Intelligent Information Extraction using Logical Approach: Skolemize Clauses Binding

Rabiah A.K.[a], T.M.T. Sembok[b], Halimah B.Z.[c]

[a]*Faculty of Computer Science and Information Technology*
Universiti Putra Malaysia, 43400 Serdang, Selangor
Tel : 03-89466537, Fax : 03-89466577
E-mail : rabiah@fsktm.upm.edu.my

[b]*Pejabat Timbalan Naib Canselor, Canseleri*
Universiti Pertahanan Nasional Malaysia, 57000 Kem Sungai Besi, Kuala Lumpur
Fax : 03-90565411
E-mail : tmts@gmail.com

[c]*Faculty of Information Science and Technology*
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor
Tel : 03-86216549
E-mail : hbz@ftsm.ukm.my

ABSTRACT

This paper is concerned with the problem of generating an automated information extraction in the context of logical knowledge representation, reasoning, and inferential processing. The research focused on Wh types of question to a restricted domain. Here, an existing resolution theorem prover with the modification of some components was adapted based on experiments carried out such as: knowledge representation, and automated information answer generation. Based on that, a Pragmatic Skolemize Clauses representation used to represent the semantic formalism for the computational linguistic was designed. The process of reasoning in generating automated information began with the execution of resolution theorem proving. Then, the answer extraction was proceed with skolemize clauses binding approach to continue tracking the relevant semantic relation rules in knowledge base, which contained the answer key in skolem constant form that can be bounded. The complete relevant information is defined as a set of skolemize clauses containing at least one skolem constant that is shared and bound to each other.

Keywords

Automated Information Extraction, Logical Linguistic Model

1.0 INTRODUCTION

This paper concerning a topic related to natural language understanding usually will only make clear its relationship to syntax, semantics, and pragmatics of the document domain and world knowledge for natural linguistic of computer science. The work is on exploring document understanding as a research domain and a method applied for natural language understanding system in question answering processes. We concentrated on the open-ended questions that involve the five *WH* questions. However, we will not tackle the negative types of question such as "WHY not" question. Negative type of questions cannot be answered directly from the traces of the reasoning processes of an expert system. The answers to the questions typically refer to a string in the text of a document and it only comes from the short document associated with the question. The information of extraction processes are restricted on the knowledge based, which are represented in simplified logical form known as Pragmatic Skolemized Clauses, based on first order predicate logic (FOPL) using Extended Definite Clause Grammar (X-DCG) parsing technique. The system used the logical reasoning approach of inference for knowledge based containing Skolemize Clauses in tracing the relevant information.

Information extraction is the question answering process of finding the relevant information to the question in a large text collection (Kim et al., 2000). In other words, the relevant information is not the whole document that is relevant to the question, but the parts of the document that can meet the questions' need more precisely. Question answering usually focuses on three areas: the content (an information repository of documents), the question, and the answer (McGuinnes, 2004). Implementation of question answering processes particularly is based on only some of the following components (Galitsky, 2003):

- Morphological and syntactic analysis.
- Semantic analysis, obtaining the most precise query representation.

- Pragmatic analysis, transforming query representation in accordance with the answer knowledge base.
- Answer knowledge base with built-in reasoning capability.
- Deterministic or statistical component which maps formal representation of questions into that of the answer.
- Information extraction component that forms the answer knowledge base.

Question asked is a method to construct meaning, enhance understanding, find answers, solve problem, find specific information, discover new information, propel research efforts and clarify confusion (Slater, 2004; Walters, 2004).

This work has been developed and tested using logical reasoning techniques by combining skolemize clauses binding to resolution theorem prover. The resolution theorem prover was implemented by Burhan in formal characteristic of answers (Burhans, 2002). The logical reasoning technique using in this work includes some changes and addition to the component such as:

- The logical inference engine – implementation of new inference of question answering called skolemize clauses binding (SCB) into existing resolution theorem prover technique. SCB module is considered as an inference technique that is used to provide explicit and implicit relevant information to the questions given by considering a theorem to be proven as a question.
- The phrase structure used to represent a clause – develop a simplified form of logical knowledge representation that is designed based on First Order Logic (FOL). The simplified form of logical-oriented model is known as Pragmatic Skolemize Clauses Representation (PragSC). It includes the event, object, properties of object, and the thematic role relationship between the event and the object in the sentence.

