Semantic Representation and Ontology Construction in the Question Answering System

Ying-Hong Wang, Wen-Nan Wang, Chu-Chi Huang, Ting-Wei Chang and Yi-Hsiang Yen
Department of Computer Science and Information Engineering, Tamkang University, Taiwan
{inhon,123670}@mail.tku.edu.tw, {895410131, 695420306}@s95.tku.edu.tw,
692192239@s92.tku.edu.tw

Abstract

To build an automatic assistance for learning and provide a self-paced learning mechanism are the objectives in the e-learning environment. The paper improves the previous proposed Semantic Question Answering System which applies Link Grammar and WordNet to form a Semantic Tree to represent the meaning of question and further find the relevant answer based on the selected Data Structure Ontology. The paper addresses the improved functions including the flexible answering method which can refine the Semantic Tree to find more relevant answers, the Ontology Extension Module is designed to acquire data from Internet to raise the ontology and the Feedback for System module is designed for the instructors to provide more sufficient knowledge.

1. Introduction

Recent years, the e-learning research issues focus on how to provide more interactive activities around learners and instructors. If system provides only static learning contents to learners, this will hardly attract learners to study in this e-learning environment. How to provide the automatic assistant-learning and self-paced learning mechanisms is the primary objective in e-learning environment. In this paper, we firstly introduce the motivation that we construct the Semantic Question Answering System in previous research. [1]

Some traditional question answering systems and famous search engines such as Google and Yahoo use keyword search, statistics and ranking mechanisms to find the possible answers from question without discussing with the syntax of the question sentence. The research results may return redundant search results since these systems do not concern about the meaning of user's questions.

For example, the synonymy means those different words may present same meaning. In QA systems, learner may use different words with same meaning to make questions. While the keywords not storing in the knowledge storage, the question answering system or search engine will never find the correct answer return to learners. Since these systems will never have able to make sense to know what the questioner asked.

Take the question as example, "How to implement a first in first out queue?" The system can only recognize the keywords set of the question sentence which contains "implement" or "queue". If learner uses the word "create" or "construct", these systems will never return the appropriate answers since they do not know those words present the same meaning. If the system data storage does not contains the information about the above keywords, the answer will never be able to be made. Besides, the question answering system will not know the question tag "how" have the meaning of finding the "method about something". For the above problems, this research applies the semantic analysis process to understand the meaning of the input question sentences and return more relevant answer back to learners.

In commonly, we know different words have different means. Before applying the semantic analysis, the syntactical information of the question sentence is firstly needed to analyze. In this section, we include the Link Grammar parser to analyze the syntactical information (the Syntactical Tree). This information will then help to construct the meaning representation of the question sentence.

In different situation, the meaning of a sentence might be different. So we know the semantic analysis should be concerned to the specific domain. In this paper, the Data Structure course is applied, the terms in the data structure are limited and we can pre-define them in the system ontology to support the functions of syntax and semantic analysis. The other kinds of course structure would be studied in one's term.

0-7695-2983-6/07 $25.00 © 2007 IEEE
DOI 10.1109/CIT.2007.199

With the introduction of the previous research, we have proposed the Semantic QA system architecture. The answering method is to build a Semantic Tree to match the Data Structure Ontology and find the relevant answer. But the main restriction is that the answering method can only be made while the Data Structure Ontology has the same architecture of Semantic Tree. [1] In this paper, we are more concerning with how to apply the flexibility.

Since learners may make some questions with the correct syntax but wrong meanings. So we can not find the same Semantic Tree in Data Structure Ontology. In this situation, the improved mapping method will remove some concept or some instance words from the node of Semantic Tree. Thus we can find related contents from the Data Structure Ontology. This can help learner to know the correctly meaning related their question especially when they are not familiar with the courses. Besides, the Ontology Extension Module is also proposed to raise the system knowledge and the feedback for system can help instructor to provide sufficient system knowledge.

The paper is organized as follows. Section 2 describes the related works used in this paper such as Link Grammar [2], WordNet[3], Ontology[4]. Section 3 introduces the improved Semantic QA system architecture and system function details. The last section is the conclusion and future works

2. Related Works

2.1. Link Grammar

Link grammar which developed by Carnegie Mellon University is a graphical grammar analyzing tool. Link grammar is a context-free formula to describe natural language. [5] This system can produce all grammar linkage from English sentence which users input and determine the sentence correctness thought the linking result. Figure 1 will show what Link Grammar is. First, input an English sentence into this system. Each word has some curves and each curve has one label on it. The curve and label is a Link which expresses a kind of linkage. After analyze and parse the sentence though Link Grammar, we may get a lot of Linkages. This information can help to realize the syntax structure of the question

2.2 WordNet

WordNet originated from Cognitive Science Laboratory in Princeton University. It is a vocabulary reference system designed by researchers who inspired by psychology theory. WordNet processed the first level classification according to part of speech (POS) tag. Driven by different word meanings and expressions, it forms several Synset. Each Synset symbolizes one vocabulary and takes down other words and expression with the same meaning.

