Issue No. 31 (vol. 11, no. 1)
Winter 2002
ISSN 0965-5476
Published three times a year

MT News International

Newsletter of the International Association for Machine Translation

In this issue . . .

Of special note ...

- ***MT Summit IX and EAMT/ CLAW Announcements*** *(pages 16-17)*
- ***MTNI 10th anniversary feature - History of the IAMT—told by the founders*** *(starts page 12)*
- ***Graham Russel reviews "Exploiting Parallel Text"*** *(page18)*
- ***Extensive coverage of 2002 conferences*** *(starts page 4)*

Spotlight on the News

MT Summit IX

Fairmont Hotel, New Orleans, USA—September 23-28, 2003

See Page 17 for details and the first call for papers, session proposals...

Lobby of the Fairmont Hotel

Continued on page 17 ▶

MT News International

Issue No. 31 (vol. 11, no. 1)

Winter 2002

EDITOR-IN-CHIEF:

Laurie Gerber
Tel: +1 (619) 200-8344
Fax: +1 (619) 226-6472
E-mail: mtni@eamt.org

CONSULTING EDITOR:
John Hutchins
E-mail: info@eamt.org

CONTRIBUTING EDITOR:
Colin Brace
Fax: +31 (20) 685-4300
E-mail: webmaster@eamt.org

REGIONAL EDITOR, AAMT
Hitoshi Isahara
Fax: +81-774-95-2429
E-mail: isahara@crl.go.jp

REGIONAL EDITOR, AMTA
David Clements
E-mail: dclemen1@san.rr.com

REGIONAL EDITOR, EAMT
Jörg Schütz
Fax: +49 (681) 389-5140
E-mail: joerg@iai.uni-sb.de

Copyright 2002 by IAMT. Permission is hereby granted to reproduce articles herein, provided that they appear in full, and are accompanied by the following notice:

"Copyright 2002 International Association for Machine Translation (IAMT). Reprinted, with permission, from the IAMT newsletter, Machine Translation News International, issue #30, March, 2002."

Electronic copies available upon request: mtni@eamt.org

MT Compendium Update

The market of computer aids for translation changes rapidly from month to month. New systems appear and old ones are discontinued. Companies merge, companies are taken over, companies cease trading. Software is adapted and revised to take account of changing operating systems, different hardware, developments on the Internet, etc. In April 1999, John Hutchins published the first version of the "Compendium of Translation Software". It is now in its 6th revision, and John Hutchins is joined by Walter Hartmann in the task of keeping the Compendium current. An overview of the compendium is available at John Hutchins website.

ourworld.compuserve.com/homepages/WJHutchins/Compendium.htm

The Compendium is an extremely valuable reference for anyone who needs to stay current with the developers and vendors of translation software, and it is free to IAMT members! Contact your regional administrator for password access to the downloadable PDF file.

ELSNET Directory of Language and Speech Technology Experts

This ELSNET website just gets better all the time! An online directory can be searched for individual experts or organizations: *www.elsnet.org/experts.html and www.elsnet.org/organisations.html*

The main objective is to offer people and organisations in need of specific expertise a facility to get access to it, and at the same time to offer experts and organisations opportunities to be found by people who can make use of their services. In April 2002 the directory contained over 850 profiles of experts and over 2500 profiles of organisations (private and public) from 56 countries all over the world.

If you are an expert in one of these fields, or part of an organization, and not yet included, you are invited to enter your data and profile via the web form.

Expert and organisation profiles remain valid for one year after the last update.

More information about ELSNET can be found at *www.elsnet.org*

Online Resource: ELSNET JEWELS

The Joint European Website for Education in Language and Speech (JEWELS) can be found at *www.hltcentral.org/jewels*

The site includes a rich resource of information on training courses, information on EU higher education programmes and recommendations on language and speech curricula. Soon to be added is a listing of all European Institutions with education in language and speech sciences and technology, and a description of materials and tools that can be used for educational purposes.

Community members are invited to visit the site and contribute as well as learn, by adding information on institutions and courses, or links to useful materials and tools.

This project is sponsored by ELSNET and Socrates projects, and supported by ISCA and EACL.

Contact: Gerrit Bloothooft Gerrit.Bloothooft@let.uu.nl Utrecht Institute of Linguistics, The Netherlands. Phone: +31.30.2536042 Fax: +31.30.2536000

MTNI Back Issues Available Online

Look back over MT history in the making! Articles from MTNI issues 1-18 covering January 1992 through October 1997 are available for browsing at *www.eamt.org/mtni.html*

The Changing Business Landscape

LG

The business climate has not been kind to the localization services or the language technology industry in 2002. Corporate IT departments have been slow to invest in new technology, and many companies seeking to cut costs have done so by slowing release schedules for new products, or generally cutting back on translation and localization projects. Here, we look at the impact on our corner of the commercial world.

