University of Sheffield TREC-9 Q & A System

Sam Scott* and Robert Gaizauskas
{s.scott,r.gaizauskas}@dcs.shef.ac.uk

Department of Computer Science
University of Sheffield
Regent Court, Portobello Road
Sheffield S1 4DP UK

1 Introduction

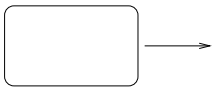
The system entered by the University of Sheffield in the question answering track of TREC-9 represents a significant development over the Sheffield system entered for TREC-8 [6] and, satisfyingly, achieved significantly better results on a significantly harder test set. Nevertheless, the underlying architecture and many of the lower level components remained the same. The essence of the approach is to pass the question to an information retrieval (IR) system which uses it as a query to do passage retrieval against the test collection. The top ranked passages output from the IR system are then passed to a modified information extraction (IE) system. Syntactic and semantic analysis of these passages, along with the question, is carried out to identify the "sought entity" from the question and to score potential matches for this sought entity in each of the retrieved passages. The five highest scoring matches become the system's response.

2 System Description

2.1 Overview

The key features of the system setup, as it processes a single question, are shown in Figure 1. Firstly, the (indexed) TREC document collection and the question are passed to an IR system which treats the question as a query and returns top ranked passages from the collection. As the IR system we used the Okapi system [8][1]. Following this, the top ranked passages are run through a text filter to remove certain text formatting features which cause problems for downstream components. Finally, the question itself and the filtered top ranked passages are processed by a modified version of the LaSIE information extraction system [5], which we refer to below as QA-LaSIE. This yields a set of top ranked answers which are the system's overall output.

The reasoning behind this choice of architecture is straightforward. The IE system can perform detailed linguistic analysis, but is quite slow and could not process the entire TREC collection for each query, or even realistically pre-process it in advance to allow for reasonable question

*Current address: Institute for Interdisciplinary Studies, Carleton University, Ottawa, K1S 5B6, Canada. sscott@ccs.carleton.ca.

[1]Software available at: http://dotty.is.city.ac.uk/okapi-pack/. For TREC-8 we both used the NIST-supplied top documents and passages from UMass's INQUERY system [2] which UMass kindly provided for us. Our switch from INQUERY to Okapi was prompted by the acquisition of Okapi and Okapi expertise in-house.