

Experiments with Open-Domain Textual Question Answering

Sanda M. Harabagiu and Marius A. Pasca and Steven J. Maiorano

Department of Computer Science and Engineering

Southern Methodist University

Dallas, TX 75275-0122

{sanda,marius,steve}@renoir.seas.smu.edu

Abstract

This paper describes the integration of several knowledge-based natural language processing techniques into a Question Answering system, capable of mining textual answers from large collections of texts. Surprising quality is achieved when several lightweight knowledge-based NLP techniques complement mostly shallow, surface-based approaches.

1 Background

The last decade has witnessed great advances and interest in the area of Information Extraction (IE) from real-world texts. Systems that participated in the TIPSTER MUC competitions have been quite successful at extracting information from newswire messages and filling templates with information pertaining to events or situations of interest. Typically, the templates model queries regarding *who* did *what* to *whom*, *when* and *where*, and eventually *why*.

Recently, a new trend in information processing from texts has emerged. Textual Question Answering (Q/A) aims at identifying the answer of a question in large collections of on-line documents. Instead of extracting all events of interest and their related entities, a Q/A system highlights only a short piece of text, accounting for the answer. Moreover, questions are expressed in natural language, are not constrained to a specific domain and are not limited to the six question types sought by IE systems (i.e. *who*₁ did *what*₂ to *whom*₃, *when*₄ and *where*₅, and eventually *why*₆).

In open-domain Q/A systems, the finite-state technology and domain knowledge that made IE systems successful are replaced by a combination of (1) knowledge-based question processing, (2) new forms of text indexing and (3) lightweight abduction of queries. More generally, these systems combine creatively components of the NLP basic research infrastructure developed in the 80s (e.g. the computational theory of Q/A reported in (Lehnert 1978) and the theory of abductive interpretation of texts reported in (Hobbs et al.1993)) with other shallow techniques that make possible the open-domain processing on real-world texts.

The idea of building open-domain Q/A systems that perform on real-world document collections was initiated by the eighth Text REtrieval Conference (TREC-8), by organizing the first competition of answering fact-based questions such as “*Who came up with the name, El Nino?*”. Resisting the temptation of merely porting and integrating existing IE and IR technologies into Q/A systems, the developers of the TREC Q/A systems have not only shaped new processing methods, but also inspired new research in the challenging integration of surface-text-based methods with knowledge-based text inference. In particular, two clear knowledge processing needs are presented: (1) capturing the semantics of open-domain questions and (2) justifying the correctness of answers.

In this paper, we present our experiments with integrating knowledge-based NLP with shallow processing techniques for these two aspects of Q/A. Our research was motivated by the need to enhance the precision of an implemented Q/A system and by the requirement to prepare it for scaling to more complex questions than those presented in the TREC competition. In the remaining of the paper, we describe a Q/A architecture that allows the integration of knowledge-based NLP processing with shallow processing and we detail their interactions. Section 2 presents the functionality of several knowledge processing modules and describes the NLP techniques for question and answer processing. Section 3 explains the semantic and logical interactions of processing questions and answers whereas Section 4 highlights the inference aspects that implement the justification option of a Q/A system. Section 5 presents the results and the evaluations whereas Section 6 concludes the paper.

2 The NLP Techniques

Surprising quality for open-domain textual Q/A can be achieved when several lightweight knowledge-based NLP techniques complement mostly shallow, surface-based approaches. The processing imposed by Q/A systems must be distinguished, on the one hand, from IR techniques, that locate sets of doc-