MT News International

Newsletter of the International Association for Machine Translation

ISSN 0965-5476 Issue no. 6, September 1993

IN THIS ISSUE:

Published in the United States with the generous assistance of Microsoft Corporation.

CONFERENCE REPORTS

MT Summit IV Features "International Cooperation for Global Communication"

Muriel Vasconcellos and John Hutchins

The Fourth Machine Translation Summit, held in Kobe, Japan, on 19-22 July 1993, lived up to its theme, "International Cooperation for Global Communication." The conference was attended by 220 participants from 15 countries. Both the keynote speech and a concluding panel addressed the theme specifically, while reports of several multinational projects highlighted the importance of international cooperation and attested to an increasing trend in this direction within the MT field. As Conference Chair Makoto Nagao pointed out in his opening speech, the International Association for Machine Translation has provided a framework within which such cooperation can be advanced and in fact accelerated. IAMT's impressive growth since its establishment two years ago at MT Summit III is clear evidence not only of interest in MT throughout the world but also of the strong desire of researchers, developers, and users to cooperate at both national and international levels.

MT Summit IV marked a new beginning in more ways than one: it was the first Summit to be organized within the framework of the International Association for Machine Translation, and it was also the first to be held again in Japan, bringing the tradition of biennial MT Summits full circle and starting a new rotation between the three regions of the world. In addition, IAMT gained a new president at the close of the conference when the Association's founder and outgoing president, Makoto Nagao, turned over the gavel to Margaret King, president-elect of IAMT and president of the European Association for Machine Translation (EAMT).

The Hotel Okura Kobe site, facing the magnificent sweep of Osaka Bay, was a fitting venue for this major event. Indeed, one of the special treats of the conference was the welcoming reception, which took the form of a twilight cruise around the harbor.

In his keynote address, Makoto Nagao stressed the importance of MT being utilized appropriately. Citing surveys of MT use by the Asia-Pacific Association for Machine Translation (AAMT) and the Japan Electronic Industry Development Association (JEIDA), he pointed out that a sizable number of translation services count on MT to handle as much as 1,000 pages a month. The studies revealed that successful user sites tend to have the following characteristics: input is already in electronic form, the subject matter is highly focused, and thoroughly customized dictionaries have been built up over time. The break-even point for cost-effectiveness seems to come after about 10,000 pages a year. There continues to be a problem with the quality of both input and output, but the advantages of MT appear to outweigh these problems. Current emphasis is on greater utilization of networks, the development of filters between different word-processing and publishing systems, and the improvement of pre- and postediting facilities. He also noted the increasing use of MT on personal computers for information purposes only. By way of conclusion, Nagao outlined future tasks for all concerned with MT. Users should build dictionaries of specialized terminology, exchange experiences with others, collect

typical corpora, cooperate in the development of guidelines for evaluation, and share all this information through user groups and clearinghouses. Developers, in turn, should clarify the information they give to the public, with specific statements about what their systems can and cannot do, and they should cooperate in the definition of common format codes for the exchange of texts and dictionaries as well as standardized evaluation procedures. And finally, researchers should explore new approaches such as statistical and example-based MT, cooperate with counterpart teams on a worldwide basis, and at the same time continue to press forward with discourse analysis, syntactic and semantic disambiguation, and the build-up of knowledge sources.

The conference featured three invited speeches. The first of these, by John Hutchins, covered the latest developments in MT technology. Since the end of the 1980s the most striking change has been the emergence of new approaches and methods in what is broadly called 'corpus-based' MT, particularly the increased use of statistical methods and explorations of example-based MT. However, the rule-based approaches characteristic of systems in the 1970s and 1980s have also witnessed considerable changes: the widespread adoption of constraint-based and unification formalisms, enabling the development of general-purpose shells for NLP and MT; the move towards 'lexicalist' approaches (illustrated at its most extreme by the 'shake-and-bake' approach), the increased attention to lexical acquisition, and the greater emphasis on the generation of idiomatic output. Other significant developments include multilingual generation from non-textual databases, research in systems for monolinguals not knowing target languages and numerous domain-specific systems developed by companies for specific purposes exploiting well-established methods and techniques of NLP and MT. He ended by speculating about the types of MT systems which may emerge from the various and sometimes divergent methodologies seen at present. He suggested that during the 1990s we may see the appearance of a "third generation" of systems, founded on a linguistic rule base (although less abstract than those typical of previous interlingua-transfer systems) incorporating a mixture of dictionary-derived lexical information, databases of domain-specific knowledge, aligned corpora of bilingual texts giving examples of translations, and the use of probabilistic approaches to lexical and structural transfer and selection.

