Natural Language Question Answering Model Applied To Document Retrieval System

Nguyen Tuan Dang, and Do Thi Thanh Tuyen

***Abstract*—**In this paper, we propose a method to build a specific Question-Answering system which is integrated with a search system for eBooks in library. Users can use simple English questions for searching the library with information about the needed eBooks, such as title, author, language, category, publisher… In this research project, we focus on fundamental problems of the natural language query processing: approaches of syntax analysis and syntax representation, semantic representation, transformation rules from syntax structure to semantic structure… Some first results let we believe that our system model can meet some strict requirements and it can be applied to develop other similar searching system.

***Keywords*—**Question Answering System, Natural Language Processing, Information Retrieval.

I. Introduction

IN Question Answering (QA) domain, many QA systems, such as START, AskMSR, NSIR ... have been developed to support the users searching the correct information about some topics. Among these systems, START may be considered as the best system that can return the good answers for users. However, START is only able to answer questions about concepts, but it could not answer the questions about causes and methods.

Furthermore, we had investigated an open source QA system. That is OpenEphyra; it is an open framework for question answering (QA). It retrieves answers to natural language questions from the Web and other sources. It was developed on Java framework. This system has: a dictionary, a set of questions, method to find out the correct answers for questions. Once receiving a question, the system classify question into one of defined categories of question to analyze and split keywords base on the dictionary. After that, OpenEphyra uses these keywords to search in data set paragraphs that contain them. The result will be estimated adequate degree and the best result will be display to user.

After considering these QA systems, we assume that current QA systems perform the question processing base on this principle:

- Match the query to existing queries form.
- Generate a set of keywords or a set of queries in knowledgebase.
- Determine the result that fit to the question and generate the answer.
- Without semantic model of query.

In our research project, we specially focus our interests on building a particular QA system model appropriately used in the domain of document retrieval.

II. Document Retrieval Based on Question Answering

Two main purposes of a document retrieval system based-on question answering model:

- Processing the natural language query of user.
- Searching the relevant information.

From these goals, the system model aims to perform the following principal tasks:

- Get simple English input query from user.
- Process the query.
- If the system cannot process a query, it will ask user to express this query in another way or it suggest user to choice one of some similar queries.
- Searching eBooks.
- Answer to the user.

The schema of system functioning is introduced in Fig. 1.

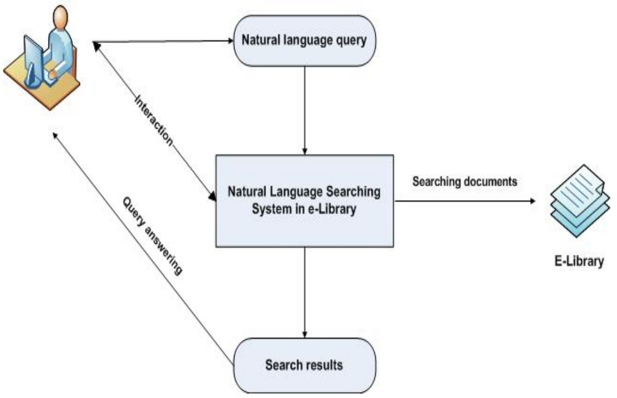

Fig. 1 System functioning

For processing the above tasks, the system is designed with three main parts:

1. Natural Language Query Processing: resolve the syntax and semantic representations of the natural language query.
2. Document Catalog Database Searching: transfer semantic performance of natural language query to a set of database query, and then implement them.
3. Query Answering: filter, organize, generalize and return the search results by defined criteria according to user's query.

Nguyen Tuan Dang is with Faculty of Computer Science and Do Thi Thanh Tuyen is with Faculty of Software Engineering, University of Information Technology, Vietnam National University – HCMC (e-mail: {dangnt, tuyendtt}@uit.edu.vn).