2.3 Ontology

People already explore some ways to express the meaning of data thought the data processing in information technology field. Although human explore the meaning of knowledge from Hellenistic Age to the present age. Ontology is a data model that represents a domain and is used to reason about the objects in that domain and the relations between them. Ontology is used in artificial intelligence, the semantic web, software engineering and information architecture as a form of knowledge representation about the world or some part of it. Ontology generally describe: individuals, classes, attributes and relations. Ontology provides an explicit conceptualization that describes the semantics of the data. [6] There are many tools to support Ontology, such as DAML [7], KAON[8], etc.

3. Semantic QA System

In previous research, a Semantic Question Answering System Architecture is proposed for e-Learning Environment. [1] In this work, we improve the system with more functionality. Figure 1 shows the modified system architecture. The following section describes the improved function to achieve the system design. There are four functional blocks in the system architecture shown in figure 1. Each functional blocks and its objectiveness is designed to:

- Syntactic Analysis: analysis the syntactical information from learner made questions.
- Semantic Analysis: find the semantic meaning to represent the meaning of the question according to the syntactical information.
- Ontology Process: query the course ontology and return relevant answer to learners.
- Feedback for system: the instructors can make the feedback for system to overcome while the information is insufficient.

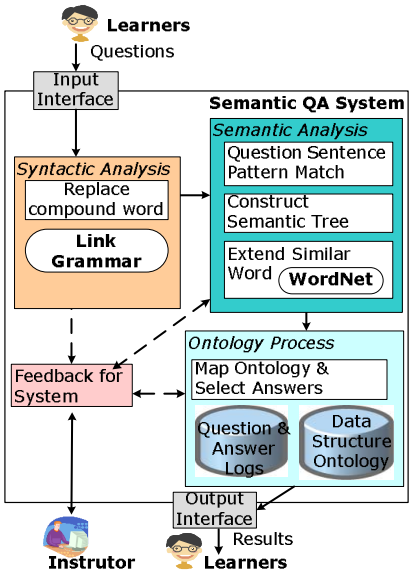

Figure 1. Semantic QA System Architecture

3.1. Syntactic Analysis

To reason a question meaning, it is important to know its syntax structure. But the learner may not familiar with the language. It will effect on the correctness of answer. With the correct syntax, the meaning of question can be analyzed with the following semantic analysis. But the limitation of syntax analysis will be restricted in the Link Grammar parser. The detail of syntactic analysis process has addressed in previous search. [1]

3.2. Semantic Analysis

This section addresses the improved semantic analysis.

3.2.1 Question Sentence Pattern Match

There are lots of question types in natural language; each of them may have different meanings. But there still have some different question types have the same meaning. This module provides possible question types and returns the corresponding Target Phrase which can represent the meaning of question made by learner. [1]

3.2.2 Semantic Tree Construction

In this process, the Semantic Tree is constructed from Target Phrase to represent the semantic meaning of question in order to represent the meaning of question sentences. The following shows the improved answering method.

A basic unit is defined as a Term, which can be a word or a compound word in the NP or VP form the Target Phrase. There will be at least one Term in a NP or VP. The following shows the construction steps:

Step1. The trace direction is bottom-up and left to right from the Target Phrase.

Step2. A node here is made up of each term in a NP or VP. In this step, a definite article will be discarded. The POS tag of each term is recorded as the node information. There are two categories of term in the node information: concept and instance. If the term appears in the concept layer of data structure ontology schema, it defined as the concept type of term. If not, the term belongs to the type of instance. Each node contains at least a term with concept or instance type. Figure 2 shows the node presentation.

$$\text{Node j}\begin{matrix}\text{Concept of TermB.POS}\\ \text{Instance of TermA.POS}\end{matrix}$$

*Note. j is the No. of Node

Figure 2. The Node presentation

Step 3. In the target phrase, the partial syntactic tree may contain some preposition phrases (PP). The preposition modifier can represent the relation between two words or phrases such as: in, on, of. In our research, we only concern about those preposition modifiers in the sentence that can help to identify which phrase is hypernym or hyponym. The syntax structure may be {VP PP NP} or {NP PP NP}. If there are no preposition modifier appeared we construct the link with no direction between two nodes. Besides, the hypernym phrase can represent the super class. Then the Semantic Tree is constructed.