Localization Service Providers

In January we noted the acquisition of ALPNET by SDL International. That combination brought SDL International into the top tier of international localization companies in terms of size. But this merger was dwarfed by the September acquisition of Berlitz Globalnet by Bowne Global Solutions, which makes Bowne Global Solutions the largest international provider of localization services, with annual revenues in the neighborhood of US$200 million.

Language Tools and Technology

Starting in 2000 or so, a number of people had the inspiration that the localization needed project management and workflow tools. This category of software came to be called "Globalization Management Systems" (GMS). In some cases the tools combine with content management systems. The result has been a glut of hard-to-distinguish products hitting the IT market at a time when companies are reluctant to spend on new technology. Two notable entrants who sold out were Uniscape and eTranslate. Uniscape merged with Trados in May. Trados has now incorporated Uniscape's workflow and project management capabilities into its own offerings under Trados Enterprise Solutions as Trados GXT. eTranslate shifted its business focus from translation and localization services to software and changed its name to Convey Software in early 2002. However, in July Convey sold its services business to Translations.com, selling it's globalization software to the same company in December. The other two players who emerged in 2001 to define the category: Idiom and Globalsight, are still going, and many other software developers and service providers also still have GMS offerings as well.

Familiar MT Developers

Globalwords, which quietly acquired Logos Corporation's MT technology in 2001, has been active and is still developing and licensing the Logos MT engines: *www.globalwords.com/gwt/m2t.html.* A recent flier from Globalwords advertises the integration of Globalwords MT technology into Star-Transit XV, the latest version of the translation memory system.

SDL International is marketing Transcend, purchased from Transparent Language in early 2001.

The Barcelona technology acquired by Bowne Global Solutions as part of Mendez Translations, which it bought at auction in the L&H collapse, continues to be developed by a team in San Diego. A new version is rumored to under development, but Bowne so far has not been advertising or promoting the technology, and it is not clear what their intentions are.

In summary, the industry is in a state of flux and remains unstable as the market remains weak and indecisive in the face of a growing number of competing technologies and services. □

The "GILT" Professional Community

LG

In 1990 LISA, the Localization Industry Standards Association, incorporated and provided just what the emerging localization market needed – a forum where both vendors (translation tools and service providers) and clients (originally just the IT industry: software publishers hardware manufacturers) could work out the standards and best practices needed to rapidly and efficiently prepare computers and software for international markets. The movement rapidly grew to include an increasingly diverse client population, eventually anyone building products of any type for export that required translated packaging, documentation, or even design changes in order to appeal to foreign audiences. The LISA cohort gave us terminology (globalization, internationalization, localization—sometimes abbreviated GIL, or GILT, incorporating translation) to capture the emerging picture of the problems and processes involved in selling and supporting customers internationally in the digital age. The terms are nicely defined on the LISA website at *www.lisa.org/info/faqs.html*.

LISA has historically focused on corporate membership, only adding an individual member category in 2002. LISA membership offers access to their resources and discounts on LISA events, but is expensive compared to professional associations aimed at individuals. In the last two years, LISA has been joined by a number of other associations that try to meet the needs of the constituent groups that first emerged in LISA, and at the same time, LISA has been branching out to maintain its relevance, though many of the founding goals have been satisfied. LISA continues to be the primary hub around which special interest groups form when it comes to industry standards. Hot topics currently are terminology, and translation memory exchange.

Following is a list of the other associations to watch and maybe join. Presented in chronological order by founding date.

The Unicode Consortium

www.unicode.org: Not a recent phenomenon, but one with ongoing importance. The Unicode Consortium was incorporated in 1991, and is the forum where the big software and hardware vendors hammer out an emerging international standard for representing character sets for all of the worlds languages.

W3C Internationalization Activity

www.w3.org/International: Part of W3C (the World Wide Web Consortium), "the mission of the W3C Internationalisation Activity (started in October 1995) is to ensure that W3C's formats and protocols are usable worldwide in all languages and in all writing systems." Individuals and organizations can join the task forces and interest groups that define standards and plan for the future.

Continued on page 21 ▶

Conference Reports

AMTA 2002: From Research to Real Users

LG

The fifth AMTA conference, with the theme, "From Research to Real Users" was held October 8-12, 2002 in Tiburon, California. The main conference sessions filled two and a half days from Thursday, October 10 through noon of Saturday, October 12. Except for the keynote speaker presentations, there were parallel sessions going the whole time, with an excellent lineup of research papers in one room, complemented by user studies, system, presentations, and panel discussions in the other.

