SESSION 11 - NATURAL LANGUAGE III

Mitch Marcus

Department CIS
University of Pennsylvania
Philadelphia, PA 19104

INTRODUCTION

The five papers in this session, as well as the ten papers in the other two natural language sessions, can be classified into three broad categories: (1) statistical approaches to natural language processing and the automatic acquisition of linguistic structure (2 out of 5 papers in this session; 8 out of 15 overall); (2) robust processing of texts by combining multiple partial analyses (2 out of 5; 5 out of 15); and (3) fundamental issues in linguistic analysis which will become bottlenecks to processing of more complex texts and interactive dialogues (1 out of 5; 2 out of 15). It is quite remarkable that the largest two of these categories, (1) and (2), would probably not been represented in an NLP session at such a meeting just three or four years ago. From the papers sampled here, it is clear that a revolution in natural language processing is occurring and that the strong surge of work in these two areas is closely linked: both of the text analysis systems discussed in this session bring together statistical and non-statistical techniques. Both the rapid change in research direction and the results achieved are surprising and worthy of special note.

The overview that follows draws out some more detailed commonalities of the papers in this session, and then presents a brief tutorial introduction to two of the problems which several of these papers address.

COMMONALITIES

Two of the papers in this session focus on the automatic extraction from large annotated corpora of rules which can be used to accurately determine two different aspects of linguistic structure. The issue that these two papers address is to what extent can the set of linguistic facts which any natural language processor needs to encode be discovered automatically through the examination of a large corpus of text rather than requiring careful hand programming. A paper by Brent and Berwick of MIT demonstrates a technique which automates a crucial aspect of building the lexicon which lies at the heart of any NLP system. The technique they describe automatically determines which subcategorization frames a particular verb takes, i.e. that "hit" takes a direct object (as in "he hit it") but "die" does it (as in "he died it"). A paper by Meteer, Schwartz and Weichedel of BBN reports on experiments with a stochastic part of speech tagger. Such a system automatically determines for every word of an input text stream what part of speech (noun, verb, adjective, etc) that word is in that particular context. This paper demonstrates (among other results) techniques by which the part of speech of words completely unknown to the system can be estimated with 85% accuracy. This paper also demonstrates that the overall error rate of the system (for both known and unknown words) drops only marginally if the training corpus is reduced from a million words to only 64,000 words. The tagger described here has an overall error rate of 3.3% when trained on the larger corpus, rising only to 3.9% when the corpus size was reduced to 64,000 words. This means that a sufficient corpus to bootstrap a stochastic tagger for a new domain can be annotated in less than the equivalent of one week of full time work for a good annotator, given that the average production rate for part of speech tagging for a single Penn Treebank annotator is 3,500 words per hour.

Papers by Jacobs, Krupka and Rau of GE and by Strzalkowski and Vauthey of NYU demonstrate two different applications for systems which can partially analyze text. The GE paper demonstrates that a partial analysis of unconstrained text, carefully targetted, suffices to extract the information necessary to fill in predetermined frames about particular kinds of events, while the NYU paper shows that the partial analysis of document abstracts can be used to advantage to drive a system which uses classical information retrieval framework techniques. While the GE system uses a fairly superficial syntactic analysis phase to feed fragments to a fairly powerful set of lexico-semantic patterns, the NYU system uses a fairly powerful parser capable of producing fragments if under severe time pressure. A fairly superficial statistical clustering routine is then used to extract information from the resulting perhaps partial analysis.

Finally, a paper by Allen presents a framework for analyzing the discourse structure of a newly collected corpus of task-oriented dialogues in which pairs of participants cooperate to solve simple transportation problems in a simulated world. Allen finds that a fairly small taxonomy of types of interactions based entirely on the intentions of the speaker suffices to categorize all the interactions in the collected corpus.

While these papers are on three different topics, they actually all share a common approach. The research underlying all five of these papers relied on a corpus-based methodology; the research that each of these papers presents is built on fairly large corpora in different domains, some annotated and some not. Furthermore, all but the last of these papers presented systems which have at least one statistical subcomponent. Both the GE and the NYU text analysis systems utilize stochastic part of speech taggers as front ends; indeed the NYU system uses the BBN tagger discussed in this session. The NYU system, as discussed above, also drives a statistical clustering algorithm. It appears that our new statistical tools are already being successfully incorporated in larger systems.

DISCOURSE STRUCTURE

To understand the importance of the paper by Allen (at the University of Rochester) and other work in discourse structure, one needs to understand that every dialogue has an implicit structure which must be recovered to extract the information being conveyed by the dialogue. While this structure is simple and minimal in discourses of two or three exchanges, the

discourse structure of a dialogue involving multiple exchanges can be quite complex. Greatly oversimplfying the problem, one can view this problem as a parsing problem, requiring a characterization of the set of nonterminals, i.e. a set of different kinds of entities which discourses are made of, and some characterization of the form of the grammar itself. Allen's characterization of discourse structure is based upon the view that the basic units reflect the intentions of the participants in the discourse, with larger units reflecting such chunks as a simple Request, a simple Inform, or a Clarification of an earlier point. The "grammar" of discourse is known to involve nested recursive structures, much like context free and more complex forms of grammar. Thus, a Clarification might be decomposed into a Request-for-Clarification, followed by an Inform, followed by an Accept of the Clarification. Casting this into a context free rule, one might have:

Clarification --> Request-for-Clarification Inform Accept

It must be stressed that this oversimplification distorts what we know about the structure of discourses. Allen's paper describes the process of "parsing" the discourse as a what is called a plan recognition process in the artificial intelligence literature. The grammar metaphor used here is oversimplified but makes clear the main point: that a discourse has an elaborate structure, and that any model which will successfully extract that structure must do much more than simply look back at the last sentence or two in Markovian fashion.

DISCOVERING VERB FRAMES

The work presented in the paper by Brent and Berwick of MIT is unlike any presented at previous Speech and Natural Language Workshops. To correctly analyze the full range of sentences which use any particular verb, an NLP system must have some knowledge about the range of verb frames that cooccur with that verb. To be able to analyze the full range of sentences that might occur with the verb "know", for example, a system must know that this verb can occur with a simple direct object ("He knows her."), or a single clause ("He knows she left") or a noun phrase followed by an infinitival verb phrase ("He knows her to be honest"). In systems to date, this information has been laboriously and carefully hand coded for every verb that the system might encounter. The MIT paper presents a new technique which opens up the possibility that a key aspect of the lexical entry for a given verb can be automatically determined given an appropriate corpus. The research reported here demonstrates a new technique which successfully determined sets of verbs which are used with five different subcategorization frames with a false positive rate of no more than 3% per frame. While there are far more than five subcategorization frames, and while the paper does not analyze the percentage of verbs which occur in a given frame that the technique does not detect, the approach presented here appears to be extremely promising. Note also that the sets of verbs which take similar sets of verbs frames are often closely related semantically. The set of verbs which take the three verb frames for "know" presented above also includes "believe", "understand", and "suspect", all closely related in meaning. Thus, if a system can automatically determine the set of verb frames that cooccur with a particular verb, it may well be able to approximate its semantics as well.

A TECHNIQUE FOR IDENTIFYING PROPER NOUNS

One simple new technique for identifying proper nouns was presented by Ken Church of AT&T Bell Labs during the discussion period which should prove of value to others building stochastic taggers. Church's tagger augments the capitalization heuristic used within the BBN tagger and presented in their paper with one other simple check: words which are capitalized and which occur only capitalized more than once within a fixed size text window are taken to be proper names.