3.2.3 Similar Word Extension

This design can help to raise the possibility to find the relevant answer, if the answer stored in course ontology is not the word learner asked. We can iterate the similar word list to find the alternate answers in the course ontology. For example, the question: "How can we perform a last in first out operation in Queue?". The Semantic Tree with the similar words information shows in figure 3. And the word operation is the concept of one Node. Here the similar words of operation have not need to be enhanced since it had matched the ontology schema.

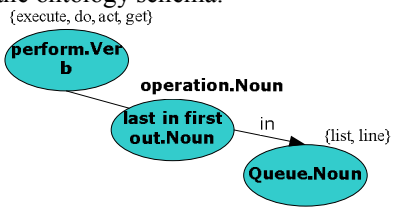

Figure 3. Add similar words to Semantic Tree

3.2.4 Mapping Ontology and Answer Selection

The constructed Semantic Tree in previous section contrasts with Data Structure Ontology. If we can find the same Semantic Tree structure from Ontology, the element data is the solution. Different to previous section, we will find any Semantic Tree corresponds to the question sentence Semantic Tree in Data Structure Ontology. The mapping sequence is top-down, the

starting node is the super class node of the Semantic Tree. The following shows the mapping steps:

Step1. Query the ontology to check if each word in node exists in the ontology:

Each node may contain the concept word or instance word. Use these two information to find if there exists an element appeared in ontology whose concept layer match the concept word of the node and instance layer contains the instance word of the node. The concept word of node should be checked firstly. If the word has been matched from ontology then return the URI.

Step2. Check the dependency relation to ontology:

In the previous section of Semantic Tree construction steps, the nodes dependency has been constructed according to the preposition modifier. In this step, the dependency relation between nodes is applied to find if there contain any node relation of Semantic Tree that can also be applied to the elements of the ontology.

But the Semantic Tree structure may not fulfill the layers of ontology schema. In a question sentence, there may have only two or more words can be used to represent the semantic meaning. While mapping the Semantic Tree to the ontology, there may be some dependency relation between two nodes which have to cross two or more nodes in the ontology structure. Since the ontology schema contains the hierarchy structure, the upper layer element is the supper class of the bottom layer element.

For the above problem, transitivity is applied to resolve this problem. With the following assumption, $Node_A$ is a subclass of $Node_B$, $Node_B$ is a subclass of $Node_C$. According to the property of transitivity, we can know $Node_A$ is also a subclass of $Node_C$. Then we can apply this theorem to the ontology mapping. From this step, we can find the relation between the elements of ontology with the corresponding dependency of nodes in the Semantic Tree.

Step3. Return Full Mapped Answer:

This process maps the ontology to check if it has the same hierarchy to Semantic Tree. If the mapping process succeeds, the bottom element data will be returned to be the answer. Different from the previous research, when the mapping process fails, this process will not stop at here. [1] From the Semantic Tree shown in figure 3, we can know in actually the Data Structure Ontology will never contain the same hierarchy of the Semantic Tree. The reason is that we can not perform the last in first out operation in Queue. But learners may make this problem due to they do not familiar with the Queue operation in Data Structure. And we should find some answers which can help the learner to know what are the correct answers or the relevant answers. If the mapping process is failed, then the next process goes to step 4.

Step4. Refine the Semantic Tree to Return Partial Mapped Answer:

From the above steps, there may have some nodes or relations mapped since the meaning of the question may not semantically correct. Besides, if the node contains no word matched to the concept and instance layer of ontology then the node should be eliminated from Semantic Tree. This can help to find partially mapped answer to learner. Take the Semantic Tree in figure 3 as the example.

In step 1, we can not find any element whose ancestor is "Queue" and contains the word "last in first out" in instance layer and word "operation" in concept layer. So the process should take off the instance "last in first out" from this node.

In step 2, the node "perform" and its similar words {execute, do, act, get} are not existed in the ontology concept or instance. Then it should be also eliminated from the Semantic Tree. The refined Semantic Tree structure is shown in following figure 4.

