The world: a global village called Babel

June 1 2010

(PhysOrg.com) -- European scientists have developed groundbreaking technology to enable machine translation using statistical analysis. Now linguistic diversity can be found in translation.

We live in a global village, and its name is Babel. As information and communication technologies unite the world into a global village, so the diversity of our global linguistic landscape creates new barriers. The smaller the world becomes the larger the language barrier looms.

Europe is an excellent example in microcosm. Political and social cooperation draw the diverse peoples of Europe ever-closer together, but language often separates them. Fully one half of the European population is incapable of conversing in a second language.

The issue is even starker on the World Wide Web, where English has

become the lingua franca. But that status quo is under threat as China and India ramp up their scientific and engineering expertise and simultaneously produce more and more essential information in their native languages. How can we help people communicate now, and how can we overcome the emerging language barriers of the future?

Smart answers

The SMART project believes it has the answer. SMART stands for Statistical Multilingual Analysis for Retrieval and Translation, and the project sought to make statistical methods a viable alternative to current paradigms. In just three years the project has made the technology a robust alternative.

Machine translation is not new. In fact it is one of the oldest problems in computer science. “It was one of the first problems tackled, with work starting in the 1950s,” notes Nicola Cancedda, researcher with Xerox and coordinator of the SMART project.

“Trained bilingual linguists would encode the rules of given languages into a computer program, and the software would use these rules to offer a best-guess at a particular translation.”

Statistical machine translation is not new either. It began in the early 1990s and lets a machine ‘learn’ translation between two languages by looking at thousands of real translations. SMART took that work further by producing robust technology that can match the state of the art in traditional methods. But their platform has not yet had the ‘fine-tuning’ applied to traditional techniques, and so SMART has opened the way to a very promising research path in statistical machine translation.

Inspirational in a network

Their work was inspired in large part by the efforts of the Pascal Network of Excellence (NoE), which sought to develop cooperative ties among Europe's leading players in pattern analysis, statistical modelling and computational learning.

"Seven out of our ten partners come from the PASCAL NoE," reveals Cancedda, "And the impetus for the SMART project came from PASCAL's work. We sought to develop more effective statistical learning methods, apply them to machine translation, and then prove the platform through rigorously measured case studies."

Those case studies focused on computer-aided translation (CAT) and cross-language information retrieval (CLIR). Computer-aided translation is used by professional translators, with the software suggesting possible translations for individual sentences in the target language.

"In our case study, the SMART platform increased words per hour by 5 to 40 percent. Most interestingly, the greatest improvement was seen among the slowest translators," stresses Cancedda.

Enormous boost

This result alone represents an enormous boost in productivity and justifies the project's work. But SMART went much, much further.

While CAT might have the largest commercial potential, the project's work on CLIR will probably have the widest societal impact. CLIR takes place where people try to acquire information from a foreign language document. In the SMART case study, Slovene students, with varying competence in French, sought to extract information from the French Wikipedia.

In the project's subsequent tests, students using SMART's CLIR system

could answer a significantly higher number of questions accurately than those students using currently available tools.

Another allied work effort saw SMART develop confidence estimation to accompany the statistical machine translation. The confidence estimate indicates the likely appropriateness of the translation.

“This is an essential element,” emphasises Cancedda, “Because software providing inaccurate translations is worse than no translation. A translator is better off working alone with his or her dictionary than reading and correcting inaccurate suggestions.”

What makes this work even more valuable is that it could be applied to existing software to make that software even more accurate. Confidence estimation was also an important, and exciting, technical challenge in itself. How do you teach a machine to assess itself?

Back to statistical methods

Again, statistical methods are applied, and the relevance and power of SMART’s confidence estimation varies enormously between different texts. The questions of context and specialist knowledge play a huge role.

Although these are early days in the technology’s development, it can already achieve up to 90 percent estimated confidence in some cases - nine out of ten machine translated sentences are relevant.

In any case, SMART also advanced new research to tackle the problem of context. “Imagine you have a million sentences on one topic, say software. In this case, you can easily use statistical machine learning to create a statistical machine translation software for that topic,” argues Cancedda.

But what if you only have several thousand sentences on another topic, for example airport computer security systems? “In this case, it is difficult to do statistical machine learning. You do not have a sufficiently large sample,” says Cancedda.

“So we developed a tool that can learn the bulk of its translation from one set of documents, and then be specialised to a particular topic with another, much smaller, set of documents. It is not perfect yet but early work shows that this approach could be very promising.”

Real-time learning

And finally, SMART’s last contribution to statistical machine translation is perhaps the most valuable, particularly in cases where only a small set of initial translations exists. SMART developed real-time learning tools that can ‘teach’ software new terms and translations.

“Normally, software is developed and that is the way it will stay for months or years. We have developed tools where the software learns all the time, so it becomes much, much better over time, and so much more valuable,” states Cancedda.

It is a packet of results for a three-year project with just a €3.5 million budget, €2.3 of it from the EU, but it illustrates the kind of focused research that can emerge from a Network of Excellence.

Now, elements of the SMART software will begin appearing in commercial products, notably those supplied by SMART partners Xerox and Amebis. Two open source machine translation systems developed in the NoE - the Sinuhe and the Max-Margin Based Translation (MMBT) systems - were released to the research community and are available for download from the project website.

In all, it means that our global village will be found, rather than lost, in translation.

More information: SMART project - www.smart-project.eu/node/1

Provided by ICT Results