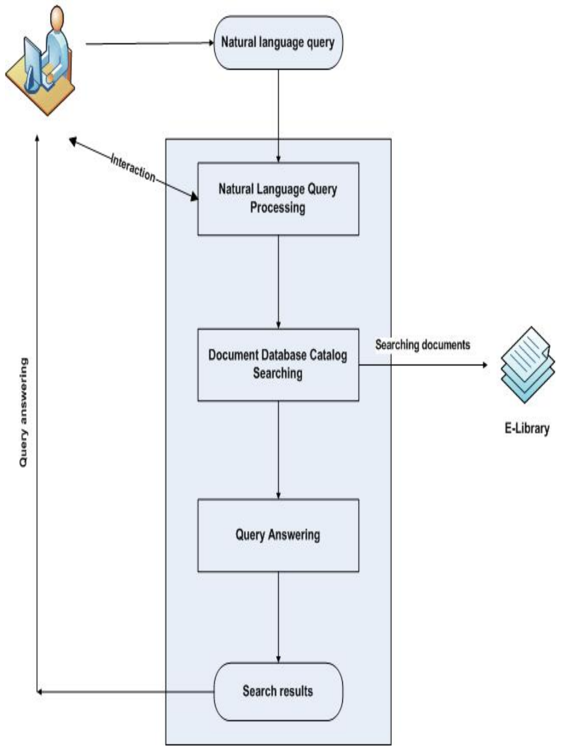

Fig. 2 System Architecture

Overview architecture of the system is presented in Fig. 2.

III. Question Answering Model

The objective of natural language query processing is to build the syntax representation of query and transform it to a semantic representation.

We have built a model of natural language query processing that is illustrated in Fig. 3.

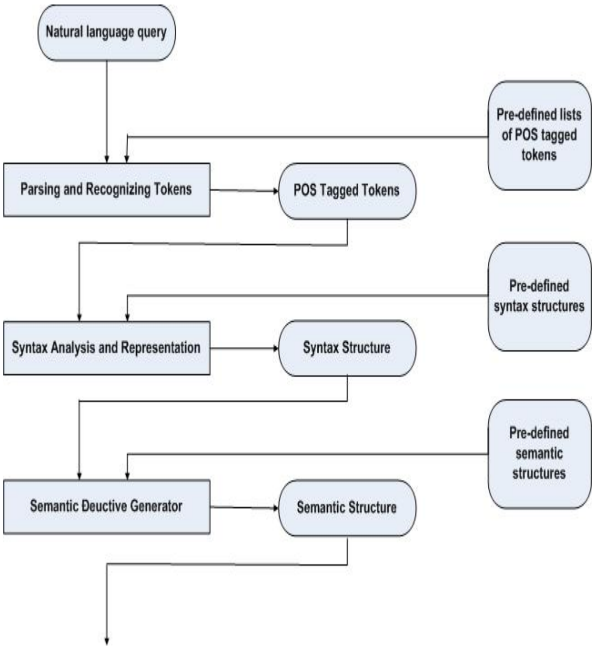
Natural language query
Pre-defined lists of POS tagged tokens
Parsing and Recognizing Tokens
POS Tagged Tokens
Pre-defined syntax structures
Syntax Analysis and Representation
Syntax Structure
Pre-defined semantic structures
Semantic Deuctive Generator
Semantic Structure

Fig. 3 Natural language query processing

The kernel of a question answering model is the processing of natural language search queries. Basing on the fundamental ideas of W. Chafe (1971), we suggest a method to process the simple English queries. In his language theory, Chafe proposed some main concepts about the syntax structures, semantic structures and transformation rules between them.

According to Chafe, the syntax model is built on the concept of relationship between words, part-of-speech and between syntax factor such as subject, verb, object, complement. The semantic model is built on Chafe's point of view about semantic structure; it is defined as the relationship between verb and its arguments. These arguments are nouns that have a semantic role in semantic structure, for example: agent, object, patient... Transformation rules between syntax model and semantic are defined base on some scenarios of search.

We proposed a method for analyzing English queries that is based on the following principles:

- Identifying a quite large number of user queries about information of eBooks: title, author, language, category, publisher ... (about 300-500 queries).
- From the listed queries:
  - Defining the set of words (vocabulary) can be used. Next, classifying the words in: nouns, verbs, prepositions, conjunctions and interrogatives ...
  - Determining all forms of query. These query forms will be mapped to some pre-defined syntax structures.
- For nouns, continuing to distinguish and label them by: author, book, publisher, title, language ...
- Pre-define:
  - Some structure to perform the syntax model of queries.
  - Some structure to represent the semantic structure of queries.
  - Transformation rules used to convert from syntax into semantic structure.
- Analyzing the search query and recognize the tokens.
- Reorganizing and mapping the recognized tokens into one of syntax structures.
- Transforming the syntax structure into pre-defined semantic structure.
- Generating a set of database queries from the semantic structure.
- Executing the database queries and gets a set of results.
- Filtering and organizing the search results.
- Generate the answer to return to user.

These principles are described on the details as follows:

A. *Pre-processing*

This phase prepares the vocabulary, pre-defined syntax and semantic structures and transformation rules between syntax structures and semantic structures.

1. Building the Word Set

The task of this step is to determine a list of words in categories: nouns, verbs, prepositions, conjunctions and interrogatives.

