The Development of a Question-Answering Services System for the Farmer through SMS: Query Analysis

**Mukda Suktarachan,
Patthrawan Rattanamanee**
Department of Computer Engineering, Kasetsart University, Bangkok, Thailand, 10900

naist_da_da@yahoo.com,
tiptop317@hotmail.com

Asanee Kawtrakul
Department of Computer Engineering, Kasetsart University, Bangkok, Thailand, 10900
National Electronics and Computer Technology Center, Thailand

asanee_naist@yahoo.com
asanee.kawtrakul@nectec.or.th

Abstract

In this paper, we propose the development of the Question-Answering Services System for the Farmer, through SMS, by focusing on query analysis and annotation based on a similar technique previously applied to language generation, thematic roles, and primitive systems of the Lexical Conceptual Structure (LCS). The annotation places emphasis on the semantics model of "What" and "How" queries, lexical inference identification, and semantic role, for the answer. Finally, we show how these annotations and inference rules contribute to the generalization of the matching system over semantic categories in order to have a large scale question-answering system.

1 Challenges and Goals

In the era of Information and Communications Technology (ICT), mobile is a fast and convenient way to communicate over a network. Knowledge service via a mobile as "a right information for a right man" is a challenging task. However, this means of interchange between persons has the limitation of personal timing. Therefore, Short Message Service (SMS) is a better way for giving knowledge service, especially automatic interchange of short text messages, by providing the information from an automatic Question & Answering System.

From the results of the statistical ICT data survey concerning the number and percent of the population 6 years of age and over who use information and communication technology: 2003 - 2007 by the National Statistical Office[1], Thailand, it was found that 47.2% of people in the entire kingdom have owned their mobile(s). Consequently, communicating via SMS facilitates an effective knowledge service for supporting the farmers in problem-solving, decision making, and early warning, and also supports the government, or a related organization, in order to e-communicate to the farmer by changing the model of "Training and Visit" to e-service and changing the collective to support cooperative problem solving. This kind of communication will provide the necessary long-term cost reductions to the agricultural economy in the areas of travel, visiting, productivity, etc.

Nowadays, providing a knowledge service through SMS is not limited to only a Question-Answering Services System, but also for such one-way services as early warning systems, for example, a Tsunami Alert System[2], a FloodSMS – Early Detection and Warning of Catastrophic Flooding via SMS[3], etc.

The development of a Question-Answering Services System through SMS is not the design of a new technology. There have been several theories developed earlier, in the context of NLP or cognitive sciences, such as Natural Language Information Retrieval (NLIR), rule based Q&A, etc. Nevertheless, some former theories of Q&A relied on complex semantic information. For instance, a Wireless Natural Language Search Engine [6] was implemented using a system resid-

---

[1] http://web.nso.go.th/en/survey/keystat/keystat08.pdf
[2] http://www.wap.ait.ac.th/tsunami.html
[3] http://www.netsquared.org/projects/floodsms-%E2%80%93-early-detection-and-warning-catastrophic-flooding-sms

ing on a server, which can translate questions or phrases into search engine queries or queries to SOAP Web services, where a gateway mediates between the mobile network and the Internet. Also, [15] developed the SMS for Question-Answering in the m-Learning Scenario System by using the Simple Matching Algorithm to match the learners' answer messages with the original answer string, thus facilitating the learners to get the necessary feedback and assessment.

In this paper, we propose the development of the Question-Answering Services System for the Farmer through SMS by focusing on query analysis and annotation, as well as on selected text matching utilizing lexical inference and semantic roles. The annotation emphasizes the semantics model of "What" and "How" queries. Finally, we show how these annotations and inference rules contribute to the generalization of the matching system over semantic categories in order to have a large scale question-answering system.

