[MT Summit IV, July 20-22, 1993, Kobe, Japan]

The Future of MT Technology

Christian Rohrer
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstrasse 12
7000 Stuttgart 1

I do not expect spectacular breakthroughs in the coming years. Instead there will be continuous improvements of existing systems. These systems will become more robust, will be better adapted to the needs of the users and therefore more widely used than today and furthermore users are adapting their needs (Ken Church). The following areas seem to me especially attractive and we will be doing active research at the University of Stuttgart in most of these areas.

Development of Dictionaries Using Corpora.

One of the greatest bottlenecks in MT are dictionaries. Dictionaries normally do not contain enough information about subcategorization, selection restriction and domain of application. It has been shown that subcategorization information can be extracted semi-automatically from parsed corpora using statistical methods. (D. Hindle and M. Room, 'Structural ambiguity and lexical relations', ACL 1991). In the meantime methods for producing dictionaries even from untagged texts have been presented. (Chr. Manning, 'Automatic acquisition of a large subcategorization dictionary from corpora', ACL 1993). These statistical methods can be applied not only to extract subcategorized constituents in the strict sense (like subject, object, etc) but also to some types of frequently occurring adverbial or adjectival modification (like German 'nachdrücklich' in 'nachdrücklich empfehlen' 'strongly recommend'). Such examples of modifier/head could of course also be subsumed under 'collocations'.

A typical example of selectional restriction are verbs denoting the action of putting on a piece of clothing. In German you have to use the verb 'anziehen' to refer to the act of putting on a jacket, whereas if you put on a hat you have to say 'aufsetzen' instead of 'anziehen'. It is of course well known that in Japanese there are even more different verbs and that their use depends on the denotation of the object of the verb. Such well-known examples will of course be listed in dictionaries. Selectional restrictions of verbs used in sublanguages however are not coded in the dictionaries. I looked up the dictionary definitions of the verbs denoting wear (like 'erode', 'abrade', etc), which were examined in detail in the TACITUS project at SRI (J. Hobbs et al.,'Commonsense metaphysics and lexical semantics' CL 1987) and found that they do not contain the information about selectional restrictions which is necessary in order to chose the German equivalent. With large aligned bilingual corpora at our disposal we can extract this selectional information automatically. (It would be very helpful for MT research and development, if large bilingual corpora could be made available at low cost!)

MT for Specialized Applications.

At least since Bar-Hillel's famous example (the box is in the pen), we know that good translation

requires knowledge of what the text refers to and not just knowledge of what the words mean. With the exception of a few experimental systems, like KBMT at CMU by Nirenburg et al., no one has seriously attempted to build working models of personal computers, copiers or cranes in order to facilitate the translation of the corresponding user manuals. However if one could use models which are built independently by knowledge engineers for expert systems, which are being used for fault diagnosis for instance, then translation based on extensional as well as intensional knowledge of the application domain would appear more realistic.The main effort required will then be the construction of an interface between the model and the semantic representation. There is of course no established methodology for the construction of such an interface nor for the construction of the relevant lexica. The investigation of these problems could lead to important results. If we ever want to reach the stage where examples of the Bar-Hillel type can be disambiguated automatically, we have to embark on this enterprise. The success of this investigation would have far-reaching consequences for all areas of natural language understanding.

MT with Human Intervention

One of the reasons why MT has not achieved the success it deserves is that MT researchers have aimed at fully automated MT. They wanted to be revolutionary instead of evolutionary. On the continuum which ranges from pure human translation to fully automated translation we have to divide the tasks in such a way that the human translator does what he can do best and the computer solves the problems which computers are best suited to solve. One possible approach consists in the interaction between the author of the source text and the MT-system. This dialogue takes place in the source language and serves to eliminate grammatical and lexical ambiguities before the translation. This is a variant of MT with controlled input.Such a realistic attitude, which has been advocated by M. Kay for many years and also at the first MT-Summit at Hakone, is now rapidly gaining ground in the European NLP research community. It was the dominant theme at the conference "Language Technology 2000" organized by the GMD in Bonn last month. Industry in Europe gladly follows this trend because they see a possibility to finally make money in the chronically deficient area of MT.

MT with Controlled Input.

Controlled language is being used in the creation and translation of technical documents. Controlled language can help to improve readability, analyzability and translatability of technical documents. Previous work on controlled language concentrated mostly on English and Japanese. Efforts are now under way in Europe to investigate EC languages like German, French and Spanish. Translation with a controlled sublanguage is recommended by the CEC especially in the context of integrated document creation and management.

Translation without Original

Efforts are now under way to represent the content of documents in a language-independent way, that is, not depending on any natural language. This representation can then be used to generate natural language surface forms. R. Kittredge has been advocating this approach for several years and has already achieved good results. In Germany D.Roesner (Stuttgart/Ulm) is applying this method to the description of the content of manuals for car maintenance. He can generate German and English texts in parallel from this underlying representation. This approach also fits well into the scenario of integrated document creation and management.