- Nouns: information about title, author, language, category, publisher ...
- Verbs: write, publish, have, is, create ...
- Prepositions: in, by…
- Interrogatives: what, when, who …
- Conjunctions: and, or …

2. Pre-defining some of Syntax Structures

This step aims to pre-define some of syntax structures, for example:

- Noun + verb + noun
  Creator is Clinton
- Noun + preposition + noun
  The book about computer science
- The passive voice
  The book was written by Clinton
- Noun + conjunction + noun
  Computer or Information
- Noun + verb + preposition + noun
  The book created after 2005

…

The pre-defined syntax structures will be used in order to map the syntax structure of query into them.

3. Pre-defining some of Semantic Structures

The sematic structure of query is based on the meaning of verb. Each verb requires the presence of agruments in the relationship with verbs. The structure including verb, nouns and semantic relationship between verb and nouns is called semantic structure. For example, the verb *write* requires a semantic structure with two nouns (arguments): an *agent* (author) and a *object* (book).

Thus, we pre-define some of semantic structures, for example:

- Book – written by – author/authors
- Book – written in – language/languages
- Book – has – title
- Book – created on/in – date

...

4. Pre-defining Transformation Rules

These transformation rules are used to map syntax structures to semantic structures.

B. Parsing and Recognizing Tokens

The system analyzes and recognizes the tokens. These recognized tokens will be matched with pre-determined words in nouns, verbs, prepositions, conjunctions and interrogatives lists. All unmatched tokens are eliminated and ignored.

C. Syntax Analysis and Representation

The words determined and labeled by their part of speech in last step will be mapped into pre-defined syntax structure forms.

D. Semantic Deductive Generator

Syntax structures will be transformed into sematic structures by the Semantic Deductive Generator component. The important principles are described as follows:

- Determining verbs in syntax structures.
- Listing the defined semantic structure that corresponding to determined verb. The system only performs on the verb in active voice and passive voice. Tenses of verbs are not considered at present.
- Depending on number and positions of nouns in the syntax structure, eliminate the semantic structures that have number of agruments which is inapproriate to the syntax structures.
- Depending on number and positions of prepositions, conjunctions in syntax structure to eliminate the inappropriate semantic structures.
- Determining the most appropriate semantic structure for a syntax structure.

E. Document catalog database searching and query answering

The Document Catalog Database Searching component base on two parts:

- Database Queries Generator (DQG) module.
- Database Retrieval (DBR) module.

From the semantic representation model of query, the DQG will generate a set of database queries (SQL commands in eLSSNL). The set of database queries will be executed to get results.

The Query Answering component is composed of two parts:

- Search Results Filtering (SRF) module.
- Database Queries Generator (DBQG) module.

The results of database search will be filtered, organized before the AG generates the answer in order to return to user. The answer bases on sorting of found books by titles or grouping them by some of attributes: author, published year, publisher...

IV. Experience

We developed a testing system eLSSNL (eLibrary Searching System by Natural Language) working on the free eBook library Gutenberg. This system accords with the document retrieval system model in Fig. 2.

At present, the eBook library of Project Gutenberg has more than 25.000 free eBooks. This project uses e-texts (electronic texts) in “Plain Vanilla ASCII” to store the content of books. Information about the eBooks is updated frequently and stored in an only file (*Catalog.rdf*).

