A Breadth of NLP Applications

Elizabeth D. Liddy, Professor
Director, Center for Natural Language Processing
School of Information Studies
Syracuse University
Syracuse, New York, USA
www.cnlp.org

Introduction

The Center for Natural Language Processing (CNLP) was founded in September 1999 in the School of Information Studies, the "Original Information School", at Syracuse University. CNLP's mission is to advance the development of *human-like, language-understanding software capabilities* for government, commercial, and consumer applications. The Center conducts both basic and applied research, building on its recognized capabilities in Natural Language Processing. The Center's seventeen employees are a mix of doctoral students in information science or computer engineering, software engineers, linguistic analysts, and research engineers.

Opportunities

We are finding that today is a particularly opportune moment for NLP due to the confluence of a number of factors:

1. Sufficient R & D in the field of NLP has been accomplished in past years to provide solid baseline NLP capabilities,
2. Computational resources have caught up to the requirements of complex NLP systems, and
3. The bulk of textual information that forms the basis on which all organizations conduct their business is now in electronic format.

These factors are key to both research centers and commercial vendors. They contribute to the open, welcoming reception that NLP applications are now receiving with both funders and customers. NLP is proving itself as a powerful enabling technology for a range of applications supported by the Center's technology, including:

- Document Retrieval
- Question-Answering
- Information Extraction
- Text Mining
- Automatic Metadata Generation
- Cross-Language Retrieval
- Document Summarization

bebee, Inc.

To meet the many opportunities before us, CNLP recently licensed in the <!metaMarker> technology of be-bee, Inc. (www.be-bee.com) that will enable us to provide a wide range of implementations by building on this solid commercial NLP technology. With the increasing request for sophisticated language-handling applications from both government and commercial funders, we needed to focus on just those aspects of the technology that are research-related. The licensing in of be-bee's <!metaMarker> advanced commercial capabilities has sped our time to delivery by bootstrapping our language processing modules with accurate interpretations of input text and providing scalable and reliable language processing.

While there are other commercial technologies available, CNLP chose be-bee Inc.'s <!metaMarker> because of the quality and depth of its language processing - they are a real NLP shop. This is what they do, and they do it well. Also, compared to the other technologies we looked at, be-bee's technology is much more flexible and enables us to specialize the output in ways that matter to us for various applications, three of which we will describe.

Automatic Metadata Generation

Together with be-bee Inc., we are working on an automatic metadata generation project under funding from the National Science Foundation's Science, Math, and Engineering Digital Library Program. The goal of our project is to run our combined natural language processing on learning resources (lesson plans, classroom activities, etc) in order to assign values to the twenty-three metadata attributes that have been accepted as the metadata standard for education. It is these metadata values, which are then matched in response to users' queries, or browsed by users to familiarize themselves with a digital library's resources. The goal here is to break the metadata generation bottleneck which human assignment of metadata has caused. We will accomplish this by improving the speed with which educational resources can be made available; increasing the number of educational resources which are available electronically, and; providing improved access to a digital library for users via richer and more complete metadata values.

In a second NSF Digital Library Project, we will be extending the metadata information we extract to include the educational standard that the resource can be used to accomplish. By working with a master list of educational content standards, which the state standards map to, this new metadata attribute will enable teachers and administrators in any state to select those teaching materials which will assist their students in meeting the required standards of their state.

Automatic metadata generation is an area of numerous opportunities because metadata standards are being developed and agreed to in many domains, including business, geography, education, and biology, and for use by various technologies, including statistical table browsers, digital libraries, and peer-to-peer technologies. If there is text associated with an object, NLP can provide the means to interpret the text to understand

whether the information for each metadata element is present in the object and then extract from the object the values to fill the appropriate standard's metadata record. This uniform description will then make the object findable and accessible by users.

Question-Answering

Question-Answering (QA), is a very hot topic right now and has a breadth of appeal because it is an application with many flavors based on the context in which the QA system is used. Building on our existing eQuery system, we are currently developing a QA capability for NASA for use within a collaborative learning environment for distance education in aeronautical and mechanical engineering. The students will use the collaboration technology for class interactions and for group project work. At any time, a student can pose a question, phrasing it as naturally as if they were asking the professor or another student. In turn, the QA system interprets the query at all the levels of language at which humans extract meaning.

For example, for the query:

"What is the best material for the wings of reusable spacecraft?"

the Language-to-Logic module of our eQuery system would produce the following representation:

material* AND wing* AND reusable_spacecraft OR reusable_launch_vehicle OR RLV

in which the correct logical relations are understood, conceptual phrases are recognized, and single words are stemmed to match morphological variants, as well as being expanded to include their synonymous phrasings. The query representation is searched against the NLP-indexed technical papers, class lectures, questions and answers accumulated over time (Previously Asked Questions / PAQs), and transcripts of problem-solving interactions of prior classes. eQuery then presents the student with one or more answer-providing passages, which the system has ranked according to their likelihood of containing the answer to their question.

Information Extraction

Government and competitive intelligence is another application area where we have utilized the rich interpretation provided by our NLP for Information Extraction (IE). We have found that both sets of customers need to extract a broad range of entities from text. Our IE includes the more typical person, place, and location, but more specialized extractions (e.g. perpetrator, drug, weapon, backer) as well - the specialization depending on the domain of interest, for a total of 165 entities. Our technology also recognizes and extracts the relations between or amongst entities, and most recently we have been focusing on using a frame-based representation to recognize and extract the multiple aspects of events. In our current Evidence Extraction and Link Discovery Project for DARPA, we are focusing on scenario extraction, which requires learning the set of events

which predicted an incident of interest in the past, and then recognizing when this scenario appears to be playing itself out again.

A project that we recently completed for Unilever, a large international conglomerate, focused on adapting our IE technology for their use in competitive intelligence by enabling them to track their competitors' activities from web and subscription database sources. The IE capability processes through large-volume daily news-feeds, recognizing, interpreting, and extracting entities, relations, and events of interest and feeding them into a visualizer to make it possible for their strategic-intelligence staff to recognize patterns that may not be apparent when individual documents are processed.

Future Work

While the existing set of applications for NLP is broad, varied, and interesting, our belief is that there will be an even larger number of viable applications in the near future. As speech-understanding technology improves and voice-input applications multiply, the need for full NLP capabilities will growth exponentially. When we think of it, a very high proportion of what humans accomplish is either accomplished through language or is reported in language – and therein lies the future of NLP applications.