The conference was attended by approximately 120 people. The main conference was preceded by a workshop day, and a day of tutorials. The conference exhibitors included IBM, LogoMedia, Lingvistica, Multilingual Computing Magazine, PAHO, Systran, and Terminotix.

The town of Tiburon is located about 30 minutes north of San Francisco, at the end of a small peninsula that extends down into the San Francisco Bay. This brings it quite close to the city, and provides a beautiful view of San Francisco and the bay. The Tiburon Lodge, where the conference was held, was just 1 block from Tiburon Harbor. The harbor includes a ferry terminal where Tiburon commuters board the express to San Francisco each morning. The harbor is also fronted by many convenient and pleasant restaurants where one can watch the boats.

Each AMTA conference is different. The first AMTA conference in 1994 was very exciting for me because it was the first time I had been at an event that included users, developers, and researchers together. I've attended all 5 of the AMTA conferences, and I admit it is one of my favorite events. AMTA 2002 showed how things have changed, and how things remain the same.

How things change

It was rather sad to notice that when people mentioned major established MT vendors, the only vendor they mentioned this time was Systran. At all of the past events, when people – in presentations, or informally – are making statements about MT vendors they would mention Logos, and Lernout & Hauspie (or Globalink before it), and SAIL Labs (or Metal etc. before it). This time, Systran was the only well-known vendor left to mention. It is surprising that IBM has not entered this list of major commercial players in the popular mind. At the same time, the exhibits hall was filled with systems which, if not as prominent yet as Systran, are no less ambitious.

However, there is a trend to counterbalance the disappearance of the old stalwarts. Four emerging MT developers described their technologies at the conference. None have commercial products yet, but they are representative of a cohort of new MT systems that are currently incubating in various parts of the world: Language Weaver is an effort by researchers Kevin Knight and Daniel Marcu, both of the University Of Southern California, Information Sciences Institute, to commercialize many years of research in statistical machine translation. Microsoft has a hybrid MT system that was recently born from the pre-existing components of grammar checkers. Although Microsoft has no current plan to commercialize the system, they have created a number of functional language pairs, with English->Spanish already in use to translate Microsoft technical support documents for users on the Microsoft website. Fluent Machines is a pre-prototype system that appears to be based on some statistical methods, together with phrase-to-phrase mappings. The company has recruited CMU's Jaime Carbonell to their board of directors after impressing him with their innovative approach. And finally, Any Language Communications presented what appeared to be an interlingual MT system.

On the applications side, one of the keynote speakers, Jaap van der Meer, who has managed a number of localization companies, including Alpnet, brought more of a business/localization perspective than we have seen before. Mr. van der Meer particularly focused on the forces that have shaped the translation industry, and in particular the time-pressure that multinational companies feel in product distribution, which is expected to increase the interest in and demand for machine translation.

The program also included four excellent user studies which brought largely good news from the front lines of MT deployment. Systran seems to have been exerting itself to make some really ground-breaking applications possible for innovative users in the U.S. (Cisco Systems, Ford, and NCR) And Corporate Language Services, a Swiss translation company that was also a conference sponsor, described their work with Comprendium (the Metal system) for the banking industry.

The technical papers reflected an almost exclusive focus on empirical methods. This seems to be the completion of a trend that has been underway for many years. Ken Church, the outspoken researcher from AT&T, warned in the opening keynote address, that the pendulum has swung too far in the direction of empiricism and automatic methods. It was surprising to hear one of the prominent innovators from the empiricist camp declaring that we must not neglect linguistics in training students and approaching the problems of translation.

How things stay the same

Two of the three panels included users of machine translation, "Taking MT from Research to Real Users", moderated by Ed Hovy, and "Who's making/saving money with MT and how are they doing it?" moderated by Mary Flanagan. In both, panelists commented that MT developers do not listen enough to users' requests or needs. This is an important subject, but hardly a new one. While there have been many changes that improve usability of MT, the changes have not kept pace with users' desires.

The closing keynote was delivered by Yorick Wilks, of the University of Sheffield, who revisited his "Stone Soup" speech from TMI-92, with support and evidence from recent trends in information extraction, information retrieval and question answering. Wilks cautioned that purely linear statistical methods will never overcome data sparseness, and even the loudest proponents of these methods, whose battle cry is "more data!" eventually incorporate more conventional linguistic information into their systems. He concluded that many natural language tasks will continue to be largely "symbolic", even if large chunks are machine learned. □

Human Language Technology 2002

by Kurt Godden

The HLT conference, held March 24-27, 2002 in San Diego, California, featured talks that ranged widely over areas as diverse as speech recognition and understanding, document summarization, text understanding, information retrieval, government sponsorship of HLT and even the impact of NL technology on genome sequencing. However, for purposes of this report I will focus only on those papers that related directly or indirectly to MT.

