Integrating translation technology

Andrew Bredenkamp reports on the EAMT workshop

Andrew Bredenkamp is a senior researcher at the German Research Centre for Artificial Intelligence (DFKI) in Saarbrücken, and Research Fellow at the University of Essex in the U.K. He also works as a consultant on translation and natural language processing technology for the translation industry in general and Praetorius Limited in particular. You can mail him at andrewb@dfki.de.

On April 2-3 the European Association for Machine Translation (EAMT) held the third in a series of annual workshops on machine translation (MT)-related issues at the World Health Organisation (WHO) headquarters in Geneva. The aim of the workshop was to share the experiences of some users of translation technology, discuss when and how such technology can be integrated into an existing organisation's workflow, and to look at some of the state of the art developments in the area of multilingual document management architectures.

The meeting was well attended, some 40 participants enjoying the pleasant surroundings of the WHO headquarters in Geneva - the meeting being hosted by their Computer-Assisted Translation and Terminology Unit.

The first main talk of the workshop was a report by Achim Blatt (presented by Dimitri Theologitis) on the experiences and future plans of the Translation Service of the European Commission. He described the way the service is managing the huge volumes of translation work they process, how they are introducing technology and how using this technology was changing the way they worked. Probably the most interesting aspect of the talk was the description of the way in which the roles of personnel in the service needed to change. Specifically a new type of job was required (indeed, this was to become a recurrent theme), namely for someone who could run various pre-processes on texts before passing this on to translators.

These pre-processes involved deciding what kind of technology should be used on the text, and retrieving, from various sources, terminology and possibly related parallel texts. The service has invested considerable effort in developing and integrating tools, including the development of their own text alignment tool. They also make use of a large translation memory resource (some 30 GB) which can be filtered according to the text domain and type, the DG (Commission Directorate) and even the translator! The Commission has been a large-scale user of machine translation for some time, and the amount of text being sent for MT (using their Systran system) is increasing annually - last year 260,000 pages were translated, up from 230,000 in 1996.

Dimitri Theologitis also reported on changes in the organisation of the service personnel, whereby groups of (about 20) translators were "fed" texts by (one-two) staff who had been secretarial and were now pre-processing and classifying texts using these tools. Currently some 250 of the 1,200 translators are using the system they have set up, and they plan to make it available to all translators by the year 2000.

The talk gave an encouraging picture of how a range of off-the-shelf and bespoke technologies could be combined with good management to really find, or at least begin to find, the best path from source to target - interestingly he pointed to plans to develop some kind of expert system to automate the selection of the path for documents with respect to the appropriate technology, similar to the EU-funded TransRouter project.

There followed a presentation by Susan McCormick from SAP Germany, which described the work of SAP in relation to the EU-funded OTELO project. The aim of this project is to develop more openness to resources for translation memory and machine translation, and the main focus of this talk was to tell us how the terminological resources of SAP were gradually being converted to a standard format for use with a number of different technologies.

SAP is currently using both METAL and LOGOS, for English-German and English-French MT respectively, and storing its terminology in both

Trados and the "own-brand" SapTerm termbanks. The problem then was that these resources were incompatible and hard (in practice impossible) to synchronise. The solution is OLIF - the OTELO Lexicon Interchange Format - and SGML based format for encoding terminology. The SAP team have both Oracle and Access 8 applications for managing these resources and have spent considerable effort in converting their existing resources to this new format. The company is also working on the various tools for the extraction or "mining" of terminology.

The morning finished with another talk by another experienced MT and TM user, Lou Cremers from Océ Technologies in the Netherlands. They are translating long (400+ pages) documents into four or five different languages with English as the source. The native format for the documents is Framemaker, and the main problems faced have been the integration of the various technologies. The company started off by using XL8 for their terminology management since it was supported by their MT vendor; however they decided more recently that Trados' Translator's Workbench and MultiTerm - which are clearly approaching the status of industry standards - were more suitable. The moral of the story, and others, might be that users should be wary of looking for a single integrated architecture for using translation technology, but rather should insist that suppliers of products maintain openness between their technology and that of others. This battle, and it often is a battle, needs to be fought partly by market forces and partly through user organisations. An interesting aspect of the work at Océ is the amount of effort spent on getting the source text right for subsequent processing.

