SESSION 9: NATURAL LANGUAGE PROCESSINGS

Kathleen McKeown, Chair

Department of Computer Science
Columbia University
New York, N.Y. 10027

Traditional approaches to interpretation in natural language processing typically fall into one of three classes: syntax-driven, semantics-driven, or frame/task based. Syntax-driven approaches use a domain-independent grammar to drive the interpretation process and produce a global parse of the input, accounting for each word of the sentence. Semantics-driven approaches use knowledge about the case frames of the verbs to drive the interpretation process. Early semantic parsers often ignored syntax altogether [1, 2] although more recent systems tend to integrate the two components whether primarily syntax or semantics driven (e.g., [3]). Frame or task based parsers use information in the underlying domain to guide the parse. Script based parsers are one example of this class [4]. A more recent example was presented at last year's DARPA Workshop [5]. These systems use the underlying ATIS domain frame that must be built to form a database query to guide the parse, relying on key words and templates to identify information in the sentence that can fill slots of the frame.

Any one of these approaches, however, has drawbacks for the spoken language systems and large text understanding systems being developed today. These systems must be robust. Spoken input is often ungrammatical and speakers use words that are unknown to the system. Text understanding systems must be able to process large quantities of novel text which are likely to contain syntactically complex sentences, ungrammatical sentences, and unknown words. Given the large number of novel sentences that both types of systems encounter, extragrammaticality (i.e., sentences that are grammatical but fall outside the scope of the system grammar) is also an issue. While syntax-driven systems have the advantage of domain independence and provide useful information for further analysis, they are unable to handle ungrammatical sentences since they must produce a complete parse of the sentence. Both semantics-driven and frame-based systems have the advantage of being able to handle ungrammatical and extragrammatical sentences, but they break down on more complex sentences and are not easily transferrable to new domains. All three approaches fail when unknown words are encountered.

The first four papers in this session present three language understanding systems that address these problems. These are the MIT ATIS system ("A Relaxation Method for Understanding Spontaneous Speech Utterances" by Seneff), the BBN DELPHI system ("Fragment Processing in the DELPHI System" by Stallard and Bobrow and "Syntactic/Semantic Coupling in the BBN DELPHI System" by Bobrow, Ingria, and Stallard. These two papers were combined into one presentation), and BBN PLUM ("A New Approach to Text Understanding" by Weischedel, Ayuso, Boisen, Fox, and Ingria). The first two of the systems are spoken language systems, while BBN PLUM is designed to extract data from text. All four papers include an evaluation of their methods. The final paper in the session presents a new approach to evaluation that does not involve testing through task application.

The three systems take a remarkably similar approach to developing robust techniques involving integration of the traditional approaches in a single framework. All three systems are primarily syntax-driven, but have modified their parsers to allow for the production of partial parses, or *fragments*. BBN DELPHI extracts most likely partial parses from its chart, MIT ATIS allows for relaxation of constraints when a full parse cannot be produce, and BBN PLUM uses a modified version of a Marcus deterministic parser where constituents do not need to be attached to a parent node. All three systems use frame or event based knowledge to combine the fragments into a single interpretation. BBN DELPHI and MIT ATIS both use the ATIS frames or templates to guide this task. BBN PLUM uses knowledge of common events in the domain. Integration of semantics (i.e., knowledge of verb case frames) and syntax also plays a role in BBN DELPHI. Semantics is used to reduce the application of plausible syntactic rules by selecting only those rules that produce semantically acceptable interpretations. Case frames are also used to rule out implausible fragments in both BBN DELPHI and BBN PLUM. Finally, both BBN systems also integrate probabilistic language models. For example, statistical models of the likelihood of each syntactic rule are used to select the partial parses that are most likely.

MIT ATIS and BBN DELPHI showed through analysis of the DARPA ATIS evaluation that robust/fallback parsing substantially improved their results. BBN PLUM was evaluated through two additional experiments in addition to the MUC-3 evaluation. Their additional experiments showed that recall grows linearly with lexicon size, while precision remains flat. These experiments support their claim that porting to a new domain can be achieved relatively easily.

The final paper in this session ("Neal-Montgomery NLP System Evaluation Methodology" by Walter) presents a very controversial new approach to evaluation, as subsequent discussion showed. Walter's claim is that task-based evaluation methodologies are too man-power intensive, requiring excess expense and time when porting to a new domain. Furthermore, due to inadequacies in both the

port and in the evaluation metrics, current evaluation methodologies do not reveal an accurate picture of system potential. She reports on an evaluation methodology developed by Neal and Montgomery that provides a descriptive profile of a system's linguistic capabilities. This methodology involves the development of a scorecard, in which each linguistic feature is defined. Systems are then scored against this list of features by a human evaluator who checks whether the system could succesfully produce output when provided with a sentence containing a specific linguistic feature. While Walter indicates that linguistic features can be syntactic, semantic, pragmatic, or lexical, it should be noted that most of the features listed in the example scorecard in the paper are syntactic (e.g., what-questions, what as determiner, what as pronoun, who-questions both with verb and with DO, etc.).

The final session discussion focused entirely on the proposed Neal-Montgomery Evaluation techniques, with many pointing out flaws and inconsistencies in the approach. Several points seemed to emerge repeatedly. Many felt that the evaluation could be not be used for system comparison. Systems work on different tasks and different domains. Whether a particular linguistic phenomena can even be tested depends on whether it is used within that domain. For example, Hobbs pointed out that the SRI parser tested using this approach failed on imperatives, despite the fact that its parser had extensive coverage of imperatives. The problem was that imperative sentences were not used in the terrorist domain on which the system now works and therefore the evaluator could not think of an imperative sentence for the test. The wide margin of disagreement among evaluators (20-100%) over whether a given system could or could not handle a given feature was noted and this raised questions about the value of the methodology.

Many felt that evaluation of fine-grained linguistic phenomena simply could not be done using a black box evaluation. When a sentence fails, it is not possible to tell what caused it. People pointed out that failure could be due to interaction between the linguistic feature being tested and other linguistic features, to other linguistic features in the sentence, or to interaction between linguistic processing and task based processing (e.g., for some tasks it is not necessary to record possessives and thus from the output one cannot tell whether the system handles them). Moreover, success could be due to quirks and ad hoc procedures, thus raising the question of whether black box methodology tests anything at all about syntactic processing. In contrast, people felt that the methodology could be useful for glass box evaluation. Developers could use the check list internally while constructing a parser for intermediate benchmarks. The extensive nature of the list of phenomena identified by Neal and Montgomery was cited as a positive aspect. However, even so, people felt the list does not account for interactions between the linguistic features listed. Many noted that interaction between linguistic phenomena is probably the most difficult part of parser development. There were some who saw the need for a more descriptive approach to evaluation and an appeal was made to involve those who know about evaluation to get involved in order that a good evaluation system could result.

REFERENCES

1. Hendrix, G. G., "Human engineering for applied natural language processing", *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Los Altos, CA, 1977.

2. Schank, R. C. and Riesbeck, C. K., *Inside Computer Understanding*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1981.

3. Woods, W. A., "Cascaded ATN grammars", *American Journal of Computational Linguistics*, Vol. 6, No. 1, 1980.

4. Schank, R. C. and Abelson, R. P., *Scripts, Plans, Goals and Understanding*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

5. Jackson, E., Appelt, D., Bear, J., Moore, R., and Podlozny, A., "A Template Matcher for Robust NL Interpretation", *Proceedings DARPA Speech and Natural Language Workshop*, Asilomar, Ca., 1991, pp. 190-194.