- The quantifier for the answer set – modify the skolem arguments to broaden the notion of answer literal to all contexts of question conducted, including universal quantifier and ground term. There are two symbols; f_n represent quantified variable names, while g_n represents ground term variable names.

2.0 IMPLEMENTATION

This work has been developed and tested using logical approach by combining skolemize clauses binding to resolution theorem prover. The resolution theorem prover was implemented by Burhan in formal characteristic of answers (Burhans, 2002). The logical reasoning technique using in this work includes some changes and addition to the component such as:

- The logical inference engine – implemention of new inference of question answering called skolemize clauses binding (SCB) into existing resolution theorem prover technique. SCB module is considered as an inference technique that is used to provide explicit and implicit relevant information to the questions given by considering a theorem to be proven as a question.

- The phrase structure used to represent a clause – develop a simplified form of logical knowledge representation that is designed based on First Order Logic (FOL). The simplified form of logical-oriented model is known as Pragmatic Skolemised Clauses Representation (PragSC). It includes the event, object, properties of object, and the thematic role relationship between the event and the object in the sentence.

- The quantifier for the answer set – modify the skolem arguments to broaden the notion of answer literal to all contexts of question conducted, including universal quantifier and ground term. There are two symbols; f_n represent quantified variable names, while g_n represents ground term variable names.

3.0 LOGICAL APPROACH IN INTELLIGENT INFORMATION EXTRACTION

In this section, a theoretical implementation of logical inference engine approach to question answering which refers to logical reasoning techniques is presented. Logical reasoning techniques by combining skolemize clauses binding to resolution theorem prover is a complete inference engine for knowledge base containing Pragmatic Skolemize Clauses representation. Providing information in a form of pragmatic skolemized clauses is just a method to collect the relevant answers. Proof start with the required goal, then resolution theorem prover is applied to provide the answer key by keeping track of variable as a proof proceeds. If the question asked has the logical form $\exists x P(x, y)$, then a refutation proof is initiated by adding the clause $\{\neg P(x, y)\}$ to the knowledge base. When the answer key is employed, the clause $\{\neg P(x, y), ANSWER(y)\}$ is added instead. The y in the answer key (ANSWER(y)) will reflect any substitutions made to the y in $\neg P(x, y)$, but the *ANSWER* predicate will not participate in (thus, will not effect) resolution. Then, the answer extraction proceed with skolemize clauses binding approach to continue tracking any relevant semantic relation rules in knowledge base, which contain the answer key in skolem constant form that can be bounded, formulated as $x \; ? \; P(x,x1) \; Ù \; P(x1,x2) \; Ù \; \ldots \; Ù \; P(xn\text{-}1,xn) \; Ù \; P(xn)$. The normalize skolem constant or atom is a key for answer depending on the phrase structure of the query. Given a key of skolemize clause in negation form and a set of clauses related in knowledge base in an appropriate way, it will