In the current stage, we have designed Q&A schema with thematic roles and have borrowed some primitive systems of the Lexical Conceptual Structure (LCS). Also, we are annotating 1000 questions and text related to the query (but we randomly choose 100 pairs of Q&A for the experiment). In the same time, we are generalizing inference rules in order to match a question to its answer. This is particularly crucial when there is no straightforward response, e.g. when they require some form of lexical inference, elaboration, and reasoning or when the response is not a simple item, but a well-formed fragment of text, e.g. a chain of events leading to a consequence, a procedure, etc.

The project we present here emerged from a need of the real end-users, the Agricultural Land Reform Office, Ministry of Agriculture and Cooperative, Thailand, in the project of ALRO CyberBrain [3], which is a social network framework that combines approaches based on knowledge science and engineering with language engineering, consisting of an ontology-based search engine, information extraction for Q&A system, knowledge aggregation through a knowledge portal and visualized in a browser with semantic links between problems, methods of problems solving and man who is the problem solver (PMM map Model) [1]. The main goal is to develop tools for e-Farming, in particular rice farming, so that farmers can easily get information on farming rice and rice diseases. Now, it has been

extended to provide question-answering services for the farmers through SMS [2].

2 Problem Statements

There are two main problems in Q&A analysis: semantic interpretation for a question word and answer identification.

2.1 Question's Semantic Roles

2.1.1 Question Word Interpretation.

In general, when we query for the answer by a traditional search engine system, we might get many answers at different levels, depending on the role of the question: Definition vs Fact or set of Facts. For example, with the question

Q1: "โรค|ไหม้|คือ|อะไร"

 Rice| Blast| is| what
 "What is a Rice Blast?"

The answer can be returned as the definitions, fact or a set of facts, which are:

A1.1: Blast, also called rotten neck, is one of the most destructive diseases of Missouri rice. Blast does not develop every year but is very destructive when it occurs.[4]

A1.2: Disease of Leave Burnt caused by Pyricularia Oryzae can destroy all rice growing period from start until harvest period.[5]

The answer can be returned as the characteristics detail or set of facts, such as the following:

A1.3 : Blast symptoms can occur on leaves, leaf collars, nodes and panicles. Leaf spots are typically diamond shaped, with gray- white centers and brown to red-brown margins. Fully developed leaf lesions are approximately 0.4 to 0.7 inch long and 0.1 to 0.2 inch wide. Both the shape and color vary depending on the environment, age of the lesion and rice variety.[6]

2.1.2 Variety of Question Forms.

In natural language, the question can be asked with different words and styles, for example:

Q2.1:ลักษณะ|การระบาด|ของ|โรค|ไหม้|เป็น|อย่างไร

 situation| outbreak| of| disease| Rice Blast| is| how
 "What is the situation of Rice Blast?"

Q2.2:การระบาด|ของ|โรค|ไหม้|มี|ลักษณะ|อย่างไร

outbreak| of| disease| Rice Blast| is| characteristic| how
 "How does the rice blast outbreak look like?"

---

4 http://aes.missouri.edu/delta/muguide/mp645.stm
5 http://www.sotus.co.th/article_4.html
6 http://aes.missouri.edu/delta/muguide/mp645.stm

Q2.3:โรค|ไหม้|ระบาด|ได้|อย่างไร

    Rice Blast| disperse| able| how

    "How can the rice blast disperse?"

The reply can be returned the same answers with a descriptive set of events, as the following:

A2.1: To prevent the Rice Blast: for the places that we often found the disease, use the disease-resistant rice variety. Don't sow the rice seed too densely. Don't use to much Nitrogen. If it is severe outbreak and it is the state of young plant, plow and sow again. If it was the epidemic state, use Fungus-Removal chemical as Carbendasim.

A2.2: Brown spot may be reduced by balanced fertilization, crop rotation, and the use of high quality planting seed. Seed treatment fungicides reduce the incidence and severity of seedling blight caused by this fungus.

The examples above show that using different verbs or noun phrases can be represent the same meaning. Moreover, there is non-correspondent focus word between Q and A.