As such, there were four talks in the plenary session that related directly to MT, another three in the poster session, and seven system demonstrations that related to translation.

Let us begin with an MT talk that generated as much animated discussion during the Q&A period as any talk during the entire conference. Kishore Papineni of IBM Watson presented on the topic of corpus-based MT evaluation on behalf of himself, Salim Roukos and Todd Ward of IBM and John Henderson and Florence Reeder of MITRE. The audience was the beneficiary of a talk that contained not only satisfying content but also great delivery of the paper entitled "Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results."

Papineni reported on their application of two metrics for MT evaluation, BLEU and NEE, previously reported in the literature, and which meet the "desired characteristics for … MT evaluations," viz. that metrics be automated or inexpensive, replicable, correlative with human judgments, predictive of MT usage, and diagnostic for MT performance improvement. For the record, these characteristics are regarded as desirable by assumption, not by demonstration. Another assumption that holds for both metrics is that evaluation can be effectively accomplished using reference translations prepared by professional translators.

Without duplicating large amounts of the paper, the Bleu metric essentially matches various length N-grams of MT to reference translations, matching word choice and word order to characterize precision, and using recall of translation length as a demerit. Nee matches only named entities from MT to those in reference translations using aligned corpora.

The key message of the paper relate to experiments comparing the application of these automated metrics to more traditional human MT evaluations. Initial experiments involving French-English and Spanish-English showed that Bleu correlates between 0.85 and 0.9958 with human evaluations of fluency, adequacy, and informativeness and Nee correlates from 0.61 to 0.98.

These are interesting results to be sure, but the talk was not without controversy. In particular, Bob Moore from Microsoft Research pointed out that if system developers know that their system will be judged on a particular metric such as Nee, then that system will suffer from an "extreme Heisenberg" effect as the developers optimize for that metric. There was also some discussion whether word error rate, which so highly influences Bleu, is appropriate. Ed Hovy touched on this point (during Q&A of the next presentation) when he pointed out that these metrics ignore the relative importance of individual terms in a particular text, referring to the word 'not' in the sentence, "Do not touch this button or the device will explode."

This presentation was immediately followed by George Doddington who continued the same topic in his paper "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics." Doddington reported on a NIST study commissioned by DARPA to test the automatic evaluation methodology as proposed by IBM. As part of this, some 1994 human evaluation data from both French and Spanish translations under a previous DARPA program was obtained and compared to the automatic N-gram evaluations. Doddington reports that the correlations with the human translation evaluations ranged from 91% to 98% for Adequacy and Fluency, while the correlation for Informativeness was somewhat less at 71% for Spanish and 78% for French.

I also very much enjoyed the paper by Yaser Al-Onaizan (presenter) and Kevin Knight of USC's ISI, called "Named Entity Translation." Al-Onaizan reported on algorithms to generate translations of named entities from Arabic into English in the absence of aligned corpora, focusing on names of persons, locations, and organizations. Included in the problems discussed is the choice of transliteration vs. translation for locations and organizations. The algorithms were informally discussed, drawing on a variety of information sources such as bilingual dictionaries, monolingual dictionaries, and web texts. Very roughly, candidate translations are produced, sorted into an N-best list, and then re-sorted according to the information obtained from the various data sources. Examples included N-best lists that contained 'Bill Clinton' (vs. Bell Clinton) and 'Bay of Pigs' (vs. Gulf of Pigs).

Another problem with an interesting discussion involved UN President Kofi Annan, whose name was not among the original N-Best choices that did include Coffee Annan, Coffee Engen, and others. To handle this kind of situation, a web search is performed looking for documents that contain any of the proposed first or last names. From these documents, all named entities that contain any of the first or last names from the original list are added to the new N-Best list, which is then re-scored. Assuming that the correct translation is found in one of these new named entities (which apparently is true more often than not), then the re-scoring generally ranks it as the preferred candidate.

