Natural Language I

Bonnie Lynn Webber
University of Pennsylvania

Except for the final presentation by Hovy, this session focussed on the use of superficial features of Natural Language in text processing (messages, in the case of the first two presentations, unrestricted text in the case of the second two). This is a very brief summary of a moderator's view of the action.

The first presentation was given by Ralph Grishman (NYU) describing work done by himself and John Sterling in preparing for the second ARPA workshop on message understanding (MUCK-2). Up to that point, their approach to message understanding had been to build a very rich semantics for a domain, which then could be accessed by their message processor in understanding ellipses, noun-noun complementation and other implicit relations, etc. Its very rich semantics gave their system a great deal of power. However, the restricted time participants were given to adapt their system to a new domain for MUCK-2 did not allow Grishman & Sterling to construct a uniform rich semantics for the whole domain. In his presentation, Grishman described their use of Wilk's *Preference Semantics*, which allowed them, in a short time, to capture *some* of the semantics of *all* the domain, rather than *all* the semantics of *some* of the domain, and thereby achieve a greater overall success in the MUCK-2 challenge.

In the second presentation, Jerry Hobbs (SRI International) compared a variety of parse preference strategies that made use of syntactic criteria alone rather than attempting to draw upon semantic and pragmatic criteria as well. For each of the preference strategies, Hobbs presented examples that would fail under that strategy. As one might expect, there was no one purely syntactic strategy that was found to improve results all around.

The next two presentations discussed language statistics and their application to processing unrestricted text. The first of these was given by Ken Church (AT&T Bell Labs) who, in his alloted 15 minutes, presented the results of two separate pieces of work. He first described experiments carried out by himself and his colleagues (Gale, Hanks and Hindle) over several millions of words of text, to characterize co-occurrence relations among words in English texts. He then described other work by himself and Gale, in which they investigated two different methods of estimating the frequency of given bi-grams in a test corpus, given a training corpus – one based on a method due to Good and Turing, the other, a purely empirical method they call *Categorize-Calibrate* or Cat-Cal. Their results led them to advocate the latter in the case of small counts, the former in all other cases.

Julian Kupiec (Xerox PARC) then described a stochastic method for assigning part-of-speech categories to unrestricted text, in a way that eliminates the need for a pre-tagged training corpus and allows some word dependency across phrases.

In the final presentation of the session, Ed Hovy (USC-ISI) advocated a new US effort in machine translation (MT) that would meld current *transfer* and *inter-lingua* approaches into a single approach that would take advantage of recent advances in grammatical theory. He characterized a two-phase effort that would begin with a single modest MT project, and then move to a few small 3-5 person efforts working on limited application domains. Hovy emphasized that the need for MT (including machine-aided human translation, and human-aided machine translation) has not only not gone away, but can only increase, given the changes in Europe brought about by the European Community's plans for 1992 and beyond, and the increasing economic inter-dependency of the Pacific Rim countries. He ended by asking that people interested in getting this new MT effort started contact him.