2.2 Answer Type Identification

2.2.1 Ambiguity between subtopic and answer form

To identify the answer, sometimes there is an ambiguity that verb phrases occurring after the focus word of the question can be both subtopics and the answer, like a procedural answer, for example,:

Q3: วิธีการ|ป้องกัน|โรค|ไหม้|ทำ|ได้|อย่างไร

    method| control| Rice Blast| to do| how

    "What method can be used to control Rice Blast?"

A3: วิธีการ|ป้องกัน|โรค|ไหม้|มี|ดังนี้

    method| control| Rice Blast| have| such as

    "Methods for preventing the Rice Blast are:"

- ใช้|สารเคมี|ที่|เหมาะสม

    use| Chemical Substance | that| appropriate

    "Use appropriate Chemical Substance."

- ใช้|พันธุ์|ที่|เหมาะสม

    use| type of rice | that| appropriate

    "Use appropriate type of rice."

- ใช้|กลไก|ใน|การ|ป้องกัน

    use| mechanism| in| prevent

    "Use mechanism to prevent."

- ใช้|วิธีการ|ผสมผสาน

    use| methods| hybrid

    "Use hybrid methods."

The examples above convey the 4 types of method for Rice Blast control or names of methods, but it is not the process or the answers that represent how to control the disease.

2.2.2 Non-correspondence between Q & A:

Sometimes, the question and answer were not matched because the clue words or focus words in the question have never appeared in the answers. This makes the question not correspond to the answer and also causes difficulty in finding the expected answer. For example,

Q4: สามารถ|ควบคุม|แมลง|ศัตรู|ข้าว|เหล่านี้|ได้|อย่างไร

    can| control| pests | rice | these | How

    "How can these rice pests be controlled?"

A4: แมลง|ศัตรู|เหล่านี้|สามารถ|กำจัด|ได้|โดย|ใช้|วิธีการ|หลาย| ประการ|รวม|กัน| |คือ| |ปลูก|ข้าว|พันธุ์|ต้านทาน|โรค|,| |ปลูก| ข้าว|ตาม|ฤดูกาล|, |ควบคุม|ระดับ|น้ำ|ใน|นา |และ|ไม่|ควร|ใช้| สาร|ฆ่า|แมลง|ที่|ทำ|ให้|เกิด|การ|เพิ่ม|ระบาด||ของ|โรค|

    "These pests can be managed through integrated approach including sowing insect resistant rice varieties, sowing rice crop at recommended time, proper water management conservation and augmentation of bio-control predators."

From the example, the focus word of the question is "control," but there is no word "control" in the answer. For this kind of Q&A matching solution, WordNet and ontology are necessary.

3 Outline of the Project and Methodology

The needs of the Thai Ministry of Agriculture have been specified in a simple way via a corpus composed of (1) questions raised in real life by farmers (about 1000 questions), (2) the responses which have been provided by experts, based on existing documents (possibly several responses per question) and, quite often, (3) the texts they originate from. In general, the response is found in a unique text: there are no multiple answers, since most texts are not redundant, although some responses, in particular complex (e.g. evaluative questions) or indirect ones, may involve the taking into account of several independent texts. We will not address here the problem of message length reduction so that it fits into an SMS format (although this is also an important semantic problem).
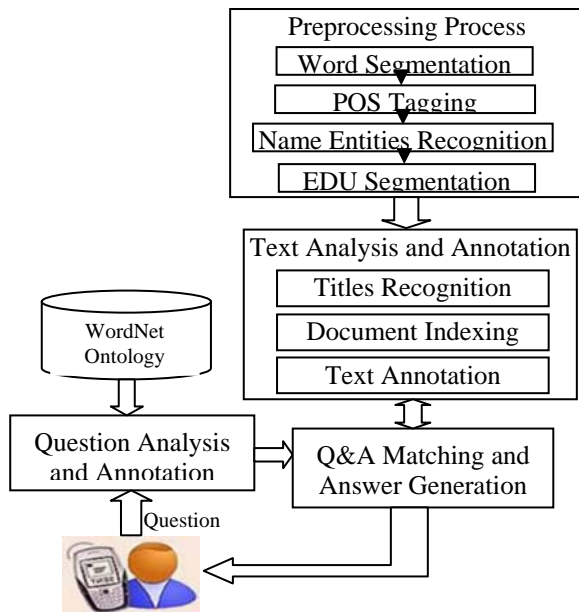
The system overview is shown in Figure 2.

