SESSION 5: NATURAL LANGUAGE I

James F. Allen
Department of Computer Science
University of Rochester
Rochester, NY 14627

INTRODUCTION

The first natural language session concentrates on issues of parsing sentences into representation structure. Roughly, there are two parts to this problem:

- finding a description of the structure of a natural language such as English, namely specifying a *grammar* that adequately describes the structures in the language; and
- assigning structure to sentences according to this grammar, namely the parsing process.

This session contains papers that address certain issues from both the perspective of defining better grammars and developing better parsing algorithms.

Figure 1 outlines the space of problems. There are three particular problems that are being addressed here. The first, which involves work in both grammatical development and parsing, is dealing with Robustness. How can we specify a system that does not collapse in the face of disfluencies, unknown words, and structures and words that are simply not known to the system. The second concerns grammatical coverage. Almost any formalism can cover the simple sentences in English, but none can handle complex constructions such as co-ordination and ellipsis very well. The third issue concerns parser efficiency. How can we develop parsing algorithms that can operate in a reasonable amount of time. Each of these issues are discussed below in more detail, and the papers that concern them will be identified.

By far, one of the most pressing issues in parsing is the issue of robustness. In some sense, every paper in this session has a contribution to make to this issue. There are two issues that need to be dealt with. We need to generalize the notion of a "grammar" so that they can describe a wider range of sentences, including many traditionally viewed as "ill-formed", and we need to develop parsing algorithms that can handle such generalized grammars and introduce additional techniques for handling sentences that the grammar still does not "accept".

The first paper in this session concerns generalizing the notion of a grammar. Bobrow, Ingria and Stallard introduce a mechanism called *Mapping Units*, which allow one to more concisely describe the possible variations in word order found in English, and which use semantic constraints rather than syntactic constraint to define the notion of "well-formedness".

Jackson, Appelt, Bear, Moore and Podlozny, on the other hand, attack robustness by introducing a system based on domain specific template matching that can be used to interpret sentences that may not be parsable by a traditional grammar. Rather than replacing the traditional parsing approach, they view this as an additional mechanism that can be used when the traditional techniques fail. This template matching approach was shown to be highly successful in the last evaluation in the ATIS task.

Weischedel, Ayuso, Bobrow, Boisen, Ingria and Palmucci also address robustness by considering techniques that can be used if traditional methods fail. In this case, they are considering techniques of extracting phrase fragments from the text and using semantic techniques to attempt to interpret the utterance from the interpretation of the fragments.

Joshi & Schabes' paper present a new formalism for handling co-ordination. this is a particularly difficult area for grammar writers. If co-ordination is handled at all in a system, it is usually done by a rule that says two constituents of the same type can be conjoined to form a new constituent of that same type. This runs into problems with sentences such as *(John likes) and (Bill hates) beans.* Traditional syntactic theories do not have constituents corresponding to the bracketing parts of this sentence. To handle this, some researchers such as Steedman have generalized the notion of constituent. Joshi and Schabes have taken a different approach. they retain the traditional classification into constituents, and generalize the co-ordination rule with the TAG framework, producing an elegant approach to this difficult problem.

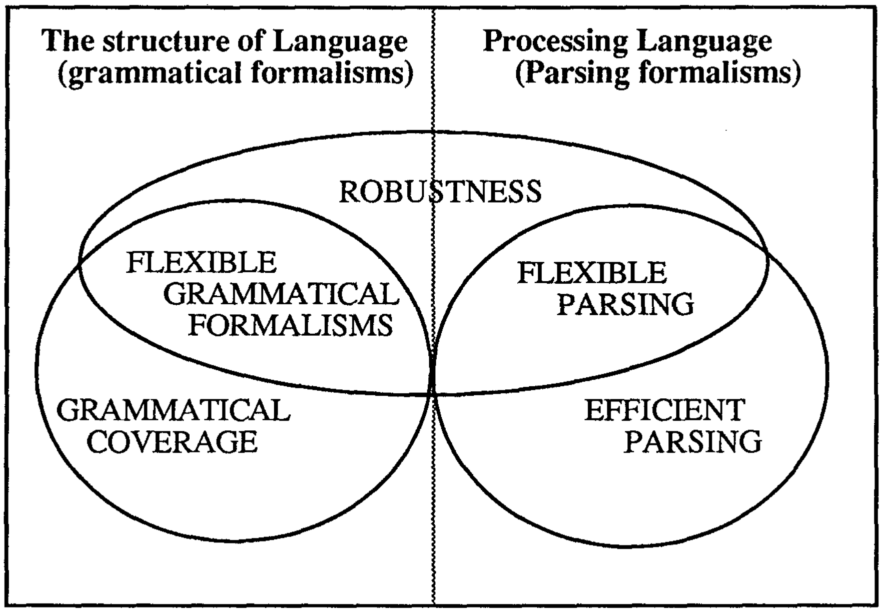

Figure 1: The space of research issues in this session

Finally, Moore & Dowding describe a series of experiments in trying to produce an efficient bottom-up parsing algorithm. In order to handle robustness, it seems that bottom-up parsing techniques are needed for top-down approaches may produce little partial analysis for sentences that fall beyond the scope of the grammar. They describes a series of techniques for speeding up bottom-up techniques, and then introduce a new technique that uses some prediction techniques (i.e. some "top-down" information) that produces a considerable faster algorithm.

Summary of the Discussion

The most important issue that came up in the discussion was the role of limited "ad-hoc" techniques such as the template matcher and their role in research. It is clear, looking at the latest evaluation results, the the systems that use template matching in the ATIS domain are more successful. Yet most everyone is in agreement that such techniques are limited and that there are many examples where it will simply fail. As the test domain becomes more complicated, these deficiencies may eventually show through. But because these techniques are so effective on sentence fragments and ungrammatical utterances, they are clearly here to stay. From an engineering standpoint, these techniques currently yield the best results. But a more interesting possibility should also be considered. These techniques are clearly filling a gap that current syntactically-based formalisms can't address. Interpretation strategies that are strongly driven by semantics and domain expectations about the domain probably will always play a role in a fully robust system.

As a result, an important research issue involves finding methods of combining the more general syntactic models with the domain-specific template matching techniques. The syntactically-based models can handle the more complex relationships that need to be found in some sentences, while the template matching techniques can handle sentence fragments and garbled input. I expect that there will be several papers on this very issue at the next workshop.