Question Answering using FrameNet

Ashish Agrawal, Amitabha Mukherjee
Computer Science and Engineering
IIT Kanpur

February 18, 2009

Abstract

In this project we plan to improve upon existing Question Answer systems by considering a probabilistic approach to question answering. The FrameNet lexical resource would be used to tag different semantic roles in the text. Then discourse understanding would be used to construct an abstract semantic graph for every page. This would be linked to other pages to create an abstract representation of the knowledge in the corpus. A question would then be answered by comparing the question representation to a subregion of the corpus knowledge.

1 Introduction to Question Answwering systems

Question answering refers to the task of automatically answering a question posed in natural language. This requires a large collection of documents called a corpus. There is an increasing interest to integrate question answering with web search and natural language search engines are possibly the future of current search engines. The usefulness of question answering systems in information retrieval serves as a strong motivation for the project. Further, there is significant potential for improvement in the current question answering systems, few of which effectively use deep semantic information.

2 FrameNet

FrameNet [1] [2] is a lexical resource for English which is based on frame semantics and is supported by corpus evidence. Each semantic frame in the FrameNet corpus consists of semantically similar lexical units(Applyheat frame includes bake, blanch, boil, broil, brown, simmer, steam etc. as possible lexical units). Each sentential phrase is identified by a lexical unit and frame elements. For example- COOK could have FOOD and HEATINGINSTRUMENT as frame elements. Each frame element is actually a triple consisting of a frame element(eg Food), a grammatical function(eg. Object) and a phrase type(eg. Noun Phrase). Consider an example for the FrameNet annotation of a sentence-

($_{Cook}$ Matlide) fried ($_{Food}$ the catfish)($_{Heating_instrument}$ in a heavy iron skillet)

The tagging could have a single target as in the lexicographic work or multiple targets as in a full text annotation. We would be using the lexicographic work. The frame elements are divided into core and non-core elemets. The latter are only optional. FrameNet also provides for inter frame relations, which can be used for paraphrasing and better linking the question to the answer. The FrameNet tagging would help us understand the semantic information in a sentence. We expect that a deep understanding of the semantic information is helpful in developing a efficient question answering system.
The FrameNet data used for tagging is being licensed under a non- commercial license.

3 Previous work

Question Answering systems in the past have tried to match potential answers with questions by reducing the surface differences. There have been some syntax-based QA systems where a comparison between a tree representing the question and a subtree of the answer candidate is done[3]. Narayanan and Harabagiu[4] were amongst the first to stress the importance of semantic roles in answering complex questions. Kaisser [5] [6] proposes a question answering system based on FrameNet. Questions were assigned semantic roles by matching them with those in the FrameNet annotations. A reformulated question is created and a simple web search is run on it. However, the coverage is limited if the assignment is not probabilistic and relies on strict matching. Some work on using discourse structure in question answering systems has been done by Chai and Jin[7]. However, from preliminary analysis none seem to be aiming towards the creation of a knowledge set from the corpus and neither aim towards integration of discourse understanding and FrameNet tagging. The annual TREC conference has witnessed a number of question answering systems and previous participants in TREC have had an accuracy of upto 77% on the trec data set.

4 Proposal

We propose to design a question answering system using the FrameNet lexical resource. The first step is to annotate the data corpus with its semantic roles. Some of the possible approaches for this could be to use a role labeller[5] or to have a dependency relation based semantic analyser [8]. Once the semantic tagging of both the question and the data corpus has been done(possibly a bipartite graph for both), a similarity measure between two graphs can be taken over the entire data corpus as in[8]. This is sometime sufficient to answer simple questions like.
Text - Mahatma Gandhi was born in 1869.
Ques - When was Mahatma Gandhi born?

This would serve as the first level of question answering ability. In the second level, we would like to create cross references between the graph representations of different sentences. For eg.
Text - Mahatma Gandhi was born in Porbandar.
Text - Porbandar is in the country of India.

The NP Mahatma Gandhi would be cross references between the two sentences and hence this would enable us to answer questions of the form -
Question - Which country was Mahatma Gandhi born in?

This would be enabled by creating linkages between the nodes of two bipartite graphs(if the semantic analyser approach mentioned in [8] is considered). This would serve as a second level of question answering ability. Finally we would like to answer more open questions which possibly require more than a single phrase answer and also take into consideration linkages between different documents. This would correspond to creating an ontology for the question answering system. The approach of having cross-linkages in bipartite graphs might not scale up for large corpusses and other alternatives need to be evaluated. One possible method could be to keep a database of key words and have linkages in the database to the occurences of the keyword. For eg. All the occurences of the keyword Abhrahim Lincoln in the corpus could be pointed to by the databse entry in the name of Abrahim Lincoln and all these would constitute our world knowledge about Abrahim Lincoln.
With this we expect to answer questions desiring a summary on Abrahim Lincoln.

The algorithms and techniques developed in this project would be tested on the TREC - QA track data set[9].

References

[1] Josef Ruppenhoffer. *FrameNet II - Extended Theory and Practice*. 2006.

[2] Collin Baker, Charles F. Fillmor, and John B. J. Lowe. The berkeley framenet project.

[3] Dragomir Radev, Weigua Fan, Hong Qi, Harris Wu, and Amardeep Grewal. Probabilistic question answering on the web. 2005.

[4] Srini Narayanan and Sanda Harabagiu. Answering questions using advanced semantics and probabilistic inference. 2004.

[5] Michael Kaisser and Bonnie Webber. Question answering based on semantic roles. 2007.

[6] Michael Kaissner. Qualim at trec 2005:web-question answering with framenet. 2005.

[7] Joyce Y. Chai and Rong Chin. Discourse structure for context question answering. 2004.

[8] Dan Shen and Mirella Lapata. Using semantic roles to improve question answering. 2007.

[9] Text Retrieval and Extraction Conference: Question Answering track dataset, 2007.