

University of Macau

Abstract

iQA: Intelligent Question Answering from the Web

by Chan Mei Pou

Thesis Supervisor: Dr. Gong Zhiguo

Department of Computer and Information Science

Question answering (QA) is the study on the methodology that returns exact answers to natural language questions, rather than a list of potentially relevant documents.

The challenge with QA system is how to return answers to user's natural language questions. The whole process is quite complicated as it involves quite a number of different techniques to work closely together in order to achieve the goal, including query rewrites and formulations, question classification, information retrieval, passage retrieval, answer extraction, answer ranking and justification. The end-to-end performance of a complete QA system hence depends on each of these independent factors. Over the past few years, individual research groups have been continuing to refine each of these steps with the intention to increase the coverage and accuracy of QA systems.

Several such question answering systems have been developed and made available for use in the Web. However, these systems have not taken questions with terms written in abbreviations into consideration. For example, they can present the relevant information among the returned answers in response to the question "*Where is University of Macau?*". However, when the

abbreviation for “University of Macau” is used instead, that is, when “*Where is Umac?*” is submitted, though the returned result set contains information related to “umac”, it has nothing to do with “University of Macau” because the abbreviation “umac” can stand for many different things besides “University of Macau”. In this case, it so happens that the semantics of the returned answers does not match with the expectation of the users.

This work addresses this problem by attempting to reduce the semantics gap between questions with terms written in abbreviations and their potential answers. To achieve this objective, the work includes (1) identifying terms that might be abbreviations from the user’s natural language question; (2) retrieving documents relevant to that abbreviation term; (3) filter noun phrases that are considered to be potential long forms for that abbreviation within the returned result. A complete QA system will be developed in order to evaluate the performance of this technique.

The work is mainly divided into three parts:

Part I gives a brief introduction on the development of QA systems with a summary on the work of TREC.

Part II details the system architecture and implementation of the proposed QA system, including the discussion on the solution to narrow the semantics gap between questions with terms written in abbreviations and the potential answers.

Part III presents the evaluation and conclusions of this work.