[Human Language Technology, Plainsboro, 1994]

SESSION 4: MACHINE TRANSLATION

Eduard Hovy

USC Information Sciences Institute
Marina del Rey, CA 90292

About $2\frac{1}{2}$ years ago, ARPA initiated a new program in Machine Translation (MT). Three projects were funded: CANDIDE, built by IBM in New York; LINGSTAT, built by Dragon Systems in Boston; and PANGLOSS, a collaboration of the Computing Research Laboratory at New Mexico State University, the Center for Machine Translation at Carnegie Mellon University, and the Information Sciences Institute of the University of Southern California. All but one of the papers in this section is directly related to these systems. In one way or another, each paper addresses one of the following two major dimensions of variation: *basic approach* (i.e., operation and data collection by statistical vs. symbolic means) and *depth of analysis* (i.e., direct replacement, transfer, or interlingual). This overview first explains these terms and then describes the import of the papers.

Basic Approach

Over the past six years, the CANDIDE group at IBM has gained some impressive results, and considerable notoriety, by performing MT employing only statistical, nonlinguistic, methods. Using cross-language correspondences collected statistically from 3m sentences of the Canadian Parliamentary records, which are bilingual French and English, CANDIDE operates by replacing portions of each French input sentence with the statistically most appropriate English equivalent, taking the whole sentence into account, and then "smoothing" the resulting words and phrases into the most probable grammatical English sentence.

In contrast, the PANGLOSS system takes a more traditional symbolic approach, involving linguistic and semantic knowledge resources such as grammars of Spanish and English, a library of "semantic" symbols that can be composed to represent the "meaning" of each sentence, and a variety of process modules, such as sentence parsers, analyzers, and generators, that employ these resources to convert information from one form (say, a Spanish sentence) to another (say, a syntactic parse tree of that sentence).

The LINGSTAT system, as its name suggests, is a hybrid, involving linguistic-symbolic information for some subtasks and statistical information for others.

Depth of Analysis

The basic theoretical underpinnings of MT involve the amount of analysis performed on the input (source language) sentence during the process of converting it to the output (target language) sentence (since almost all MT systems work on a sentence by sentence basis, multisentence complexities are ignored here). In the simplest possible translation method, a system simply pattern-matches (portions of) each input sentence against a bilingual replacement dictionary and replaces each portion with its target language equivalent. The result is usually massaged in various ways in order to achieve some degree of grammaticality.

A major problem with this approach is the immensity of the replacement dictionary required: since no generalizations are represented, the dictionary needs distinct entries for each form of each word (*see, sees, saw, seen*, or *book, books*, etc.). Even a rudimentary generalization (e.g., storing in the replacement dictionary only the root forms of words) can have a large effect. However, this move comes at the cost of creating two new programs: one program on the input side that recognizes each source language inflected word and replaces it by its root form plus a pseudo-word that carries the additional (number, tense, etc.) information, and another program on the output side that appropriately inflects the replacement root form according to the information in the pseudo-word.

Once you have embarked on this route of analysis, the next step is to notice that languages exhibit syntactic regularities that map regularly. For example, in English the "active" verb appears in what one can call "second position", while in Japanese it appears at the end of the sentence. Without knowing which word is the verb, a direct replacement system has no way in general of repositioning it correctly during translation. However, at the cost of writing two more programs, a syntactic parser (input) and a generator (output), and creating rules that specify how the syntactic wordclasses (verb, noun, etc.) map over from one language to another, you can greatly improve the quality of the translation, since now you get grammatical sentences. This move involves some effort: though by now basic parsing and generation technology is fairly well understood, writing adequately large grammars of languages is a daunting task; no complete grammar has so far been written of any natural language.

