Report from the Text Analysis Techniques Topic Session

Robert E. Stumberger
Language Systems, Inc.
Internet: res@lsi.com

INTRODUCTION

Presentations in the Text Analysis Techniques session covered areas related to:

- disambiguation of temporal expressions (Lois Childs, GE/Martin Marietta),
- trainable correction of Japanese word segmentation (Sean Boisen, BBN),
- principle-based parsing (Robert Belvin, Language Systems Inc.),
- and disambiguating with semantic tags (Jim Cowie, NMSU).

DISAMBIGUATING TEMPORAL EXPRESSIONS

Lois Childs discussed GE's efforts to extract temporal expressions from text through the identification of relevant patterns. The Shogun system used 37 patterns in English, and 7 for Japanese. Patterns were context dependent, and referenced a dateline in order to handle relative time. The patterns were able to perform temporal calculations, and the system computed a temporal structure from reference points on the dateline. The system was able to handle temporal references which were spread throughout a message. This approach allowed the Shogun system to have a good coverage of time fills; extensions to this approach will provide improved handling of ambiguous dates.

AUTOMATICALLY TRAINABLE ERROR CORRECTION OF JAPANESE SEGMENTATION AND PART-OF-SPEECH ANALYSIS

Sean Boisen presented BBN's work on a learning algorithm which was used to improve the performance of the Juman Japanese word-segmentation system provided by Kyoto University. BBN used AMED (Automatic Morphological Error Detection) as a segment correction model between Juman and BBN's POST tagger. Using hand-produced segmentation and tagging for training purposes, the system was able to acquire transformations from tags, and learn rules for segment correction, in order to reclassify words, put words together, and take words apart. The system produced a chart of possible corrections. The supervised training used Treebank software. The AMED/POST combination was able to improve segmentation and tagging performance with little data; this approach will be extended to parsing in the future.

SOME ASPECTS OF PRINCIPLE-BASED PARSING IN THE MUC-5 TASK

Robert Belvin described the use of principle-based parsing in LSI's parser, in particular the use of principles of grammatical theory and parsing principles (based on empirical knowledge of language). LSI's parser incorporates a number of features of the Government-Binding theory of syntax, including projection and thematic principles, in an essentially head-driven parser which employs bottom-up and expectation-based characteristics. The parser is designed to be language independent, to produce syntactic structures which facilitate semantic processing, and to be robust enough to produce partial parses which are usable in later semantic processing. Robert discussed the handling of empty categories, with respect to passive constructions and embedded infinitivals. Robert concluded with a discussion of the insertion of special structures as a means of providing a "quick fix" for constructions which are not completely handled by the principles which have been implemented.

DEALING WITH AMBIGUITY

Jim Cowie discussed experiences with NMSU's reference resolution module in their Diderot system. The system attempted to disambiguate text into a list of sense tokens. Disambiguation was performed in parsing and semantic tagging stages. Tagging was done using word-lists with semantic and type tags. Parsing used tags in conjunction with co-specification patterns. Jim discussed various problems which occurred, including a lack of sense-tokens for Japanese, multiple-tagging problems, the need for a lexical database featuring compound terms, the need for domain-specific markers, the need for combinatorial rules, and the need for negative blocking information. Future experiments will focus on the use of machine learning techniques for acquiring semantic tagging information and deriving semantic patterns.