5

Figure 2. System Architecture

To develop a Thai QA system, the preprocessing of Thai morphology and syntax is necessary. The NAIST lab at the University of Kasetsart has basic tools to manage morphological analysis, parts-of-speech recognition, simple syntactic analysis, as well as Thai parsing, and an Element Discourse Unit System (EDU). These tools were designed as basic tools in natural language processing applications. (accessible on http://vivaldi.cpe.ku.ac.th:9292/ with a recommendation to use the Mozilla Firefox browser)

A few examples of question-answer pairs are:

Q5: วิธีการ|ป้องกัน|กำจัด|ข้าว|วัชพืช|ทำ|ได้|อย่างไร

"How to prevent the Weedy rice"

A5.1: ควร|เว้น|ปลูก|ข้าว|บาง|ฤดู

"Skip some seasons when growing rice,"

A5.2: ปลูก|พืช|อายุ|สั้น|อื่น

"Grow hydrotonics plants."

Q6: โรค|ใบ|ขีด|โปร่ง|แสง|มี|วิธีการ|ควบคุม|อย่างไร

"How to control the Bacterial Leaf Streak Disease"

A6: ไม่|ควร|ใส่|ปุ๋ย|ใน|โตรเจน|มาก|เกินไป|

"Do not put too much Nitrogen."

Q7: ถ้า|พบ|เพลี้ย|ไฟ|ระบาด/   /สามารถ|กำจัด|ด้วย|สาร/ ฆ่า|แมลง|ชนิด|ใด

"How to eradicate the rice thrips"

A7: ฉีด|พ่น|ด้วย|มาลาไทออน|หรือ|คาร์บาริล|ทุก|อาทิตย์| | ใส่|ปุ๋ย|และ|ให้|น้ำ|ทุก| |ๆ| |สอง|วัน

"Spray with Malathion or Carbaryl every week, add fertilizer and water every two days."

Questions are essentially factoid questions (e.g. best periods for rice planting, rice varieties suggestion, symptoms of a disease), why ques-

tions, where responses are chains of events (reasons for something to happen) and a large number of procedural questions [4], in particular for treating diseases. There are relatively few comparative or evaluative questions besides general questions, such as: What are the major rice pests?

In most cases, questions do not have responses which can be immediately found in the texts by standard term matching techniques. For example: *"How does the Sheath Blight affect the rice growth?"* has the following response in a text: *Plants heavily infected at these stages produce poorly filled grain, particularly in the lower portion of the panicle. Additional losses result from ...* Therefore, some lexical semantics devices (e.g. a semantic link between *affect* and *infect*) or more elaborated reasoning schemas, based on domain knowledge, are needed to allow appropriate question-text matching [11, 7]. The kind of domain knowledge at stake may be quite unexpected (i.e., not the main topics that everyone knows, but more subtle pieces of information, as will be seen in 4.3). This is the major challenge of this work, which we try to resolve via a full annotation of the matching process, from question parsing to response production, identifying matching and reasoning aspects.

Complex questions may, e.g., require the elaboration of a diagnosis from premises given in the question before finding the response, either factoid or procedural (*My rice has weedy leaves and some yellow spots, what should I do?*). This question requires one to select all texts where such a symptom is identified, and then, e.g., to enter into a dialogue with the user if there are several possible diagnoses, leading to different treatments.

The second aspect of this problem is to be able to extract the complete text portion that responds to the question. For that purpose we are developing an annotation methodology whose goal is to identify the different processes at stake and the needed resources. This method allows us to identify relevant text portions and then to delimit them appropriately.