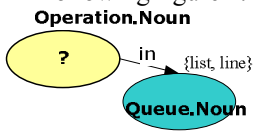

Figure 4. The Refined Semantic Tree

And figure 5 shows the mapped Data Structure Ontology has satisfied the refined Semantic Tree. So that the possible answers can be the instances belonging to "enqueue" or "dequeue".

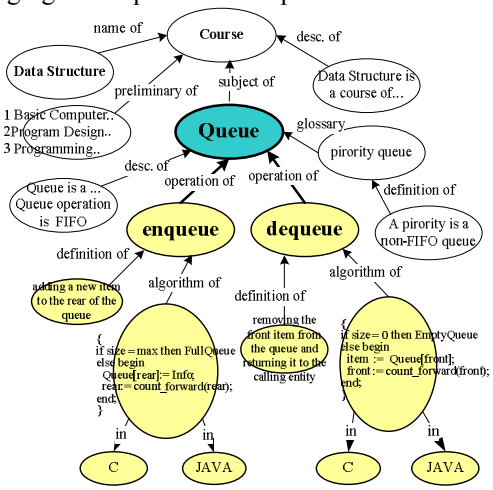

Figure 5. The Partially Mapped Answers

3.3. Instructor Feedback for System

If there are some answers cannot be answered from the system, the reason may come from the insufficient system knowledge. In this situation, instructors can

manually add information to the system through this interface. The information may contain: (1) Compound Word List: The proper nouns of selected data structure course are limited. But there may still have some proper nouns not listed. (2) Question Sentence Pattern: There are some simple question sentences analyzed in this table. There should be also providing an interface to extend other question sentence patterns to classify other question types. (3) Similar Word List: There are some used words and their synonyms are listed in the table. This list can help the instructor to query the WordNet to add more words. (4) Course Ontology Refinement: Each question and answer will be recorded in Question and Answer Logs. If there exist some question cannot be automatically answered, there may cause of the insufficient knowledge in the data structure Ontology. Instructors can check the logs and add the needed content to the Data Structure Ontology in this interface.

4. Semantic Ontology Construction

The previous research has proposed a chat room to grade the leaner the dialogue relevant to the course topic. The grade judgment of relevance is come from the keywords of learner dialog have been related to the designed XML form of Data Structure Ontology. [9] In this paper, we use RDF to improve the previous QA system. [1] The following sections show the Ontology construction steps.

4.1 Build Domain Terminology

The Ontology is designed to be versatile and semantically rich. At the preliminary work, we build terminology used by Ontology from the Wikipedia, which is an open-ended online encyclopedia. People are free to contribute and edit contents to Wikipedia. The quality of content could be unofficial, but it is a good place to gather technical terms there. Each page can be logically separate into three parts. First is the term to be defined; second is the description of that term; last is some See Also links. While we inspect the HTML source, we found that every link is titled to be the term it is going to define.

The solution is quite simple that we just follow links starting at Data Structure page and extract its title attribute to form a set of our domain terminology.

We further constructed a directed graph to capture the relationships of terms. The graph forms a knowledge hierarchy which implicitly embedded in the website. The hierarchy captured is quite interesting but not to be used at this time. By separating Ontology into Conceptual and Instance Layer, we encoded Whole-Part relations of concepts into RDF format and mapped technical terms into Instance Layer.

Conceptual Layer is actually acting as classes of terms. The Conceptual Layer we designed reflected the building blocks of Data Structure textbooks. One reason is that we are familiar with the structure and presentation of textbooks. Another is that since textbooks are written for teaching's purpose, our system is built for the same purpose.

4.2 Ontology Extension Module

Figure 7 is the overview of Ontology Extension Module. We put emphasis on the automated aspect of Ontology population i.e. contents for the WWW. The Scheduler manages URLs and spawns Page Savers to fetch the content into Document Pool.

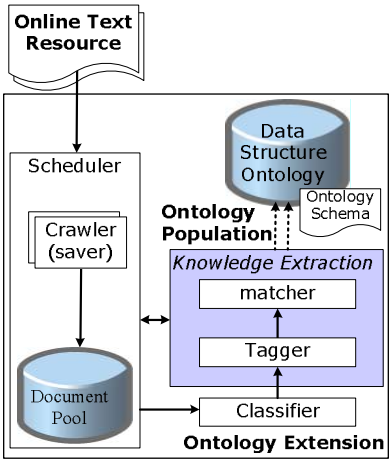

Figure 7. Ontology Extension Module

The Classifier analyzes texts enclosing in every paragraph and identifies phrases in domain terminology. The assumption here is that the phrases used in a page are proportional to the phrase linkages in the domain terminology. We use this assumption to simplify our system design and it's empirically correct for us to use this assumption.

