[*Language Industry Monitor* 17, September-October 1993]

TMI ’93 Notes

The biannual Trends, Methodologies, and Issues in Machine Translation (TMI) conference has a strong theoretical orientation, but this year’s conference in Kyoto, held the week before the MT Summit, was notable for news about two very concrete systems under development. Microsoft’s Steve Richardson discussed and demonstrated how he and his colleagues at Microsoft are exploiting existing electronic dictionaries in their NLP engine, while CMU’s Eric Nyberg detailed the acquisition of an English-French lexicon for the system CMU is building for Caterpillar. While these two groups clearly have different agendas, they do share one noteworthy ambition: to find ways of at least partly automating the development of knowledge bases for NLP.

Like many NLP research groups, the Microsoft group has licensed the Longman dictionary (LDOCE) and has been experimenting with ways to extract information from its entries for such things as sense disambiguation and determining prepositional phrase attachment. Recently, the Microsoft group has extended these explorations by developing procedures for rendering explicit a vast network of implicit basic semantic relations, such *hypernym, part of,* and *location*, found within the entries. Confining themselves (for the moment) to nouns and single word verbs, Richardson and his colleagues find upwards of 150,000 semantic relations, which they call *backlinks.* Richardson demonstrated Microsoft’s development tools while giving his paper, and the complex trees he could display graphically for entries provided strong visual impact. Parsing some simple test sentences, Richardson switched on a kind of natural language debugger contributed by his colleague George Heidorn, in which the connections pursued by the parser through this network are displayed in a window in plain English.

The Microsoft group has also licensed the 3rd edition of the American Heritage dictionary in electronic form and is currently applying the same techniques to it, and Richardson expects the results to be even more gratifying. Whereas the LDOCE restricted entry language tends to generate loads of connections for these two thousand words – and this is something of a limitation – the AHED should produce a richer, more stratified network because of the wider vocabulary used. Richardson says it is plausible that multiple dictionaries, eventually even encyclopedias, could be incorporated within this architecture. As the system evolves, it could prove to be a viable methodology for partly automating the development of NLP systems for other languages as well as directly contributing to the development of an NLP system for English. Do we see the first stirrings of future MT systems here?

The TMI audience was impressed by Richardson’s display and clearly envious of the Microsoft group’s software tools. Commented one researcher afterward, “it’s bold of Steve to give a live demonstration of his stuff in front of this kind of audience.” Said another, a commercial MT developer, “I guess we have to watch out for these guys at Microsoft.” While it is not known yet when Microsoft’s NLP technology is likely be deployed in commercial software, Richardson did remind us that Microsoft’s programs applications have well-documented API’s for linguistic resources; when the group’s software is ready, it can simply be plugged into these applications.

Meanwhile, at CMU, the first language pair, English-French, of the MT system being developed in collaboration with the Carnegie Group for Caterpillar is nearing completion and should be “production-ripe” by the end of the year. The CMU MT group has focused on Knowledge-based Machine Translation (KBMT) for many years and it represents one of the most promising new directions in MT. However, KBMT will not deliver the much-coveted translation goods unless techniques are found for at least partly automating the process of developing the substantial knowledge bases that will be required for commercial systems.

The first step is supplying the system with a comprehensive lexicon for the Caterpillar domain, and at the TMI, Eric Nyberg related how the CMU team is building a bilingual lexicon on the basis of a 53 MB corpus of Caterpillar documentation. First, they developed a set of tools which generated basic templates for words (based on word frequency statistics and part-of-speech information) that their lexicographers could then refine. This identified a domain vocabulary of 9,000 single words and 50,000 multiword noun phrases. To generate the source language lexicon, they aligned the English source text with the French translation and provided their lexicographers with a bilingual corpus browser. Because many of the noun phrases contained previously translated words, a function was included for scanning for these entries. Nyberg says that an initial bilingual lexicon was thus created in just a few weeks with the assistance of "undergraduates with a background in French but with no expertise in the domain." But because they were only able to extract aligned examples for 12% of the source lexicon, the remainder will need to be handcoded by a professional translator familiar with the domain, obviously a slower and more costly method. As with much corpus work, this promising approach largely depends on the availability of corpora, in this particular instance corpora of well-aligned translations.

Like the Microsoft group, the CMU group is not only aiming to build a working system for a specific context – an English-French MT system – but also to develop a methodology which it can apply to other languages. It certainly needs to: CMU and Carnegie group have another ten language pairs to go for Caterpillar.