4   The Question-Answering Process Annotation

Since the task is quite large (a large group of students are annotating a set of 600 questions and related texts), we need to establish norms and annotation guidelines. Using the research conducted at IRIT on annotating procedural questions and instructions based on semantic roles

(TextCoop project) and a few rhetorical relations (e.g. elaboration, example, explanation), we first annotated the questions and their corresponding responses in texts provided by the Thai Ministry of Agriculture. One of the challenges was to identify relevant linguistic marks or patterns [9, 10, 14].

There are many attempts to annotate arguments by means of primitives; our approach, here, is oriented towards the precise task at stake and the specific actions. Therefore roles are not as standard as they are in general. An earlier attempt with a similar technique applied to language generation was carried out in, e.g., [10, 7]. Semantic tags are either close to thematic roles (instrument, location, etc.) [8], or borrowed from the primitive systems of the Lexical Conceptual Structure (LCS) [13], in particular, to establish useful links between arguments or between a large variety of constituents, which thematic roles cannot do. For example, in *the first Thai university* we have a link between 'first' and 'Thai university' which is either loc_{temp} or $loc_{+char+ident}$, depending on the interpretation of *first* (oldest or the best). However, in a majority of cases, semantic roles based on thematic roles have a sufficient granularity, and these are the ones which are used in the examples in 4.1.

The main roles we consider are: agents (for humans and animals like insects, and metaphorically for diseases and natural forces), themes (undergoing actions, basically plants and soils, and artificial products), location (spatial), time (covering dates and also periods), instruments (from tools to chemical products), manners, means, conditions (under which to realize an action, or related to observation e.g. of a disease), cause, goals, and results.

Besides, the tags <action>…</action> or <fact>…</fact> were considered to tag the verb with it's arguments or adjuncts.

In the remainder of this section we briefly report the different steps of the process as they stand at the moment, i.e. almost at the end of the experimental stage, before automating knowledge acquisition, and implementing the application.

4.1 Dealing with Questions

As in most systems dealing with complex types of questions, questions are represented by a triple: the question type (which can be in our case polymorphic), the question focus (usually an NP or a VP in case events or procedures are induced)

and the question body, annotated by means of semantic roles, as indicated above.

The main types of questions we have identified from our corpus are the following; they are quite different from standard classifications, but they correspond to more operational views:

F: fact, with subtypes: temp (temporal, time, date), loc (location) or product,

E: an event (with a subtype event: cause)

SF: set of facts

SE: set of events (not related, and without any form of sequence: different from SqE below)

PROC: procedure, more or less complex, it may be just a single instruction; it can also describe the use of an instrument.

SqE: sequence of events, which follow each other.

EVAL: evaluation, making value decisions about issues or resolving controversies or differences of opinion.

DEF: definition, the description of object.

Some questions may bear several non-conflicting types, in particular when the nature of the response is not straightforward to determine from the question. For example, *"What is the symptom of Bakanae?"* would get the types SF and SE.

An annotated question is, for example:

<question type=" SF or SE" focus=" symptom of *Bakanae*"> What <fact> is <theme> the symptom of *Bakanae* </theme> </fact> ? </question>

As can be noted, the response is the set of those facts that contribute, together or independently, to the spreading of the disease.

By the observation from 100 random interrogative sentences corpus analysis, we found that the semantic types of questions correspondent to the question words are the following

Q-Types	What	When	Where	Why	Who	Which	How
F	11	6	1		1	9	2
E				1			6
SF	3					15	3
SE	7					2	5
PROC							15
SqE							7
EVAL	2				1		
DEF	3						

Table 1 the correspondence between questions and semantic types of questions

From Table 1, it is clear that "what" and "how" questions vary in types of question, because they have many forms to use, for example, *"how + verb to be + noun", "how + do(es) + noun + verb", "how to", "how can", etc.* or *"what + verb to be + noun", "what + noun + auxiliary verb",* etc. This is why we point out the "What" and "How" questions.

