Interfacing with Big Data Repositories

Boris Katz

The Infolab Group's research of bigdata@csail was focused on creating new ways for accessing and interacting with massive scale data. We believe that natural language interaction provides a powerful mechanism for pairing specific user requests with specific data analysis, data access, and data visualization techniques: language makes it possible to "compress" data by focusing on what people find important.

We concentrated on interfacing the START Question Answering system (http://start.mit.edu) with large datasets which contain huge amounts of valuable data regarding a large number of different classes of objects. Traditionally, the START system models heterogeneous semi-structured resources on the Web using our Object–Property–Value (OPV) data model which makes it possible to view and use large segments of the Web as a database. This model regards any such resource as a collection of objects, each possessing a number of properties with values. Some examples of object–property combinations are a country's population, or a famous individual’s place of birth. When a user asks START a natural language question, this question is first matched against a knowledge base of manually written "annotations" which tell the system what knowledge sources are currently available to the system. A successful match makes it possible for the system to recognize the question as a request for finding a value for a particular property of a particular object, identify a Web resource containing the answer, and then run a script that extracts the required value from that data source. A particular strength of sophisticated language processing for these types of questions is the ability to recognize numerous ways in which language can be used to ask an OPV query.

The approach described above makes it possible to create a powerful and high-precision question-answering system; however, it suffers from a serious drawback: it is labor-intensive and time-consuming, which makes the incorporation of very large information resources of certain types infeasible.

In this effort, using Wikipedia as our example dataset, we combined the START system's language processing and syntactic and morphological analysis to dramatically increase START's coverage: we implemented the capability to answer OPV questions from Wikipedia infoboxes without the use of manually written annotations or scripts. The techniques we developed not only allow the system to answer vast numbers of new and interesting questions from Wikipedia infoboxes, but they are also directly applicable to all large datasets that contain labeled properties.

We see great opportunities for further research in this area. In our current system, START determines how the language of the question can be mapped to a noun that expresses a property, such as occupation or director. However, Wikipedia infoboxes contain many properties that aren't expressed as nouns. They may be expressed as verbs (released), verb–preposition pairs (succeeded_by, known_for), and in many other ways. In the next stage of our research, we plan to
use START's parser to analyze these more complex labels for properties of objects and create new methods for mapping various verbalizations of OPV queries in natural language to property names in Wikipedia infoboxes and other Big Data repositories.