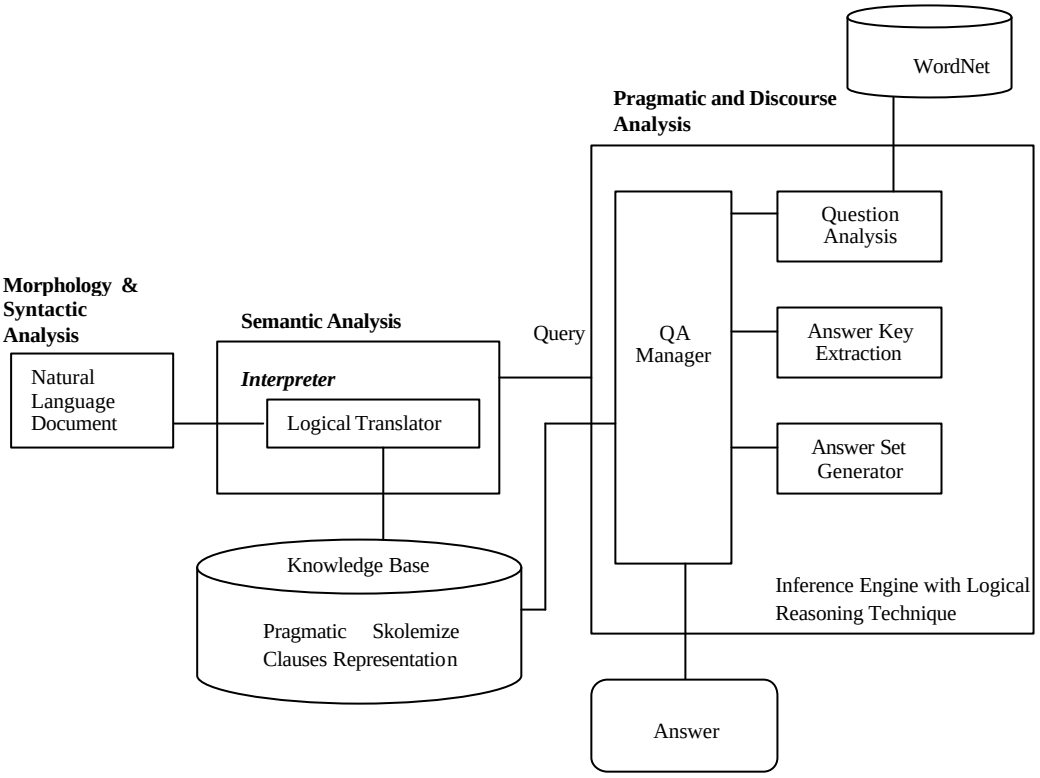

Figure 1: Description is placed right below the figure

generate a set of relevant clauses that is a consequence of this approach.

The system architecture, depicted in Figure 1, integrates different forms of knowledge analysis. This includes morphological and syntactic knowledge analysis for capturing the words in correct sentence form, semantic knowledge analysis for the meaning of words, and pragmatic and discourse knowledge analysis for preceding sentence affect the interpretation of the next sentence.

3.1 Morphological and Syntactic Analysis

Morphology analyzed the internal structure of words. Lexicon represents individual word in a sentence and words are generally accepted as being the smallest units of syntax. Syntax is the analysis of the rules that govern the way words are combined to form phrases and the way phrases are combined to form sentences. It also determines the structural role that each word plays in a sentence, and phrases as subpart of other phrases.

3.2 Semantic Analysis

The semantic knowledge analysis examines the meaning of each individual word. It covers the classification and decomposition of word meaning, the differences and similarities in lexical semantic structure, and the relationship of word meaning to sentence meaning and syntax.

For every single input in natural language, the interpreter translates the input into simplified logical form based on FOL. The conversion of a logical form into PragSC representation will be done for each question and answer document which will be used in extracting inferences using logical reasoning technique. The translation of logical form into PragSC involve six main stages; removing implications, moving negation inwards, skolemising, moving universal quantifiers outwards, distributing '$\wedge$' over '$\vee$', and putting into clauses.

3.3 Pragmatic and Discourse Analysis

The Pragmatic and discourse analysis part consists of a variety of modules: QA manager for handling a question

answering execution detail and also act as an intermediate module, the question analysis module for capturing the meaning of the natural language of question, the answer key extraction module for unification agreement, and the answer set generator module for generating a set of relevant answers. These integrated modules perform logical reasoning technique which can be considered as a primary part of the reading comprehension system architecture. The detail description of each module will be discussed in the subsequent sections.

3.3.1 Question Analysis Module

For each question, the question analysis module determines the question categories are conducted in this experiment and semantic of any existing hypernyms by consulting WordNet. The 5*Wh* questions categories as conducted in this experiment are the most straightforward approaches to comprehension.

If the key skolem clause of question is matched to skolemize clauses in knowledge base, the module returned *TRUE* to the QA manager. Otherwise, there is a problem such as some semantic of question does not really have an equivalent semantic relation rules to be matched, so the system attempt to establish hypernyms matching procedure.

3.3.2 Answer Key Extraction Module

A question to be answered is in negated form and the unification agreement proceeds between the key of skolemize clause and semantic relation rules in knowledge base. Two skolemize clauses are considered unified if their predicates and arguments match. This step may yield one or more skolemize clauses. An answer key is then added to each of these clauses. The answer key has the form *ANSWER(x`)*, where *x`* are all skolem constant or atom from the original question. This experiment proposed an expanded notion of literal answer that may be used with quantified and ground term variables. The key answer enables a resolution theorem prover to keep track of variable binding as the proof proceeds. Resolution can be visualized as the bottom-up construction of a search tree, where the leaves are the clause produced by knowledge base and the negation of the goal.

