An Application of Lexical Semantics Annotation to Question-Answering in e-Farming

Mukda Suktarachan (1), Patrick Saint-Dizier (2)
(1) NAIST, Kasetsart University, Bangkok, Thailand
(2) IRIT, Toulouse, France
naist_da_da@yahoo.com, stdizier@irit.fr

Abstract

In this poster we present an approach to responding to complex questions in the agriculture domain, from specifications given by experts. We present in particular a semantic annotation procedure that would allow us to define accurate and domain dedicated forms of lexical semantics inference, in order to be able to match non factoid questions (i.e. questions whose response is a significant text portion) with documents on a large scale. This project is designed to help farmers to get advices via question answering on SMS in order to improve rice farming.

1 Challenges and Goals

Question answering (Moldovan 2000, Maybury 2004) operates on top of search engines of classical textual database querying tools, by providing a layer that has natural language understanding and generation as well some reasoning capabilities in order to provide users with responses which are much more accurate and cooperative than what search engines provide in general. This is particularly crucial when responses are not straightforward, e.g. when they require some form of elaboration (synthesis of data, consistency checking, etc.), 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 from the Thai Ministry of Agriculture. 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, for example via SMS. The Thai Ministry of Agriculture has

Proceedings of the 8th International Conference on Computational Semantics, pages 338–341,
Tilburg, January 2009.

large text databases on the way rice can be planted, on how to prepare and fertilize soils and on the numerous diseases rice may be subject to, the effects, the treatments, etc. Question answering is a particularly well-adapted approach to allow farmers to query in Thai (via SMS short messages) such databases.

The NAIST lab at the university of Kasetsart has basic tools to parse Thai (stemmer, morphological analysis, part of speech recognition, and simple syntactic analysis). These tools were designed for machine translation purposes, but they turn out to be appropriate to parse queries and to retreive information in technical texts.

2 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 (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 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 SMS format (althought this is also an important semantic problem).

A few examples of question-answer pairs are (glosses from Thai, but structures are in fact quite similar to English):
Q: How to prevent the Weedy rice?
R1: Skip some seasons when growing rice,
R2: Grow hydrotonics plants.
Q: How to control the Bacterial Leaf Streak Disease?
R: Do not put too much Nitrogen.
Q: How to eradicate the rice thrips?
R: Spray with Malathion or Carbaryl every week, add fertilizer and water every two days.

Questions are essentially factoid questions (e.g. products to use, best periods for plantations, varieties to use, symptoms of a desease), why questions where responses are chains of events (reasons for something to happen) and a large number of procedural questions (Delpech et al. 2008) in particular for treating deseases. There are no comparative or evaluative questions as in other domains.

In most cases, questions do not have responses which can be immediately found in the texts. For example: *How does the Sheath Blight affects the rice growth?* has the following response: *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 are needed to allow appropriate question-text matching. 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.

3 The Question-Answering Process Annotation

Since the task is quite large (a large group of students are annotating the set of 1000 questions and related texts), we need to establish norms and annotation guidelines. Based on the research conducted at IRIT on annotating procedural questions and instructions based on semantic roles (TextCoop project), we first annotate the questions and their corresponding responses as provided by the ministry of agriculture. There are many attempts to annotate arguments by means of primitives, our approach is here oriented towards the task and the specific actions. Therefore roles are not as standard as they are in general (see IWCS 2009 Dautriche et al. this volume). An earlier attempt for language generation is e.g. (Delin 1994). Semantic tags are either close to thematic roles (instrument, location, etc.) or borrowed from the primitive systems of the Lexical Conceptual Structure (LCS), in particular to establish links between arguments, 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). Then, the text in which the response occurs is annotated (so that the boundaries and surrounding elements of the response can be characterized) and the underlying lexical semantics mechanisms at stake are given.

Let us present here an illustrative example:
Q: *How can thrips destroy the rice ?*
annotation:
<question type="manner" > How can <agent> thrips </agent> <action> destroy <theme> the rice </theme> </action> ? </question>
The response is annotated as follows:
<response> <agent> The rice thrips <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 developed to characterize the matching between the question and the answer is used to induce and develop forms of lexical inference (or other phenomena like synonymy, lexical equivalence, etc.). The types and lexical functions which are introduced contribute to the process of induction of generalizations over some semantic categories (plants, products, etc.), and verb classes.

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 exprimenting different generalization levels in order to tune lexical inference rules. This process involves (1) generalization principles over different types and categories (via a domain ontology), and (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. This task needs to be developed and evaluated gradually,; so far it is too early to evaluate the quality of the rules we get, but we believe this is a simple and close to the data approach and it should be reproducible to other areas.

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, for non-factoid questions.

References

[1] Delpech, E., Saint-Dizier, P., 2008, Investigating the Structure of Procedural Texts for Answering How-to Questions, LREC 2008, Marrakech.

[2] Delin, J., Hartley, A., Paris, C., Scott, D., Vander Linden, K., *Expressing Procedural Relationships in Multilingual Instructions*, Proc. 7th Int. Workshop on Natural Language Generation, pp. 61-70, Maine, USA, 1994.

[3] Maybury, M., *New Directions in Question Answering*, The MIT Press, Menlo Park, 2004.

[4] Moldovan, D., Harabagiu, S., Pasca, M., Milhacea, R., Goodrum, R., Grju, R., Rus, V., *The Structure and Performance of an Open-Domain Question Answering System*, Proc. 38th Meeting of the ACL, Hong Kong, 2000.

[5] Takechi, M., Tokunaga, T., Matsumoto, Y., Tanaka, H., *Feature Selection in Categorizing Procedural Expressions*, The Sixth International Workshop on Information Retrieval with Asian Languages (IRAL2003), pp.49-56, 2003.