In the afternoon Doris Marty-Albisser gave a description of the workflow of a company providing translation services to the Swiss banking industry. This was largely a non-technical talk, but raised some interesting issues, particularly that of ownership of terminological and translation memory resources. The speaker's view was that the former was shareable, while the latter was not. However, I suspect that for many users the situation is much more complicated than that, and remains for many translation companies potentially a very serious problem.

There followed a talk by Maghi King on work undertaken in the EU-funded TransRouter project. The talk showed how, even in organisations which might on face value have similar translation workflows, the needs and relevant criteria for decision-making with respect to the use of technology varied widely. The talk described in depth the workflow of three organisations, the Translation Service of the European Community, The Linguistic Services of the Swiss Federation, and the World Intellectual Property Organisation. For each case, Maghi King presented a detailed description of what the current situation was and how translation technology might best be introduced. It was clear that while existing tools generally offered the hope of gains in efficiency, the decision about "right" workflow for the job was always rather idiosyncratic. Like Dimitri Theologitis, this speaker also pointed out that "a new profession needs to be born" - she proposed the title "translation technicians" - if translation technology is to be managed properly in the workflow.

The second day started with a presentation of the PARS translation systems by their developer, Michael S. Blekhman. PARS was originally a bidirectional Russian-English MT system, which has now been extended to German and Ukrainian, and runs on Windows 95 and NT.

Jörg Schütz from IAI Saarbrücken then gave an overview of MULTIDOC, an EU-funded project involving developers and users from the European automotive industry. The aim of the project is to develop workflow architectures incorporating modern translation technologies. The project is at a relatively early phase, and much of the current work is focussing on the architecture rather than the translation technology - establishing an SGML-based system where users retain control at each step of the document production cycle will be a significant result of the project. The harmonisation and standardisation of the tools used in this framework, which is crucial to the successful integration of diverse technologies, is also a major project goal.

Drs Van Wees from The Cap Gemini Advanced Technology Services team, based in Utrecht, gave a short talk about controlled language checkers for English and Dutch they have developed. These applications which run inside Microsoft Word, use full grammatical analysis to search in texts for constructions which the user has identified as problematic for machine translation sytems. The checker worked by highlighting the text in different colours and offering the user often very precise help on the nature of the error and proposed solutions.

On a more user-oriented level, Pierre Lawalle from the World Health Organisation, gave a talk reiterating the needs of users in translation and documentation departments. He provided a useful overview of the state of the art and the problems still

faced in getting tools to work together, and for integrating them into existing document production methods.

J-P Chanod gave the last paper of the workshop, an introduction to the work of the Xerox Language Resources Group based at the Xerox Research Centre Europe in Grenoble. The group is building on the long history of work by Xerox in this area, notably in the use of finite-state technology for natural language processing, from this core technology they have developed NLP components for part-of-speech tagging, noun phrase recognition, etc. These components are then used to build applications, and Jean-Pierre Chanod described the new Xerox Translation and Authoring System, an environment including tools for sentence alignment, automatic term extraction and management, as well as translation memory.

Before the final discussion, Thorsten Mehnert from Wordnet Language Consulting and Services gave a brief mention of a project he is running to facilitate the introduction into small and medium sized companies (SMEs) of some of the complex workflow architectures which are normally only an option for rather large companies.

The conference closed with a discussion, led by Dimitri Theologitis, on what we had learnt from the conference or rather what the potential users in the audience felt were still unanswered questions. One of the most interesting problems related to the personnel needed to use translation technologies, whether they are called language assistants or translation technicians, it is clear that these people are doing a new and complex job and that organisations training translators need to start thinking about how to provide these new skills, as well as convincing students of the importance of acquiring them.

In general, the workshop was a valuable opportunity to hear users describe their experiences of using translation technology in a number of complex document production workflow scenarios. Two recurrent themes persist, firstly developers need to work harder to maintain openness both with other translation tools and more generally with other document production technologies, secondly users need to be aware of the fact that there are no "shrink-wrapped" solutions, and that making multilingual documentation production efficient was a complex, highly technical, but very necessary task.

I should express my thanks to Olivier Pasteur (local organiser), and the EAMT Committee for an interesting, well-organised and informative event. ▪