A bidirectional, English-Croation speech-to-speech translation system for the U.S. Army Chaplain School was described in a poster session by Alan Black, Ralf Brown, Bob Frederking, and Rita Singh of CMU and John Moody and Eric Steinbrecher of Lockheed Martin. Data-driven techniques were employed for the major components of speech recognition and synthesis and MT. A multi-engine, example-based MT system from CMU was used that employed very shallow analysis and a trigram model of the target

Continued on page 9 ▶

9th Conference on Theoretical and Methodological Issues in Machine Translation

Directions in MT Research

by Teruko Mitamura

The 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002) was held in Keihanna, Japan, from March 13 to March 17. Keihanna is located in the southern part of Kyoto prefecture, close to the historical city of Nara. March in Keihanna was still chilly and a bit windy at times, but the weather was generally pleasant. The main conference was held in the NTT Communication Science Laboratories building, in a setting that is spacious and clean, surrounded by lots of trees and bamboo. Unlike traditional cities in Japan, Keihanna is a relatively new "research town", home to many Japanese research laboratories. One of the local taxi drivers told me that he feels honored to be able to give rides to the many great professors and researchers who visit Keihanna from abroad.

The three day main conference was held first, followed by a Workshop and Tutorials. TMI-2002 was supported by the NTT Communication Science Laboratories as part of their 10th anniversary celebrations, as well as AAMT, ANLP and IEICE.

For the main conference, twenty papers were presented and one panel discussion was held. In general, the direction of the conference focused on the synergy between empirical and knowledge-based approaches to machine translation. Recent research on the use of corpus-based, statistical techniques to create MT knowledge bases holds promise for the rapid development of open domain MT systems for new languages. There were a number of papers that focused on Japanese MT, but there were also papers on German, Spanish, Catalan, Polish and English. The papers touched on a variety of topics in machine translation, including empirical approaches, rule-based systems and other special topics.

Some papers focused on MT systems which are integrated into larger applications, such as cross lingual document retrieval, speech translation, or even Sign Language generation. Other papers focused on improvements to existing MT systems, such as adding pronominal anaphora resolution and ellipsis resolution. A number of papers were devoted to lexicons, morphology and dictionary development or maintenance. Since the use of bilingual or tagged corpora for MT has become prevalent, a number of papers addressed how to improve an MT system by using different types of corpora as development data.

Various types of techniques were presented for more efficient acquisition of knowledge for MT systems; these improvements can be considered incremental rather than revolutionary. Nevertheless, there is a realization that researchers are now facing more challenging situations where the directions and goals for MT are diversified more than ever.

The growing amount of text and commerce on the web in various languages is a stimulus for ongoing research in open-domain MT, efficient knowledge acquisition from corpora, rapid deployment of MT systems, and integration of MT with other applications. These new areas will provide challenges for high-quality MT systems for years to come.

In the TMI tradition, the invited talk was "relevant to MT, but not of it". Mutsumi Imai (Keio University) presented her research on how Japanese children learn the meanings of novel nouns and verbs.

The theme of the TMI panel discussion was "New Applications for MT", moderated by Eric Nyberg. Four panelists, Koichi Takeda (IBM-TRL), Harold Somers (UMIST), Kevin Knight (ISI/USC), and Hitoshi Iida (Sony CSL) presented their ideas of how MT would be used in the future. The ideas presented included applications in the near future (e.g., using MT to allow different machines to communicate) as well as "dream" applications for 10 or 20 years from now (using MT to communicate with different species). The conference also included a workshop and tutorials, which were held on March 16 and 17. The workshop and tutorials were intended to make the conference more interesting and useful for a wider audience, and were held at the same time and place as workshops and tutorials in the 8th annual conference of the Association for Natural Language Processing in Japan.

On March 17, three tutorials were given: Example-Based Machine Translation (Eiichiro Sumita, ATR); Statistical Machine Translation (Kevin Knight, ISI/USC); Translation Memories (Timothy Baldwin, CSLI, Stanford University).

More information about the conference can be found at: www.kecl.ntt.co.jp/events/tmi/.

Teruko Mitamura is a research faculty member at the Language Technology Institute at Carnegie Mellon University, and served as program chair for TMI-02. <teruko+@cs.cmu.edu>

TMI-02 Conference Summary

by Laurel Fais

TMI 2002 was attended by 90 participants from 9 countries. More information about the conference can be found at the website: www.kecl.ntt.co.jp/events/tmi/. *Note that from this website, there are links leading to online proceedings with the full text of the papers described below, as well as author affiliations. –ed*

Invited Presentation

True to TMI tradition, the conference began with an invited talk, the subject of which was “relevant to MT, but not of it.” Mutsumi Imai (Keio University) addressed “Building up the lexicon: How Japanese children learn meanings to novel nouns and verbs.” Unlike (most) machine translation systems, children do not come equipped with a

Laurie Fais is a Research Specialist in the Machine Translation Group of NTT's Communication Science Labs. <fais@cslab.kecl.ntt.co.jp>

dictionary of their target language. Yet, despite the classic gavagai problem, children learn word meanings efficiently, sometimes from a single exemplar. Imai described how Japanese children acquire ontological categories, such as count vs. mass, which do not have any morphological cues in Japanese, and how children learn to assign words (verbs) to actions as well. Her talk set the tone for a number of following papers which critically examined particular language problems in MT.

