Topic Session on DISCOURSE

Damaris M. Ayuso (Chair)
BBN Systems and Technologies

Discourse processing remains one of the major outstanding issues in data extraction. The speakers in this session represented a wide range of approaches. Their presentations are briefly described below:

Wendy Lehnert (UMass) presented a view of the coreference problem from the perspective of machine learning, in particular, the idea of treating coreference as classification. To pursue this view, the overall coreference problem needs to be broken up into sub-problems of the appropriate granularity that are amenable to machine-learning paradigms. Once a problem is identified, some of the issues that arise are: finding the right set of features to use in the classifier's training vectors; determining which features are domain dependent and which are domain independent; and finding the right size for the feature set, as more features require more training data. Dr. Lehnert presented an example of this approach for learning rules which recognize appositive constructions. UMass created a training interface where all possible appositive constructions are highlighted for a human-in-the-loop, who responds yes/no to each proposal. Some of the domain-independent features used in the training were syntactic in nature ("does the NP start with an indefinite article?"), and others were domain dependent ("does it contain a corporate designator?").

Jin Wang (NMSU) contended that although achieving a domain-independent reference resolution module would be desireable, it is an impossible goal. As an example, he cited the case of company descriptions versus equipment descriptions, where semantic coreferential constraints for the two classes are very different. In addition, being able to ex[ress coreference constraints in a declarative way is also an unrealistic goal. MUC-5 had an example justifying this assertion: recognizing name aliases--different naming conventions of company names and device names required treatments which performed surgical operations on strings. A more realistic goal, it was argued, is to design a kind of pre-compiler which provides a user with a library of routines from which to draw. The porting job becomes defining the domain-specific routines specifying the merge methods and conditions which are then combined with the domain-independent components to produce a reference resolution program for the new domain.

Chinatsu Aone (SRA) described the new multilingual discourse module of the Solomon text processing system. This module is data-driven, with the goal of achieving a module with core algorithms and data-structures that are domain and language independent. It uses information in three knowledge bases (KBs): the discourse domain KB, which identifies the discourse phenomena which will be handled in each domain; the knowledge source KB which contains a hierarchy of anaphora resolution strategies, such as filtering functions and possible-referent generating functions; and the discourse phenomena KB, which identifies the knowledge sources to use for each discourse phenomenon. Porting of the discourse module involves extending core information in the three knowledge bases.

Dan Moldovan (USC) introduced a proposal for a computational model of reference resolution based on parallel marker-passing in a conceptual network. A semantic instantiation of the interpreted text is created in the network, and activators are propagated outward following relational links. The level of activation of nodes is used as an indication of discourse focus. In performing reference resolution, the most active concept which is acceptable syntactically and semantically is taken to be the referent. A major outstanding problem being studied is how far to propagate markers in the network. A parallel processing approach is proposed to handle computational complexity.

Kazunori Muraki (NEC) described the discourse processor of VENIEX. It combines fragments generated by the parser into microelectronics-capability frames. The relationships among the resulting frames are determined by resolving coreference for entities and capabilities. This is done by applying a set of heuristic rules that achieve good performance in the Japanese microelectronics texts. Reference resolution of a wide variety of expressions is treated, including anaphoric expressions, cleft sentences, and ellipsis.

Marc Vilain (MITRE) described how discourse processing in the Alembic system makes use of a propositional database, simplifying the tasks of reference resolution and event merging. The propositional database is a new component in Alembic and forms the substrate for semantic processing and data extraction. The equality system provided by this propositional substrate is utilized to seamlessly integrate facts from different places in the text when a reference is resolved. Another valuable task supported by a propositional substrate is inference. Instead of having complicated rules for determining the compatibility of event or template objects, and then specifying how they should be merged, Alembic implements an inferential strategy for the merging process.