4.2 Dealing with texts: document indexing and associated annotations

Texts are initially indexed based on the main terms they contain which are relevant w.r.t. the questions given in the corpus. Our representation resembles a frame approach, but it is more flexible since there is no predefined structure to represent indexes. This is more in accordance with the variety of texts in terms of contents. Indexes basically are formed from:

- Top-level terms that structure the domain: for example, concepts like symptom, spreading, treatment, time, place, effect, etc. where predicative (action terms) terms as well as entities are found,
- relatively generic terms, found in the questions and structured in the domain ontology: water, clean, control, eradicate, etc., which are organized w.r.t. the top concepts above,
- named entities, typed as: disease names, location names, chemical product names, bacteria names, etc.

In our representation, those generic terms (and near synonyms) are represented as predicates, while arguments are represented as attribute-value pairs (or attributes alone), include typed name entities and any kind of terms besides the generic terms.

Indexes are associated with texts in the text database. Indexes must remain general so that indexing is fast and as reliable as possible. The idea is that when a question is uttered, a small number of texts are first selected on the basis of the indexes for further analysis. An example below can be indexed and annotated [2] as the following:

Index: disease-name (Bakanae), symptoms (disease: Bakanae), origin (disease: Bakanae, place: California, date: 1999), spreading(disease: Bakanae, period: winter, medium: [soil, water]), treatment(disease: Bakanae, product).

<title type="goal" level="1" > Rice Bakanae </title>
<title type="goal" level="2">SYMPTOMS </title>
<task type = "SF">
<theme>Symptoms of Bakanae</theme> first appear about a month after planting. Infected seedlings appear to be taller, more slender, and slightly chlorotic ... The rapid elongation of infected plants is caused by the pathogen's production of the plant hormone, gibberellin.....</task>

<title type="goal" level="2">COMMENTS ON THE DISEASE </title>
Bakanae is one of the oldest known diseases of rice in Asia but has only been observed in California rice since 1999 and now occurs in all California rice-growing regions. While very damaging in Asia, the extent to which Bakanae may effect California rice production is unknown. As diseased plants .....

<title type="goal" level="2">MANAGEMENT</title>
<task type = "PROC">
The most effective means to<action> treat <theme> this disease </theme> </action> is the <instruction compound><instruction type="imperative">use of noninfested seed</instruction>. Also,<connector type="advice"> when possible</connector>, <advice>burning plant residues</advice> with known infection in fall may help limit the disease. ..... Field trials indicate that a seed treatment with sodium hypochlorite (Ultra Clorox Germicidal Bleach) is effective at reducing the incidence of this disease.... </instruction compound></task>

4.3 Matching selected texts with questions: the deep indexing level

The main words of the question focus and body are used to select a subset of indexed texts as potential candidates containing the response. Then, in each of these texts, the few sentences where the terms of the question or derived terms (closely related terms) are effectively found are annotated by means of semantic roles as for the question, for further analysis and investigations.

For that purpose, we have developed guidelines for annotating those text fragments where the response is and the associated knowledge, based on the same semantic roles as those used in the questions. These annotations remain so far exploratory, in terms of feasibility and automation. Our major concern is to develop a method for annotators so that a large number of texts can be tagged homogeneously and also so that the technique can be reproduced for other technical areas. Finally, in terms of response identification, the goal is to define a metric that defines the best match and selects the text fragment(s) that best respond(s) to the question among several potential candidates.

Let us first consider a simple example. Given the question:

Q8: *"How to eradicate Bakanae ?"*

with the following representation:

<question type="PROC or SqE" focus "eradicate Bakanae" > How to <action> eradicate <theme> Bakanae </theme> </action> ?
</question>

The main terms of the question are 'eradicate' and 'Bakanae'. The text above is therefore selected on the basis of its indexes, because 'treatment' is a closely related term (in terms of semantic relation: 'way to realize an event') of 'eradicate' in the domain ontology.