Main Conference Papers

Twenty papers and a panel discussion made up the remainder of the three days of the main conference. There was a balance between papers in which MT was treated as a mature area of research, ready to be refined or integrated into larger information systems, and those which focussed on searching for faster, less labor-intensive approaches to MT in (often quite limited) particular domains.

Exploiting MT in Applications

A number of presentations took as their vantage point the perspective that MT has reached a level of maturity had suggestions about how to take advantage of the usability of MT systems.

Harold Somers delivered the paper written by Jim Cowie and Sergei Nirenburg entitled "Two Experiments in Situated MT." The authors embedded appropriate MT systems into both a crosslingual document retrieval system and a multilingual summarization system. Once documents have been retrieved, or relevant portions of a document organized as a summary of information for a query, the documents can be translated automatically into a target language.

Koichi Takeda explored "Sentence Generation for Pattern-based Machine Translation." This work was spurred in part by the desire to enter the web page translation market, in which translation must be done over broad domains for text which is sometimes of poor quality. Since users also may not have patience for long delays in text generation, it is important to develop fast and efficient generation. Takeda suggests a polynomial-time sentence generation algorithm to achieve just that.

Tomohiro Konuma described "An Experimental Multilingual Bi-directional Speech Translation System" developed along with Kenji Matsui, Yumi Wakita, Kenji Mizutani, Mitsuru Endo, and Masashi Murata. The hand-held device takes as input Japanese, Chinese or English speech, and gives the user tentative speech recognition output and suggested translations drawn from an example database. Once the user selects an appropriate translation, (s)he can have the device "speak" the utterance.

Ian Marshall and Éva Sáfár, in their paper entitled "Sign Language Generation Using HPSG," extended the notion of translation to the qualitatively different linguistic realm of sign languages, in which three-dimensional lexical items involve most of the body in production. English text is parsed and transformed into a discourse representation structure, which is then mapped onto an HPSG sign language grammar. The use of this grammar allows the synthetic generation of signs presented by an avatar on the computer screen.

New Applications Panel

Eric Nyberg moderated a panel discussion that began with four short invited talks, all of which focused on "New Applications for MT."

Koichi Takeda (IBM TRL) started off the talks with a prediction of great growth in the PC market, especially in Asia. He outlined some key features of the MT market: MT embedded in games and chat; use of MT in web content management; and MT for translating web pages. However, he cautioned that high quality, attractive services need to be developed in order to capture the potential MT market.

Harold Somers (UMIST) described "Computer Aids for Minority Language Speakers," in particular, a computer-mediated system for interactively providing information and conducting doctor interviews, with MT for patients speaking limited amounts of English. This sort of application requires sensitivity to socio-cultural issues as well as adaptations for expert users (doctors) and novice users (patients), in languages that don't generally receive attention in the MT field. [Note: the not-quite-adequate term "minority languages" was used throughout conference, workshop and tutorial discussions to denote this concept.]

Kevin Knight (USC/ISI) mused about

Continued on page 22 ►

Roadmap Workshop at TMI-02

by Laurel Fais

The fourth day of the conference saw the Machine Translation Roadmap Workshop, organized by ELSNET (the European Network of Excellence in Human Language Technologies), and moderated by Steven Krauwer. TMI 2002's Roadmap Workshop was one in a series of such workshops in which the goal is to explore the current situation in language technology, construct a vision of where participants would like to see the field go, and outline some intermediate milestones for measuring progress. They allow researchers to identify possible collaborations, synergies, and major challenges, and funders to understand the strategic priorities of the field. Unlike the main conference, the issue of improving the quality of machine translation was a part of the workshop presentations and discussions.

Workshop Invited Presentation

After opening remarks by Steven Krauwer, Satoru Ikehara (Tottori University) delivered an invited talk entitled "Toward the Realization of Typological Semantic Pattern Dictionaries for MT." Ikehara pointed out that in 1965, the major problems identified for MT were computing power and the need to develop translation aids and semantic processing methods. While the first two of those have been adequately addressed, the problems of semantic analysis have not yet received appropriate attention.