In an invited speech on the current state of machine translation usage, Muriel Vasconcellos reported data from a worldwide study of MT users. The survey yielded 40 responses, 23 of which gave annual production figures. The data from this subset alone showed that users are enlisting MT to produce more than 170 million words (680,000 pages) a year. On the basis of these figures and the number of other known users (about 80, most of them with smaller installations), Vasconcellos proposed that a very conservative estimate of MT use in the world would come to a total of 300 million words (1.2 million pages). Of the volume reported in the survey, 60% involved the translation of technical manuals, software, and other materials related to localization. At least one of these users found MT "indispensable," and those that reported on productivity cited increases ranging from 30% to 50%, which for them meant both reduced costs and faster turnaround. They were happiest with the fact that MT keeps terminology uniform and saves the need to re-enter format codes in each translation. The most frequent complaints had to do with document preparation, the quality of input texts, and postediting. [A revised version of this paper is reproduced in this issue on pages 12]

The third invited speaker was Toshio Yokoi of the Japan Electronic Dictionary Research Institute. His topic was the problem of constructing very large-scale knowledge bases, not just for MT but for a multitude of natural language processing purposes. He began by stressing the crucial role of natural language for the representation of knowledge, not just as the interface between knowledge databases and their users. Major requirements of large knowledge bases must include: a) the capacity to expand without difficulty, to enhance the quality during expansion, and to absorb information from a variety of different sources; b) support for interpretation by both humans and by computers, for integrating representations in many types of languages (natural and formal), images, diagrams, sounds, etc.; c) demands on the development of new system architectures, where natural language is the core medium of the computer system; d) the ultimate goal of full understanding of natural language in reliable processing. The EDR electronic dictionary project was given as an example of progress towards these ends. In one respect, the electronic dictionary was itself a large knowledge base, and here Yokoi stressed the relationship between the EDR concept dictionaries and the text base providing example definitions. More important perhaps was the experience gained during the project in lexical knowledge acquisition, providing techniques and methods for building knowledge bases in general. He concluded by describing plans for the "Knowledge Archives" project, the creation of a very large database of 'knowledge documents' - in free or controlled natural language, informal knowledge representation languages, in pictures, images and sound.

The EDR project was the theme of other contributions - by Seibi Chiba and Yoichi Takebayashi - who together brought participants up to date with its current status. Hiroshi Yasuhara (another EDR member) concentrated on a description of the research on example-based MT using text data gathered for the dictionary. Research and development on text corpora was the central topic of Yorick Wilks' presentation. He gave a broad survey of developments in message understanding (SRI, BBN, Massachusetts), tagging of corpora (Lancaster, AT&T, Pennsylvania), alignment of bilingual texts (Brown et al., Church, Kay), stochastic grammars, machine-readable dictionaries (LDOCE, CoBuild, etc.), connectionist models,

large text corpora (Lexical Databank Consortium, the European Corpus Initiative, etc.), ending with speculations about the role of corpus-based approaches in MT, which was dealt at greater length in his paper in the proceedings which concentrated on an assessment of the achievement of the IBM Candide project.

As in previous years, the MT Summit provided a forum for the description of major MT and NLP projects. Susumu Funaki described the status of the CICC multilingual project involving teams from Japan, China, Thailand, Malaysia and Indonesia; and Meying Zhu (in a joint paper with Hiroshi Uchida) gave a detailed account of the structure of the interlingua being used. Ahmad Zaki Abu Bakar described the Malaysian contribution to this project and also outlined other MT activity on the translation of textbooks and of news bulletins for the Kuala Lumpur Stock Exchange.

The ATR project was described by Tsuyoshi Morimoto and Akira Kurematsu, with particular emphasis on its experimental speech translation system. In earlier versions more than 90% of utterances were recognized and translated accurately in about 25 seconds processing time. The latest version (ASURA) has been greatly extended in vocabulary and sentence types, but accuracy has dropped to about 60% and processing time has increased to 50 seconds. The speakers were confident, however, that hardware improvements will enable the achievement of near real-time processing. In brief, ATR is claimed to have demonstrated the technical feasibility of an "Interpreting Telephony System" in the near future.

The third major project is also focused on spoken language translation. This is the Verbmobil project described by Wolfgang Wahlster of the German Research Center for Artificial Intelligence. The aim of this long-term project is the development of a portable device for aiding the translation of spontaneous spoken language in face-to-face negotiation dialogues. The first version will aim to provide translations on demand for participants (German or Japanese) who both have passive knowledge of English but are not fluent speakers; i.e. where English is used as a common language in business or technical discussions. The project is planned for 8 to 10 years; in the first four years funding will amount to 60 million Deutschmarks. International collaboration is planned with ATR and with three US groups at Carnegie Mellon, Stanford (CSLI) and Berkeley (ICSI).

