Linguistic and Semantic Passage Retrieval Strategies for Question Answering

Matthew W. Bilotti
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
mbilotti@cs.cmu.edu

December 4, 2009

Abstract

Question Answering (QA) is the task of searching a large text collection for specific answers to questions posed in natural language. Though they often have access to rich linguistic and semantic analyses of their input questions, QA systems often rely on off-the-shelf bag-of-words Information Retrieval (IR) solutions to retrieve passages matching a set of terms extracted from the question.

There is a fundamental disconnect between the capabilities of the bag-of-words retrieval model and the retrieval needs of the QA system. Bag-of-words IR retrieves documents matching a query, but the QA system really needs documents that contain answers. Through question analysis, the QA system has compiled a sophisticated *information need* representation for what constitutes an answer to the question. This representation is composed of a set of linguistic and semantic constraints satisfied by answer-bearing passages. Unfortunately, off-the-shelf IR libraries commonly used in QA systems can not, in general, check these types of constraints at query-time. Poor quality retrieval can cause a QA system to fail if no answer-bearing text is retrieved, if it is not ranked highly enough, or if it is outranked or overwhelmed by false positives, text that matches the query well, yet supports a wrong answer.

This thesis proposes two linguistic and semantic passage retrieval methods for QA, one based on structured retrieval and the other on rank-learning techniques. In addition, a methodology is proposed for mapping annotated text consisting of labeled spans and typed relations between them into an *annotation graph* representation. The annotation graph supports query-time linguistic and semantic constraint-checking, and serves as a unifying formalism for the QA system's information need and for retrieved passages. The proposed methods rely only on the relatively weak assumption that the QA system's information need can be represented as an annotation graph. The two approaches are shown to retrieve more answer-bearing text, more highly ranked, compared to a bag-of-words baseline for two different QA tasks. Linguistic and semantic passage retrieval methods are also shown to improve end-to-end QA system accuracy and answer MRR.

Available online at: `http://www.cs.cmu.edu/~mbilotti/thesis.pdf`