The major problem in dealing with semantic analysis is ambiguities in the relationship between structure and meaning in language. Natural language is rational expression exchanged between human beings, which incorporates social promises or conventions about how symbols of natural language relate to the way of life in the community which uses that language. Because human perceptions and experience are three-dimensional, and language is one-dimensional, it is impossible to characterize a strict relationship

Continued on page 24 ►

International Conference on Translation Technology

by Joseba Abaitua

Organized by the Spanish Association of Translation Companies (ACT), a recently created branch of the European Union of Associations of Translation Companies (EUATC), the first International Conference on Translation Technology has taken place at the new World Trade Center in Barcelona, Spain, April 4-6. An audience of around one hundred people, made up mainly of professional translators, and translation students, attended this three-day event that had the support of the Barcelona 2004 Forum. The conference brought together a relatively large number of software companies (Déjà-Vu, Star, Reinisch (vendors of Trados), Web-Budget, TransMC of iLingua), although it was the local newspaper El Periodico and its Spanish to Catalan translation system (developed by AutomaticTrans) that played the starring role. Six round tables were scheduled with topics such as "Internet and translation in the 21st century", "Quality standards in translation services", "Translation curricula", "A world of information", and "Machine translation vs. machine aided translation: What does the future have in store for us?" Among the participants in this last round table were four scholars from the main Catalan universities plus myself from the Basque University of Deusto, and the technical leader of the Catalan edition of El Periodico.

Toni Badia, from Universitat Pompeu Fabra, argued against the old view of Machine Translation as a monolithic closed process. Contrary to this view, the technology now offers the possibility of integrating different specialized agents that negotiate the flow of translation depending on a number of factors: recognition of already translated chunks, text typology, required quality, specialized vocabulary, etc. All other speakers agreed with this view, and supported the idea, put forward by David Farwell, of a future hybridization of symbolic rule-based and stochastic example-based systems. Ramon Pique, from Universitat Autonoma de Barcelona, discussed the role of free resources and the increasing relevance of the Internet as the main working-space for the translator. He also mentioned the new profile of the localizer, as more texts in electronic form become available. Joseba Abaitua, from the Universidad de Deusto, raised the point that translation should at present be considered as part of the whole process of document-production. He also suggested the idea of sharing translation memories in TMX through internet as a way of overcoming the current coverage limitations of memory-based systems.

Ricard Fite reported on the translation procedures in the Catalan edition of El Periodico. The process is mostly automatic, carried out by a relatively simple word-by-word translator that benefits from the closeness of the two languages. Still, the process is supervised by a team of 36 people (mainly linguists and Catalan journalists), who not only post-edit the output, but also select large parts of the Spanish source text for manual translation. These manually-done parts usually concern items written by leading reporters or special collaborators, but may also include news from the Sports and other more idiomatically prone sections. The translation into Catalan takes no longer than three to four hours, with less than half an hour delay from the completely parallel original Spanish version.

For more information on the conference, see: www.act.es/congreso; www.elperiodico.es;

Dr. Joseba Abaitua abaitua@fil.deusto.es; www.serv-inf.deusto.es/abaitua; Facultad de Filosofía y Letras,Universidad de Deusto, Bilbao, Spain

Survey on Research and Development of Machine Translation in Asian countries

by Thepchai Supnithi and Virach Sornlertlamvanich

With the insufficient collaboration on machine translation research among Asian countries, the workshop "Survey on Research and Development of Machine Translation in Asian countries" was held during May 13-14, 2002 at Phuket Thailand, and sponsored by Asia-Pacific Telecommunity (APT). The objective of this workshop was to provide an opportunity for colleagues in this region to know each other's research interest and NLP-readiness, which is expected to extend a fruitful collaboration in the near future.

In this workshop, There were 50 participants from 11 countries and one region, that are, Hong Kong India, Indonesia, Japan, Korea Republic, Lao PDR, Malaysia, Myanmar, Philippines, Singapore, Thailand, and Vietnam. Presentations were composed of one keynote lecture and 17 papers. The first day and the first half of the second day were scheduled for presentations from participants. The second half of second day was scheduled for a roundtable meeting. Dr. Hitoshi Iida gave a keynote presentation, on the activities of the AAMT, such as technical and market trend survey, network for MT researchers and the trend of MT research such as MT evaluation, statistics-based and example-based MT, multi-lingual NLP, tools and practical corpora building. In the paper presentations, experienced countries gave presentations on MT research status in their own countries and MT research techniques, such as spoken language transla-

tion, semantics annotation, research on proper nouns in Asian context, etc. Inexperienced countries gave presentations on research status and infrastructure status, and then raised the problems of doing research in each country. In the roundtable meeting, there were a lot of discussions about collaboration within Asia, standardization for languages within this region, financial support problems, and the possibility of joining the existing working body as a sub-working group.