In the context of resolution theorem proving, the form of a question is often assumed to be a ground or an existentially quantified formula. Green (Green, 1969), discusses question forms in the context of applying the theorem model to general problem solving, including universally quantified questions and questions with mixed quantification, and his literal answer is designed for use with existentially quantified questions. Meanwhile, Burhans (Burhan, 2002), expand notion of answer that each reasoning step may be associated with the generation of an answer, and answer should be made available to return to a questioner as soon as they are generated. However, this experiment proposed the answer key which is designed for quantified and ground term questions. Answer Key Extraction module returns the answer key to QA manager as soon as they are generated.

3.3.3 Answer Set Generator Module

The Answer Set Generator module is to produce a list of relevant answer by executing the skolemize clauses binding. Its input is an answer key, and their associated relevant answer in skolemize clauses retrieved by connecting the key answer either in a form of normalize skolem constant or atom. Interest in classifying answer of logical reasoning technique has led to the focus on two types of answers: satisfying and hypothetical answer. Both of these answer clauses are associated with question whose logical form contains variables which are represented in normalize skolem constant. Each answer clause is considered informative with respect to a set of answer clauses.

4.0 RESULT OF EXPERIMENT

In order to show a desired answer for a variety of *wh* questions, the feasibility of intelligent information extraction system with a set of 115 documents is illustrated. This experiment also observed the effect of the logical reasoning technique, on the performance of automatic answer extraction with added skolem clauses binding as inference technique, by considering resolution theorem to be proven as a question. The scoring metric used for evaluation is HumSent, which is the percentage of test questions for which the system has chosen an exact answer in skolem clauses form. The HumSent answers are sentences that a human judged to be the best answer for each question. This metric is originally proposed by Hirschman (Hirschman et al., 1999). The experimental result in automatic answer extraction that is based on the methodology chosen will be given in Table 1, which the result breakdown of the questions answered correctly per question type.

Table 1: Performance of information extraction.

Question Type	Metric Score	Information Extraction Performance
WHO	0.896	0.861
WHAT	0.887	0.861
WHEN	0.922	0.852
WHERE	0.922	0.930
WHY	0.809	0.626
OVERALL	0.887	0.826

5.0 DISCUSSION

Since this experiment conducted a logical reasoning technique, there are two considerations related to the control of the inference engine. The first is determining when the resolution process should halt. Resolution theorem prover is designed to halt when an empty clause is generated, which makes sense when the goal of resolution is to find a proof. Therefore, when a literal answer is employed, a proof is associated with a clause containing only literal answers. Since this experiment's interest is in finding all relevant answers (whenever possible), stopping when the empty clause is generated is not appropriate. Instead, reasoning is halted when the set of support is empty. When the number of automatic answers is infinite, some alternative mechanism must be available to halt the reasoning process because the set of support will never be empty. The problem of an infinite reasoning process will appear when there is no proof. Prolog provides the semicolon option after the completion of a proof, and this gives the user the option of looking for an additional proof. This experiment conducted a new approach of searching strategy called skolemize clauses binding.

This second control issue known as skolemize clauses binding is the search strategy on which clauses should be chosen at any point to be resolved. Clauses are stored in knowledge base. In employing the search strategy, at least one clause to be resolved must be selected from the question clauses through its skolem constant binding.

The original version of the theorem prover incorporated unit preference, sorting the clauses in question clauses and knowledge base sets from smallest to largest length. The length of a clause is defined as the number of literals in the clause. Rather than relying on the built-in sorting of clauses, the user may elect to manually select the clauses to resolve at each step of resolution instead. By giving an additional proof to the original resolution theorem prover, the reasoning process allows to conduct search strategies and preferences may be examined.

6.0 CONCLUSION

This research aimed to study the best method to extract an exact answer on the information presented in the document, and any external knowledge sources that might be related to the subject. It is important that the QA system ascertained the difference between information that is stated directly in the document, and inferences and assumptions. The users of the QA system might be asked questions based on factual information found in the documents. The documents might also include information about which will be asked to make an inference.

Inferences — An *inference* is a conclusion based on what is stated in the document. You can infer something about a person, place, or thing by reasoning through the descriptive language contained in the document. In other words, the automatic system implies that something is probably relevant.

Assumptions — An *assumption*, on the other hand, is unstated evidence. It is the missing links in the text document.

