SESSION 5: NATURAL LANGUAGE, DISCOURSE

Paul S. Jacobs, Chair

Information Technology Laboratory
GE Research and Development Center
Schenectady, NY 12301

OVERVIEW AND BACKGROUND

The papers in this group cover a broad range of topics from different perspectives. They have in common an emphasis on the handling of frequently-occurring phenomena in real data sets of spoken and written language, phenomena which are in some sense outside of the scope of some of the core problems in human language technologies. We can view problems such as acoustic processing, word recognition, sentence parsing, and word sense disambiguation as core problems because they have a wealth of published literature and a set of broadly applied techniques. By contrast, the natural language papers in this session hit upon issues like recognizing speech repairs and designing "templates" to capture information and test text understanding. These issues are also central to HLT work, but have certainly not evolved into mature practices.

Within this general framework, three of the papers address written language work, including template design and processing methods, while the remaining two papers are on spoken language work, including repairs and intonation.

The two speech papers represent innovative work done independently in universities, while the three written language papers are part of a community effort to develop and evaluate systems for extracting data from free text, and are all thus all directly related to the ARPA-funded TIPSTER program. The two speech papers, therefore, are quite self-standing. However, to appreciate the three written language papers, it helps to have a general understanding of TIPSTER data extraction and why, for example, some of the themes of these papers, such as template design and linguistic motivation, are important. Therefore, we will first consider the three written language papers here, then turn to the two speech papers.

TIPSTER DATA EXTRACTION

The TIPSTER (Phase 1) data extraction project, spanning roughly a two-year period from 1991-1993, aimed to expand the state of the art in coverage and accuracy in data extraction techniques—systems that derive structured information from free text, to support "downstream" text applications such as database retrieval, trend analysis, question answering, and so on. TIPSTER Phase 2, which focuses on a common, sharable architecture for such systems along with continuing algorithm development, is just beginning.

In TIPSTER Phase 1, data extraction systems showed the ability to cover broad ranges of text in two domains (joint ventures and microelectronics) and two languages (English and Japanese), with accuracy comparable to how similar programs had done a year earlier on much easier tasks in a single language and domain. The program thus made considerable progress in scale-up and portability, which were its key goals. In addition, the TIPSTER-sponsored evaluations, including the recent MUC-5 message understanding conference, provided a testbed in which many other sites were able to participate and compare approaches.

The papers by Boyan Onyshkevych, "Issues and Methodology for Template Design for Information Extraction", and by Jerry Hobbs and David Israel, "Principles of Template Design", discuss the infrastructure within which work in data extraction is conducted. The data extraction task generally uses a corpus of texts, a template design reflecting both the domain of knowledge and the information to be extracted, and some examples of correctly-filled templates for training. Two particularly important criteria of template design help to make such work successful: (1) the design must be expressive and general enough so that a system that can do a good job at filling a particular template can be used successfully in new, real applications, and (2) the design must be both rich and intuitive, so that the task reveals interesting characteristics of different methods without making a lot of extra work for researchers. The infrastructure issues presented in these papers, therefore, are essential for progress in the field.

The paper by Damaris Ayuso and the BBN PLUM Research Group, "Pattern-Matching in a Linguistically Motivated Text-Understanding System" summarizes some

of BBN's efforts in TIPSTER, but also gives a good perspective on some of the results of ARPA-sponsored research in data extraction. As the paper points out, some of the most successful efforts in data extraction have gradually dispensed with traditional linguistic knowledge (such as the largest and most powerful grammars) and relied more heavily on pattern matching and lexically-organized knowledge. This trend has been going on for several years and has been reported, particularly by GE and SRI, in previous ARPA workshops. Lexically-oriented pattern matching is particularly good at quickly capturing the domain and corpus knowledge that is required for data extraction. The BBN paper suggests that pattern matching and linguistic knowledge work best together, a claim that would seem to be supported by the fact that systems like SRI's and GE's both use pattern matching as an approximation to more powerful analysis. BBN's system, by contrast, still includes a more traditional grammatical component.

SPEECH REPAIRS AND INTONATION

"Tagging Speech Repairs" by Peter Heeman and James Allen, addresses a critical issue in processing real spoken dialogues—that recognizing and correcting repairs, which are frequent in real speech, is necessary to process and understand many spoken inputs. The novelty of this work is that it relies heavily on a part-of-speech tagger, combining a variety of cues to spot and correct the repairs. The results reported, for both recognition and correction, are quite good. However, the discussion of the paper at the HLT meeting did raise some of the problems with comparing results of different systems and approaches on different data sets; for example, related work by Nakatani and Hirschberg used acoustic information only and tested only on examples that included repairs, and did only recognition, not correction. This raises the question of how these different approaches could be effectively compared and even combined.

"Information Based Intonation Synthesis" by Scott Prevost and Mark Steedman uses a rich linguistic model to account for problems with contrastive stress in dialogues—cases where the simple traditional rule, that previously-mentioned word are de-accented, breaks down. Although the results of this work so far are more limited and anecdotal than the other papers in this session, the approach shows promise for covering a broader range of examples of stress and intonation.