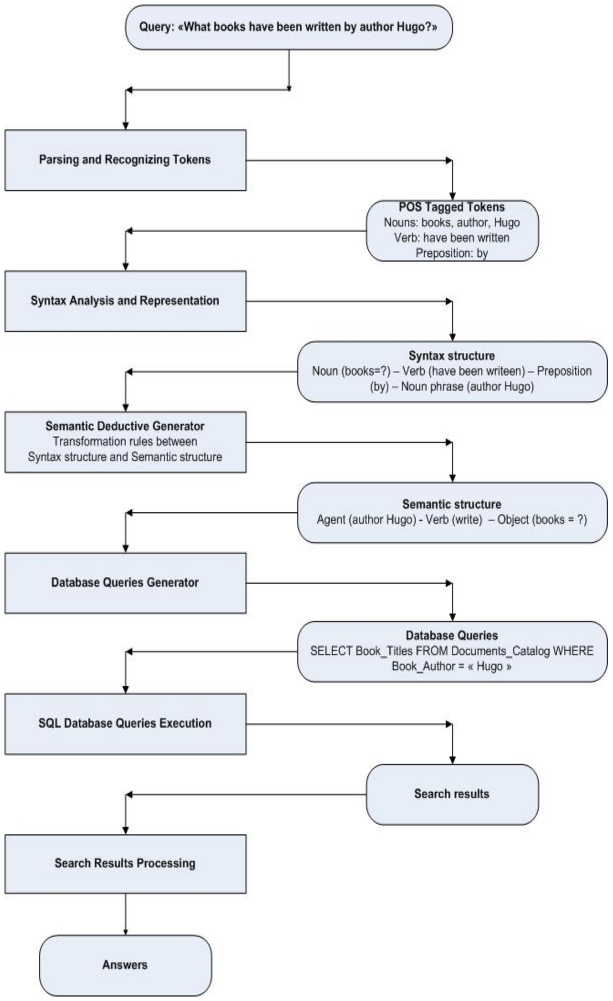
Query: «What books have been written by author Hugo?»
Parsing and Recognizing Tokens
POS Tagged Tokens
Nouns: books, author, Hugo
Verb: have been written
Preposition: by
Syntax Analysis and Representation
Syntax structure
Noun (books=?) – Verb (have been writeen) – Preposition (by) – Noun phrase (author Hugo)
Semantic Deductive Generator
Transformation rules between
Syntax structure and Semantic structure
Semantic structure
Agent (author Hugo) - Verb (write) – Object (books = ?)
Database Queries Generator
Database Queries
SELECT Book_Titles FROM Documents_Catalog WHERE Book_Author = « Hugo »
SQL Database Queries Execution
Search results
Search Results Processing
Answers

Fig. 4 Example of natural language query processing

V. Conclusion

We have developed the language theory model of W. Chafe to establish a method for processing English queries. This method is based on three main parts: a syntax model, a semantic model and the transformation rules between them. We have also applied this method into a system which supports user to use simple English queries. Some first results demonstrated that our model can meet some strict requirements and it can be applied to develop other similar searching system.

We have proposed some problem that needs to be improved in the system model:

- Some problem in word segmentation, POS tagging is needed to be performed in a more generally way in order to apply this model for wider domain of application. In our model, these problems are solved by a technical solution. It is defining a dictionary for the system and assigning POS label for words. This is only suitable in case of a specific application with some clear information about structure of data and predictable searching scenarios.
- The method of generating database queries is needed to be estimated and optimized.

Our next step in system developing focus on building the user interaction mechanic in order to the system may deal better with non-understand queries.

References

[1] Enrique Alfonseca, Marco De Boni, José-Luis Jara-Valencia, Suresh Manandhar, “A prototype Question Answering system using syntactic and semantic information for answer retrieval”, Proceedings of the 10th Text Retrieval Conference, 2002.

[2] Carlos Amaral, Dominique Laurent, “Implementation of a QA system in a real context”, Workshop TellMeMore, November 24, 2006.

[3] Eric Brill, Susan Dumais, Michele Banko, “An Analysis of the AskMSR Question-Answering System”, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002.

[4] Boris Katz, Jimmy Lin, “Selectively Using Relations to Improve Precision in Question Answering”, Proceedings of the EACL 2003 Workshop on Natural Language, 2003.

[5] Boris Katz, Beth Levin, “Exploiting Lexical Regularities in Designing Natual Language Systems”, Proceedings of the 12th International Conference on Computational Linguistics.

[6] Callison-Bruch, Chris, A computer model of a grammar for English questions, Undergraduate honors thesis, Stanford University, 2000.

[7] Nguyen Kim Anh, “Translating the logical queries into SQL queries in natural language query systems”, Proceedings of the ICT.rda’06 in Hanoi Capital, 2006.

[8] Nguyen Tuan Dang, Do Thi Thanh Tuyen, “E-Library Searching by Natural Language Question-Answering System”, Proceedings of the Fifth International Conference on Information Technology in Education and Training (IT@EDU2008), Pages: 71-76, Ho Chi Minh and Vung Tau, Vietnam, December 15-16 , 2008.

[9] Nguyen Tuan Dang, Do Thi Thanh Tuyen “Document Retrieval Based on Question Answering System”, accepted paper, The Second International Conference on Information and Computing Science, Manchester, UK, May 21-22, 2009.

[10] Riloff, Mann, Phillips, "Reverse-Engineering Question/Answer Collections from Ordinary Text", in Advances in Open Domain Question Answering, Springer Series: Text, Speech and Language Technology , Vol. 32, 2006.

[11] Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, “Exploiting Paraphrases in a Question Answering System”, Proceedings of the second international workshop on Paraphrasing, 2003.