Then, the question terms are searched in the selected text and the sentences that contain them are annotated using semantic roles. For example, the following sentence is a candidate:

The most effective means to treat this disease is the use of noninfested seeds.

It is tagged as:

8

> …<action> treat <theme> this disease </theme> is the use of <instrument> noninfested seeds </instrument> </action> .

The answer is the above sentence and the text fragment that follows (introduced by the connector also) since the response is of type *procedure*:

The most effective means to treat this disease is the use of noninfested seed. Also, when possible, burning plant residues with known infection in fall may help limit the disease.

Following [5], this structure is annotated as a single instructional compound, which is the fundamental unit in a procedural text. This is the structure which is typically returned to users.

Let us present here another illustrative example of a text fragment where the response is annotated together with the required related reasoning elements:

Q9: *"How can thrips destroy the rice ?"*
     annotation:

> <question type="SqE" focus = "destroy">How can <agent> thrips </agent> <action>destroy <theme>the rice</theme> </action>?</question>

The text fragment that corresponds to the answer is annotated as follows:

> <response> <agent> The rice thrips</agent> <action> sucks the sap <source> from the young plant. </source> </action> </response>

To match the action 'destroy' in the question with the text portion from which the response is extracted, it is then necessary to identify the inference:

> <lex_inference> <action> Suck sap of X </action> <entail> <modality> probably </modality> <action> destroy X </action> </entail> , <type> X : plant </type> <part-of> sap : X </part-of > </lex inference>

This example shows that (1) in the question and in the answer, annotations are used to identify the different components, arguments, adjuncts, but also some other components (e.g. temporal adverbs), and (2) the annotation is developed to characterize the matching steps and inferential components (either lexical or domain knowledge) between the question and the answer. This latter form of annotation, which is quite time-consuming to develop, is the means we use to induce and develop domain dependent forms of lexical inference (or other phenomena like synonymy, lexical equivalence, etc.) and relevant domain knowledge. The types and lexical functions which are introduced are then used in the process of induction of generalizations over some semantic categories (plants, products, etc.), and verb classes. This way of annotating knowledge and inferences is obviously a simple bottom-up process, with well known limitations, but we feel it may have some advantages for inducing an upper organization of knowledge, in conjunction, and as a complement to, the domain ontology. It is also simple and accessible to annotators. Obviously this remains to be evaluated.

4.4 Generalizing inferences for question-answer matching

At this level, the inferences which may be drawn are directly attached to the terms which are tagged. This is obviously too limited. We are now experimenting with different generalization strategies in order to tune the lexical inference rules. This process involves:

(1) developing various generic principles over different types and categories (via the domain ontology), We will annotation the title for matching the "theme" of the answer to the "theme" and "Focus" of the question by using word net and ontology as shown below.

Surface Form	Concept
destroy, destruct, eliminate, kill,…	destroy
treat, prevent, eradicate, protect,…	manage
suck, eat, bite, drink,…	consume
spread out, diffuse, disperse,…	spread

(2) a set of principles that limit these generalizations via, for example, the taking into account of the semantics restrictions imposed by lexical items, in particular verbs. The main words of the question focus and text body that already annotated will be considered for extracting the potential candidates containing the response. The sentences, where the terms of the question or derived terms (closely related terms) are effectively found, will be the corresponding answer by using matching function as shown below.

```
Function Matching (Question Q, Answer A){
    Match = false;
    // Relevant document
    If (Q.focus = A.index) then
        // Relevant answer
        If (Q.type = A.task type) then
            //Detect Answer for the Question
            If (Q.focus = A.title) then
                Match = true;
            Else if (Q.action = A.action and
                    Q.theme = A.theme or
                    Q.agent = A.agent) then
                Match = true;
            End If
        End If
    End If
Return Match;}
```

The tuning of the level of these generalizations is obviously one main parameter of our project. It has several conceptual dimensions that we explore and may also be domain dependent.