It is clear on a moment's reflection, however, that translation on purely syntactic grounds is bound to fail in many instances; think of simple lexical ambiguity by which "the chair called for order" is translated as "the stool called for command", still grammatically correct. Adequate translation obviously requires some sensitivity to the *meaning* of the source text. That is, to improve the quality of translation ever further, you have to write semantic analyzer and generator programs and develop an internal meaning representation notation, upon which you can unleash inference rules about how word meanings combine (disallowing "call" to take inanimate agents, for example). Here the central issue is constructing an adequately large and expressive set of meaning representation symbols. Ideally, these symbols would be independent of any human language; they would constitute the symbols of what is called in MT the Interlingua.

One advantage of Interlingual systems is the fact that they can be significantly cheaper to extend to handle new languages. MT systems that use transfer rules to specify how the intermediate representations map from one language to another require order n^2 sets of mapping rules for n languages (one set of rules between each pair of languages). Interlingual systems, on the other hand, need only order n rules, just between each language and the Interlingua. When the number of transfer rules between any two languages exceeds 1,000, the cost of developing an interlingua starts looking attractive.

The Papers

The first paper in this session, by White and O'Connell, is the one that ties everything together. White and O'Connell have been retained by ARPA to administer the frequent (on average, every 8 months) evaluations of the MT systems. These evaluations have grown from relatively small (the first involved the three ARPA contractors and three or four additional systems) to fairly large (the most recent, of January–February 1994, involved 19 systems). In their paper, White and O'Connell describe the three major evaluation criteria and then provide the systems' scores for the evaluation of May–August 1993.

PANGLOSS is a symbolic system at heart, although some of its newer components, and many of its knowledge acquisition efforts, are of a more statistical or semi-automated flavor. The paper by Okumura and Hovy addresses the core problem of linking a large wordlist used by the JUMAN module to separate words in the input Japanese text to the system's Ontology (its interlingua lexicon). Since the wordlist contains over 100,000 items and the Ontology over 70,000, this linkage cannot be done manually. The algorithms outlined employ a bilingual Japanese-English dictionary as a "bridge". PANGLOSS has undergone other changes as well. It was originally designed as a pure interlingual system, but now includes (as is described in the paper by Nirenburg and Frederking) several MT engines, one of which (lexical transfer) is essentially direct replacement technology. With such a multi-engine system, a central problem is reconciling the various engines' outputs; this paper describes how the best translation is selected by a chart-walk algorithm using dynamic programming techniques to find an optimal cover.

The paper by Smadja and McKeown describes work not at present used in an MT system, though its result is clearly intended to be. The issue is how automatically to construct bilingual phrase (i.e., multi-word collocation) dictionaries from a bilingual aligned text corpus. Like the CANDIDE system, Smadja and McKeown use the Canadian Parliamentary records for their statistically based identification first of the longest multi-word sequence that appears with high enough frequency to be a true phrase, in one of the languages, and then find its translational equivalence(s) in the other. They argue that the Dice coefficient is the most suitable for their purpose.

The third ARPA-funded system, LINGSTAT, is an interactive machine-aided translation system that helps the user compose a high-quality translation from Japanese to English. Architecturally designed along more symbolic lines, the system contains a tokenizer/de-inflector, a syntactic parser, and a word-order re-arranger. As such, it is a classic transfer system. Each of these modules, however, employs knowledge collected and used in the statistics paradigm. The bulk of the paper describes the automated construction of a lexically based word sequence grammar.

The final paper, by Berger et al., describes IBM's CANDIDE system. In its simplest original form, CANDIDE was a pure direct replacement system, for which various kinds of French-English equivalencies had been collected statistically and captured in a so-called Translation Model. Over the past two years, it has become an increasingly sophisticated transfer system, as its developers build more morphological, lexical, and syntactic knowledge into more complex models: equivalent words, fertility (the number of English words corresponding to a French one), word and word-class alignment correspondences, etc. Thus while CANDIDE remains a statistical system at heart, some of the rules or knowledge it uses, especially during early processing, are symbolic/linguistic in nature.