Question Classification Using Support Vector Machines, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and. Development in Information Retrieval, 26-32, 2003.

[19] J. Pomerantz: A Linguistic Analysis of Question Taxonomies, Journal of the American Society for Information Science and Technology 56(7), 715-728, 2005.

[20] Z. Cheung, K. Phan, A. Mahidadia, et al: Feature Extraction for Learning to Classify Questions, AI 2004: Advances in Artificial Intelligence 3339(2004), 1069-1075, 2004.

[21] http://www.isi.edu/natural-language/projects/webclopedia/.

[22] http://www.iwici.org/.

[23] R. S.Braz, R. Girju, V. Punyakanok, et al: An Inference Model for Semantic Entailment in Natural Language, Proceedings of the 20th National Conference on Artificial Intelligence, 1043-1049, 2005.

[24] R. Braz, R. Girju, V. Punyakanok, et al: Knowledge Representation for Semantic Entailment and Question-Answering, Proceedings of IJCAI'05 Workshop on Knowledge and Reasoning for Answering Questions, 2005.