Perspectives

The matching problem between questions and documents to retrieve answers in question-answering systems in concrete applicative contexts is often a difficult problem. This matching procedure often requires very accurate domain knowledge, besides ontological descriptions. It is not always easy to access this knowledge in a structured way or to extract it from texts. The present contribution, still experimental and in an early stage of development, is an attempt, via annotations, at resolving this problem, following a simple and clear methodology.

This task needs to be developed and evaluated gradually. So far, it is too early to evaluate the quality of the generalizations and the inferential patterns we get.

This approach, and the principles we have briefly outlined, allow us to introduce a working method for the development of question-answering systems for concrete applications, especially for non-factoid questions, an area which is still not very much developed in spite of its obvious usefulness. One of the reasons is that non-factoid questions require a language processing technology, analysis methods, reasoning aspects, and a conceptual approach, which are substantially different from what is used for factoid questions.

Acknowledgments

References

1. Asanee Kawtrakul, et al. Chaveevan Pechsiri, Sachit Rajbhandari, Frederic Andres, Problems-Solving Map Extraction with Collective Intelligence Analy-sis and Language Engineering , Book Chapter 18, Medical Information Science Reference in Information Retrieval in Biomedicine ISBN: 978-1-60566-274-9; pp 460

2. Asanee Kawtrakul, et al. 2009. From CyberBrain to Q&A Services: A Development of Question - Answering Services System for the Farmer through the SMS, WCCA2009, Grand Sierra Resort, Reno, Nevada, USA.

3. Asanee Kawtrakul, et al. 2008. "CyberBrain: Towards the Next Generation Social Intelligence" IAALD AFITA WCCA 2008, Tokyo, Japan.

4. Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, Vasile Rus. 2000. The Structure and Performance of an Open-Domain Question Answering System, Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL), Hong Kong.

5. Estelle Delpech, Patrick Saint-Dizier. 2008. Investigating the Structure of Procedural Texts for Answering How-to Questions, LREC2008, Marrakech.

6. Jochen L. Leidner, 2005. A wireless natural language search engine. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval table of contents: 677 – 677, ACM, New York, USA

7. Judy Delin, Anthony Hartley, Cecile Paris, Donia Scott, Keith Vander Linden. 1994. Expressing Procedural Relation-ships in Multilingual Instructions, Proceedings of the 7th International Workshop on Natural Language Generation: 61-70, Maine, USA.

8. Karen Sparck Jones, Branimir Boguraev.1987. A note on a study of cases, research note in Computational Linguistics archive, Volume 13 , Issue 1-2 (January-June 1987) : 65 - 68.

9. Leonard Talmy. 1976. Semantic Causative Types, In M. Shibatani (ed.), Syntax and Semantics 6: The Grammar of Causative Constructions. New York: Academic Press: 43-116.

10. Leonard Talmy. 1985. Lexicalization Patterns: Seman-tic Structure in Lexical Forms, in Language Typol-ogy and Syntactic Description 3: Grammatical Categories and the Lexicon, T. Shopen(ed.), 57-149, Cambridge University Press.

11. Mark Thomas Maybury. 2004. New Directions in Question Answering, The MIT Press, Menlo Park.

12. Mineki Takechi, Takenobu Tokunaga, Yuji Matsumoto, Hozumi Tanaka. 2003. Feature Selection in Categorizing Procedural Expressions, The 6th International Workshop on Information Retrieval with Asian Languages (IRAL2003):49-56.

13. Ray Jackendoff. 1990. Semantic Structures, MIT Press.

14. Robert E. Longacre. 1982. Discourse Typology in Relation to Language Typology, Sture Allen ´ed., Text Processing, Proceeding of Nobel Symposium 51, Stockholm, Almquist and Wiksell, 457-486.

15. Sadhu Balasundaram Ramakishnan and Balakrishnan Ramadoss. 2007. SMS for Question-Answering in the m-Learning Scena-rio, Journal of Computer Science 3(2):119-121.