A more general view of European activity in MT and related spheres was provided by Loll Rolling (Commission of the European Communities). He described the various projects which have been supported by the Multilingual Action Plan, the Eurotra project, the Language Research and Engineering (LRE) project, the ESPRIT programme, EUREKA projects such as GENELEX and GRAAL, and the ECLAT programme in language technology.

A panel on the evaluation of MT gave rise to a lively discussion with considerable audience participation. Under the chairmanship of Margaret King, some of the key issues in this area were addressed by Jaime Carbonell, Hirosato Nomura, Loll Rolling, and Muriel Vasconcellos. The panellists agreed that there is no single "right" way to evaluate MT and that there should be a variety of methodologies, much like a cook's *batterie de cuisine*, from which several can be chosen in combination, depending on the circumstances. They went on to discuss the differences between progress, adequacy, and diagnostic evaluation, as well as the best approaches for each. A heated debate ensued over the relative merits of glass box vs. black box perspectives. The conclusion seemed to be that a view inside the glass box is essential for progress and diagnostic evaluations and that a "gray box" may be best for evaluations that focus on adequacy: even though users may be more concerned with black box considerations, it is always important to know why a system fails.

Future developments in MT technology were the subject of a panel chaired by Hiroshi Uchida. Six experts took on the challenge of this topic: Christian Boitet, Pierre Isabelle, Hwee Boon Low, Sergei Nirenburg, Christian Rohrer, and Junichi Tsujii. The general agreement was that future MT systems will be 'hybrids', combining the best features of rule-based approaches, whether linguistic rules or knowledge bases, with the more recent stochastic and example-based methods. There would be more attention to specific user needs in the design of actual systems (i.e. fewer general-purpose systems) - selecting the 'best' methods for the purpose. Isabelle argued that corpus-based approaches were more appropriate to machine-aided translation than rule-based methods; indeed, research should concentrate on tools for aiding translation rather than systems for producing translations. Rohrer saw a major role in the future for multilingual text generation from databases; and Boitet speculated about the future for speech translation, the place of MT in multimedia communication (pointing to experiments in TV subtitling as an example), and aids for simultaneous interpretation.

In keeping with the theme of the conference, the third panel brought the substantive program to a close with an overview of current international cooperation in MT and plans for the future. Y.T.Chen, director of the NSF information center, gave the US view, emphasizing the advances in many areas related to language technologies and the growth of the global economy, multinational companies, telecommunication networks, and the appearance of new information- and knowledge-based products. He then outlined the support of the US government to research in these areas, through ARPA and NSF, and the encouragement of international cooperation, particularly the sharing of data and tools, the establishment of standards and evaluation methods, through demonstrations of technology, workshops, joint sponsorships, and the establishment of the national information infrastructure. The Japanese view of was given by Makoto Nagao. He highlighted first the areas which are still problematic and where international cooperation is desirable, even essential: discourse,

selection of texts suitable for translation, terminology problems, multilingual text corpus, learning mechanisms, parallel architectures for MT, methods of evaluation, networking MT systems, common text and dictionary formats for interchange. He ended by surveying current Japanese government involvement in the multilingual cooperative project (CICC), cooperative development of terminology dictionaries (EDR), and support for MT and NLP information exchange centres. A weakness in the Japan picture was the poor progress in building large text databases, with the main impediment being the problem of copyright. He thought also that far more discussion on standardisation was desirable in Asia; there was much activity here in Europe and the United States but Asian countries had lagged behind. Loll Rolling sketched previous CEC involvement in MT, and then went on to describe the present activities of EAGLES on cooperation with text corpora, lexical resources, evaluation and raising of awareness. He emphasized the need for projects and plans to be adequately reported and distributed, and believed that MTNI could and should play an important role in the dissemination of information about international activity. Yang Tianxing described the support for R&D in MT in the People's Republic of China, surveying briefly the current centres and their research activities, and describing China's involvement in the multilingual CICC project. Toshinori Saeki of MITI (Japan) stressed the large-scale support for CICC project by MITI, but he also pointed out the relatively low level of telecommunication networking in Japan and the need for its improvement if international cooperation is to be encouraged.

The final event was the Second General Assembly of the International Association for Machine Translation. The minutes of the Assembly are published on page 8 of this issue.
[Copies of the proceedings are available from the MT Summit Secretariat c/o AAMT, Akasaka Chuo Mansion 305, 7-2-17 Akasaka, Minato-ku, Tokyo 107 Japan. Tel: +81-3-3479-4396; Fax: +81-3-3479-4895.]