Question Type Classification using Head Words and Hypernyms

Justin Domingue – 260588454
justin.domingue@mail.mcgill.ca

Abstract

This document contains the instructions for preparing a camera-ready manuscript for the proceedings of ACL-2014. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

Users have more and more expectations for web engines, requiring them not only to understand the nuances of languages better but also to give more meaningful answers to queries. By amassing a greater amount of knowledge in the Web, these engines become increasingly smarter. However, without some deeper analysis of that internal knowledge, web engines seem reach a plateau in terms of themeaningfulness of the information-retrieved. The next step in information retrieval is to extractconcise and succinct answers to user queries rather than giving back entire documents (Sundblad, 2007).

A Question Answering (QA) system can be built to respond to natural language questions byretrieving the relevant information from stored documents. However, without some sort of question classification, the range of answers is quite large.Question type classification plays a crucial role is bounding the type of the answers given by the QA. For instance, the answer to the question ?Where is Canada? should be a Location:city while that of ?Who is the president of the United States? should be a Human:ind. It is thenunavailing to answer a question with the wrong type ofanswer.

Question type classification is then defined as "the task of determining the correcttype of the answer expected to a given query." (Khoury, 2011) There are few ways of accomplishing this task. One could use rule-based classification; or a machine learning approach – learning from a labeled data set. It seems obvious that simple rules cannot account for all the nuances of English syntax. Take for example a rule-based classification system which would attribute questions with a particular wh-word to a given category. Consider the sentences *What is the capital of Canada?* and *What is an espresso?*. The first sentence is of location (city) type while the second is of definition type. Clearly, a more sophisticated set of rules would need to be devised. On the other hand, a statistical approach has the advantage that "one can focus on designing insightful features, and rely on learning process to efficiently and effectively cope with the features." (Huang et al., 2008).

Work on machine learning approaches to question classification was initiated by Li and Roth in 2002 (Li and Roth, 2006). Other work in this area include (Pinto et al., 2002) who employed entity tagging, part of speech tagging, regular expressions and language models, (Sundblad, 2007) who carried out a performance analysis of various machine learning algorithms applied to the question classification task, (**?**) who introduced a Part-of-Speech Hierarchy coupled with informer spans.

More related to this paper, Huang et al. (2008) have built upon the work of Li and Roth and defined new features, namely head words and their hypernyms. They have found a significant increase of the model accuracy. In this paper, we follow the path of Huang et al. to confirm their results.

2 Classifiers

3 Features

4 Experimental Results

5 Conclusion

References

Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 927–936, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard Khoury. 2011. Question type classification using a part-of-speech hierarchy. In Mohamed Kamel, Fakhri Karray, Wail Gueaieb, and Alaa Khamis, editors, *Autonomous and Intelligent Systems*, volume 6752 of *Lecture Notes in Computer Science*, pages 212–221. Springer Berlin Heidelberg.

Xin Li and Dan Roth. 2006. Learning question classifiers: The role of semantic information. *Nat. Lang. Eng.*, 12(3):229–249, September.

David Pinto, Michael Branstein, Ryan Coleman, W. Bruce Croft, Matthew King, Wei Li, and Xing Wei. 2002. Quasm: A system for question answering using semi-structured data. In *In Proceedings of the Joint Conference on Digital Libraries (JCDL) 2002*, pages 46–55.

Håkan Sundblad. 2007. *Question Classification in Question Answering Systems*. Ph.D. thesis, Linköping Studies in Science and Technology.