From there on, we flatten web page and extract links for later use. The problem is that an answer can span multiple paragraphs. To solve this inter-paragraph problem, we use phrase collocation to detect whether the topic is still going on.

We select nouns in a paragraph and build a window which spans three paragraphs to represent our domain of discourse. In order to identify nouns in a paragraph, we pass each paragraph into POS Tagger[10] and extract adjacent noun class.

The heuristic is we usually explain things analogously. And we must relate things somewhere in the paragraph to perform this analogy i.e. there exists a sentence where two things collocate. Either we talk

about the subject or the things we related, they should still be in the same topic.

While two noun phrases appear in a sentence and one of them is in our domain of discourse, we invite another one into the current discourse. This process terminates when there's no noun phrases that contained in the following paragraph satisfies the current discourse. Every document contains at least a term in our domain terminology which itself in the Instance Layer. By analyzing paragraph by paragraph, we will eventually find this term. We call it Pivot Term of current scan. When we find this Pivot Term, we use 3-paragraph window to back trace previous two paragraphs in order to include these paragraphs into our discourse if needed.

Each scanning process concentrates on a single Pivot Term which has the highest frequency to be scan in the document. This information is provided by the Classifier which has previewed whole document in a bird's eye.

5. Conclusion and Future Works

This research proposed an improved Semantic Question Answering System to provide learning assistance with relevant contents in Data Structure Ontology. This previous research uses Link Grammar and WordNet to form a Semantic Tree to present the learner's question. But the answering mapping function must satisfy all the architecture of the Semantic Tree. When learners are not familiar with the courses, they may ask the question with correct syntax but error in meaning. The answer will not be returned. The improved answering mapping function can refine Semantic Tree to eliminate unrelated words and return relevant answer to learners. And the Ontology Extension Module is also proposed to raise the system knowledge and the feedback for system can help instructors to furnish sufficient system knowledge.

This research contains the following characteristics:

- Accuracy: use similar word list to extend the semantic information raise the possibility to find relevant answers.
- Flexibility: use pluggable module to accommodate multiple data sources, and unify heterogeneous data formats into uniform system Ontology scheme. The mapping function can provide flexible answering function.
- Usability: organize semantic relationship between words.
- Extension: We can apply this Ontology to support other courses. And the course material can also be extended through Ontology Extension Module.

Future research can plan and analyze the relation type between elements deeply. The semantic cognition method can deal with more question sentence type. Furthermore, the pattern matching table needs to be analyzed and expanded in the future to deal with more question sentence types. Finally, how to let Ontology learn automatically, combine and link the data correctly is worth for future research.

6. References

[1] Ying-Hong Wang, Wen-Nan Wang and Chu-Chi Huang, "Enhanced Semantic Question Answering System for e-Learning Environment", 21st International Conference on Advanced Information Networking and Applications (AINA 2007), Vol. 2, pp. 1023-1028, May 21-23, 2007

[2] Daniel Sleator, David Temperley, and John Lafferty, "Link Grammar," http://www.link.cs.cmu.edu/link/

[3] George Miller, "WordNet, a lexical database for the English language", http://www.cogsci.princeton.edu/~wn/

[4] Michael Denny, "Ontology Tools Survey, Revisited", http://www.xml.com/pub/a/2004/07/14/onto.html

[5] D. K. Sleator, Davy Temperley "Parsing English with a Link Grammar", October 1991, CMU-CS-91-196

[6] Deborah L. McGuinness and Frank van Harmelen, "OWL Web Ontology Language Overview", W3C Recommendation, 10 February 2004

[7] DARPA's Information Exploitation Office, The DARPA Agent Markup Language (DAML), http://www.daml.org/

[8] FZI and AIFB, "KAON is an open-source ontology management infrastructure targeted for business applications", http://kaon.semanticweb.org

[9] Ying-Hong Wang, Wen-Nan Wang, and Yi-Hsiang Yen, "An Intelligent Semantic Agent for e-Learning Message Communication," Proceedings of 2005 19th International Conference on Advanced Information Networking and Applications (AINA2005), held on March 28-30, 2005, Taipei, Taiwan, Vol. II, pp. 105-108

[10]Kristina Toutanova and Christopher D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger", Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hong Kong, pp.63-70

7. Acknowledgement

This research was funded by the National Science Council of the Republic of China under the Contracts No: NSC 94-2520-S-032-003 and NSC 95-2520-S-032-001.