There were several conditions of experiments conducted for improving the performance level of subjectivity analysis which extracted automated answers. Firstly, the study was designed to investigate the use of logical linguistic on question answering in reading comprehension tasks. The study used the semantic translation of natural language into a context-independent pragmatic known as Pragmatic Skolemize Clauses representation. Ideally, the semantic representation proposed adopted in its entireness to represent the semantic of passages and queries. Based on the theory of sentence understanding in reading comprehension, the semantic representation of individual sentence includes the event, object, properties of object, and the thematic role relationship between the event and the object in the sentences.

Secondly, this exploration of information extraction used the resolution theorem prover as an inference. In order to proceeded with this work an expanded notion of literal answer was used with the given question expressible in PragSC. Here, the universal and existential quantified, and ground term of literal answer within the theorem-proving paradigm was provided. In the expanded notion of literal answer, the restriction of information extraction broadened the capability of document understanding.

While the traditional function of resolution theorem proving remains that of providing a satisfying answer, it was demonstrated that a new approach to theorem proving could provide as an additional. It applied valuable information in the form of hypothetical answers by a composite of mixed with skolemize clauses binding. In general, the combination of both approaches known as logical reasoning technique was conducted and formulated to encode preferences for particular types of answers.

The foundation work presented in this paper will enable future work to be carried out, in terms of improvement to automated answer extraction using logical model construction, as well as provide a means of comparing current QA system. Despite the apparent variety of such system, the antecedents of question answering in theorem proving can be found in most, and the majority of their answers produced were categorized especially as satisfying. Information extraction as a process is viewed as a means of

generating a sequence of answer sets with a limited, converge of a complete answer set for a given query and knowledge base.

Finally, the employment of world or external knowledge sources forced the ability of the QA system. As indicated in the implementation of the experiments, previous cognitive psychologists (Golden & Goldman, 2005; Hirsch, 2003; Michael, 2007) argued that, there is a great effect of external knowledge sources on document understanding. Thus, just as is shown in this study, the QA system used without external knowledge sources portrayed the lowest percentage of ability in generating automated answer. The QA system with this phenomenon had difficulty to form a coherent situation model of reading expository text because the system was not able to generate the necessary inference.

The method used in determining the improvement of computer performance levels to produces an automated relevant information of document was based on additional of knowledge sources rather than on a well founded formulation. A better understanding of the effect of cohesion on computer system in document understanding will provide valuable insight and explicit direction on how to improve expository texts while taking into consideration the human performance skill. Therefore, this is open for further research to be undertaken, in order to produce better performance.

REFERENCES

Burhans, D. T. (2002). *A Question Answering Interpretation of Resolution Refutation.* Ph.D Dissertation. Faculty of the Graduate School, State University of New York, Buffalo.

Galitsky, B. (2003). *Natural Language Question Answering System*. 2. Adelaide: Advanced Knowledge International Pty Ltd.

Golden, T. M., & Goldman, S.R. 2005. ARCADE: Automated Reading Comprehension and Diagnostic Evaluation. *Poster presented at the REC Principal Investigator Meeting*.

Green, C. (1969). The Application of Theorem Proving to Question-Answering Systems. AI Center, SRI International.

Hirsch, E. D. J. 2003. Reading Comprehension Requires Knowledge - of Words and the World Knowledge. *American Educator*: 10 – 23.

Hirschman, L., Light, M., Breck, E., & Burger, J.D. (1999). Deep Read: A Reading Comprehension System. *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 325 – 332.

Kim, S. M., Baek, D.h., Kim, S.B., & Rim, H.C. (2000). Question Answering Considering Semantic Categories and Co-occurrence Density. *Proc. of the 9th TREC Conf.*, 317.

McGuinnes, D. L. (2004). Question Answering on the Semantic Web. *IEEE Intelligent Systems* Jan/Feb 2004: 82 – 85

Micheal, F. S. 2007. An Interview with Dan Willingham: Reading Comprehension. http://www.ednews.org/articles/6696/1/An-Interview-with-Dan-Willingham-Reading-Comprehension/Page1.html [2 February 2007].

Slater, A. (2004). Reading Comprehension: Strategies to help children unlock the meaning. http://www.nevadarea.org/presentations/comprehension-slater-group.ppt [24 March 2005].

Walters, F. S. (2004). *An Application of Conversation Analysis to the Development of a Test of Second-Language Pragmatic Competence*. Masters Thesis. Department of Educational Psychology, University of Illinois, Urbana-Champaign.