Integration of Structured Data with Natural Language: Three Test Collections

Douglas W. Oard, oard@umd.edu
CLIS/UMIACS, University of Maryland, College Park, MD 20740

Interesting and important questions are often to be found near disciplinary boundaries. Reasoning across combinations well structured data and "natural language" information sources seems to fit that mold well. Much has been published on extraction of structured data from documents that were written originally for human readers, and much of that work is motivated by the potential for integration with data obtained from other sources. But actual large-scale integration has been rare, in part because we have yet to identify canonical challenge problems for which standard evaluation resources can be constructed. This paper describes three cases in which evaluation resources already exist that could be used to investigate a range of information integration issues. Each includes a substantial amount of structured data, a large collection of natural language text, and one or more integration tasks.

The most straightforward of these is identity resolution in the Enron email collection. The raw collection is available from Aspen systems for media and processing charges (about $15,500), but MIT, SRI and CMU have made a relatively clean version of the collection freely available.[1] That "CMU" version of the collection contains about a half million messages that were found in the Microsoft Outlook mail folders of 150 Enron employees. Remarkably, the collection contains more than 133,000 unique email addresses. There are a few cases (e.g., mailing lists or shared staff accounts) in which the same address was used by more than one person, but the vast majority of the addresses identify a single individual. Some people do, of course, use more than one address, but many such cases are easily identified (e.g., when they cc themselves at another address). We can, therefore, reasonably model identity resolution as association of named, nominal and/or pronominal references with an email address for the appropriate individual. All that would be needed to build a test collection for evaluation of accuracy on that task is an answer key for an appropriately stratified sample of mentions. Of course, many mentioned people will not have known addresses. Recognizing that condition is, therefore, a natural part of the task.

The Complex Document Information Processing (CDIP) project at Illinois Institute of Technology has produced a freely available collection that contains about 7 million scanned documents from the Tobacco Master Settlement Agreement.[2] Each document has an associated metadata record that describes its content and provenance, and about half the records contain optical character recognition results. As with the Enron collection, the author and recipients of each document are often named in the metadata, and both mentioned persons and mentioned organizations are also sometimes identified in the CDIP metadata. About 800,000 of the metadata records also contain topic labels drawn from a controlled vocabulary. Thus the CDIP collection can be partitioned in ways that could quite naturally support both identity resolution and some types of semantic information integration.

---

[1] http://www.cs.cmu.edu/~enron
[2] http://trec-legal.umiacs.umd.edu

The Cross-Language Evaluation Forum (CLEF) Cross-Language Speech Retrieval (CL-SR) collection, which can be licensed for research use from the Evaluation and Language Resources Distribution Agency (ELDA), can support an even broader range of information integration tasks.[3] The CLEF CL-SR collection contains about 1,000 hours of English and Czech interviews for which speech recognition transcripts and metadata fields containing both names of speakers and of people mentioned in the interviews are available. As with the CDIP collection, controlled vocabulary topic labels are also available as metadata, and the fact that the same labels are used for both English and Czech means that the collection can support some types of bilingual information integration experiments. The CLEF CL-SR collection also includes metadata specifying the locations and time periods being described by the speaker at every point in the collection, thus potentially supporting some types of geographic and temporal information integration experiments. Temporal information integration may be particularly interesting in this context because references to time in informal speech are often both relative and indirect (e.g., "a few months after the hurricane …"), thus requiring extended reasoning chains.

These collections illustrate three trends that are increasingly evident: (1) it is becoming increasingly common to find complex metadata together with natural language, (2) automated transducers such as optical character recognition and automatic speech recognition are now sufficiently accurate and efficient to make a broad range of language use computationally accessible, and (3) the resulting complex collections are seriously underutilized. This last point is worthy of strong emphasis. Despite the presence of rich metadata in the CDIP and CLEF CL-SR collections, researchers have to date focused mostly on the natural language content of those collections. That's entirely natural, of course, since that is the purpose for which those collections were created. But as the collections we work with become more complex, there will surely be many new opportunities for interdisciplinary collaboration to explore information integration issues that demand an increasingly broad range of expertise. This workshop seems like a good place to begin that discussion!

[3] http://clef-clsr.umiacs.umd.edu