EAGLE:
An Extensible Architecture for General Linguistic Engineering

Breck Baldwin, Christine Doran, Jeffrey C. Reynar, Michael Niv, B. Srinivas

University of Pennsylvania
Department of Computer and Information Science and
The Institute for Research in Cognitive Science

email: {breck,cdoran,jcreynar,niv,srini}@linc.cis.upenn.edu

Over the course of two summer projects, we developed a general purpose natural language system which advances the state-of-the-art in several areas. The system contains demonstrated advancements in part-of-speech tagging, end-of-sentence detection, and coreference resolution. In addition, we believe that we have strong maximal noun phrase detection, and subject-verb-object recognition and a pattern matching language well suited to a range of tasks. Other features of the system include modularity and interchangeability of components, rapid component integration and a debugging environment.

The demo will feature aspects of the system currently being used to develop a coreference resolution engine in preparation for MUC-7, in addition to an information extraction task done over the summer of 1996. Two aspects of the system will be featured prominently, a diagnostic tool for evaluating system output using SRA's discourse tagging tool (DTT) and the MOP pattern matching language.

The diagnostic tool takes a coreference annotated text to be evaluated, an answer key assumed to be correct, and produces various diagnostics which evaluate system performance. Areas of evaluation include:

- Classification of coreference links into correct, sins of commission(precision errors), sins of omission (recall errors)
- Noun phrase detection errors
- Filters on what sorts of links to evaluate
- Support of system trace functions in the DTT
- Fast implementation of MUC-6 scoring algorithm

In addition, we present MOP (Mother of Perl), a pattern description language developed for use in an information extraction task and currently being used to do coreference. Patterns are described in MOP by left-to-right enumeration of components, with each component specifing at various levels of descriptive granularity. The patterns are compiled into Perl scripts, which perform back-tracking search on the input text. MOP also allows for rapid integration of a variety of analytical modules, such as part-of-speech taggers and parsers.