Fast-learning computer translates from four languages

February 18 2008

Modern approaches to machine translation between languages require the use of a large ‘corpus’ of literature in each language. Now a European project has demonstrated a cheaper solution which compares favourably with the market leaders in translating from Dutch, German, Greek or Spanish into English.

The European Union now has 23 official languages. That means documents written in one language may need to be translated into any of 22 others, a total of 253 possible language pairs. Small wonder that the institutions of the European Union, and organisations dealing with international commerce, among others, have a keen interest in automating the process where they can.

Efforts to use computers to translate languages, known as machine translation, date from the 1950s, yet computers still cannot compete with human translators for the quality of the results. Machine translation works best for formal texts in specialised areas where vocabulary is unambiguous and sentence patterns are limited. Aircraft manufacturers, for example, have devised their own systems for quickly translating technical manuals into many languages.

The EU has been active in promoting research in this field since the large Eurotra project of the 1980s. In common with other projects of the time, Eurotra used a ‘rules-based’ approach where the computer is taught the rules of syntax and applies them to translate a text from one language to another. This is also the basis of most commercial translation

software.

But since the early 1990s the new concept of ‘statistical’ translation has gained ground in the machine translation community, arising out of research into speech recognition. This dispenses with rules in favour of using statistical methods based on a text ‘corpus’.

A corpus is a large body of written material, amounting to tens of millions of words, intended to be representative of a language. Parallel corpora contain the same material in two or more languages and the computer compares the corpora to learn how words and expressions in one language correspond to those in another. An important example is a parallel corpus of 11 languages based on the proceedings of the European Parliament.

Pattern matching

“Parallel corpora are expensive and rare,” says Dr Stella Markantonatou, of the Institute for Language and Speech Processing in Athens, which coordinates the EU’s METIS II project. “They exist only for a very few languages and in small amounts and in specialised texts. So our idea was to try to do statistically based machine translation without this resource, using just monolingual corpora of the target language. For instance, to translate from Greek into English we use a large English corpus.”

To use a single corpus you need a dictionary for the vocabulary and a way to understand the syntax. In the original METIS project, completed in 2003, the corpus was processed to analysis sentence patterns and the text to be translated was then matched against the patterns.

In Greek, for example, the verb can precede the subject of a sentence. “So if you come in with a Greek sentence, ‘Eats Mary a cake’, you would like the machine to be able to translate it into English and rearrange the

words to make ‘Mary eats a cake’,” explains Dr Markantonatou. “Pattern matching is a good way of doing that because it is able to take patterns from the source language and make them like the target language.”

METIS II takes the principle further by matching patterns at the ‘chunk’ level, a phrase or fragment of a sentence rather than a sentence as a whole, as this makes the pattern matching more efficient.

It can also use grammar rules to generate alternative possibilities for the translation and then use the corpus to identify which is the more probable. For example, where English would say ‘I like cakes’, some European languages might use the form ‘cakes please me.’ So in translating into English, METIS II can test alternative interpretations against the English language corpus. In this example, 'cakes please me' would get a very low score while the closest match 'I like cakes' would score highly.

Four languages

The partners have now built a system that translates from Greek, Spanish, German or Dutch into English. Trials so far show that it performs well in comparison with SYSTRAN, the rules-based market leader in machine translation. Considering that SYSTRAN is based on half a century of development while METIS II has only run for three years, that is quite an achievement. A prototype is already available on the internet.

The problem now is what to do next. Results from METIS II are being followed up in national research programmes in Spain and Belgium, but there are no plans as yet to further develop the whole system. Some of the components created in the project, such as dictionaries and associated language tools, could be marketable in their own right, but would need an industrial partner to provide the investment needed to

turn the prototype into a commercial product.

“For Greek, it would be an excellent opportunity because there is nothing really good for [translating it] at present,” Dr Markantonatou tells ICT Results. “With a better lexicon, fixing bugs and making algorithms more efficient, this kind of thing could work. In another two or three years, METIS could be a very serious competitor to SYSTRAN. It’s a matter of funding.”

Source: ICT Results