The future collaboration summarized from this workshop was classified into three levels: standard level, language resource level and, application and research level. Collaboration in the standard level was about making standardization among Asian languages, for example, to establish an Asian chapter for working groups such as ISO, and construct a help-desk operation for standardization. Collaboration in the language resource level was about a collaboration to share language resources, for example to establish a Working Group or Liaison Secretariat for language resources such as coding description, basic descriptors and mechanisms for language resources, representation schemes, multilingual text representation, lexical databases, and workflow of language resource management. Collaboration in the application and research level was about a collaboration to exchange the information for research and applications in this region, for example, establishing a working group or liaison secretariat for MT.

Thepchai Supnithi <thepchai_s@notes.nectec.or.th> and Virach Sornlertlamvanich <virach@nectec.or.th> are researchers at the National Electronics and Computer Technology Center, Thailand. □

Godden on HLT

...continued from page 5

language to select among competing partial translations. Domain-specific dialogs, recycled DIPLOMAT texts and other text from the web were all used for the training corpus, which included some 6 million words in Croatian.

The system was built in less than one year, including the construction of a Croatian synthesizer from scratch, and the final system delivered on a 200Mz Windows system with 192MB of memory, and field-tested with the Army's Chaplain Corps in Zagreb. Just over two person-years of work went into the system, and it was determined to provide useful transfer of information between chaplains and Croatian speakers about half of the time.

Another poster by Y. Gao, B. Zhou, Z. Diao, J. Sorensen, H. Erdogan, R. Sarikaya, F. Liu, and M. Picheny displayed a speech-to-speech research system from IBM Watson, CSLR/Colorado and Texas A&M. A customized version of IBM's ViaVoice was used along with a statistical concept classer and parser to extract source utterances into a semantics-tree interlingua to translate between English and Mandarin. The presenters were careful to point out their coupling of the recognizer with the NLU analyzer in order to assist in the recognition process by using language models that are dependent upon the dialog state or turn. They conclude that a purely statistical speech-to-speech system with semantic interlingua "shows great promise."

The NESPOLE! speech-to-speech translation system was highlighted in a poster session by Alon Levie of CMU, Florian Metze at Karlsruhe, and Fabio Pianesi of ITC-irst in Italy. In NESPOLE! two conversants in the travel domain are connected by video-conferencing and the user speaks one language while trying to communicate with an agent speaking another language. Four languages are supported with the agent speaking Italian and the client speaking either English, French or German. Different ASR systems were used with vocabularies ranging from 4,000 (Italian) to 20,000 words (French). In this application, multi-modal interaction was leveraged to enhance the communicative process. As an example, back translations of the user's recognition results were used as paraphrases that could indicate by their quality whether or not the system was likely to be translating accurately for the other party, allowing the user to click a "please ignore translation" button to signal the other to wait for a new translation.

At this point I will confess to the Gentle Reader that the 12-hour days were wearing me down and I did not attend any of the evening system demonstrations. However, the proceedings inform me that I missed some very interesting demos. Langlais, Foster, and Lapalme from the University of Montreal showed TransType, which is a probabilistic text prediction tool described as an interactive MT assistant for translators.

The NESPOLE! system previously discussed was also shown, as was an extension to the Oasis system from BBN which helps translators create English transcriptions from radio and television broadcasts in Arabic. TExtractor, a multilingual terminology extractor tool for English, French, German and Spanish, was demonstrated by Valderrabanos, Belskis, and Iraola from SchlumbergerSema in Spain. Zhiping Zheng from the University of Michigan showed the AnswerBus question-answering system that supports web information retrieval from queries expressed in English, German, French, Spanish, Italian and Portuguese, while displaying the answers in English. Under the DARPA TIDES program, BBN created the OnTAP system for multilingual browsing, indexing, and other types of linguistic processing for intelligence analysts. Finally, another speech-to-speech translation system for Japanese and English in the travel domain was shown by Okumura, Iso, Doi, Yamabana, Hanazawa and Watanabe from NEC.

In summary, high-quality MT research was well represented at the eclectic HLT conference. I am sure that the next event will also prove useful for anyone interested not only in MT, but in many active fields of language technology.

Kurt Godden, a computational linguist and MT veteran, works at Lockheed Martin Advanced Technology Labs. kgodden@atl.lmco.com □

Special Feature: Speaking of MT

After 40 years of domination by rule-based design, empirical or data-driven MT systems are incubating in various parts of the world and a number will enter the market as products in the next year. Such systems derive at least some part of their capability by learning directly from text. Fluent Machines is one of these which has gotten some high-profile press coverage—including Scientific American and Red Herring in 2002. MTNI's American regional editor David Clements spoke to Fluent Machines' Mike Steinbaum to get the real story. In the next issue, we'll look at the whole crop of upstart MT systems. –ed

Fluent Machines

by David Clements

Fluent Machines, a New York-based company founded in 2001, is developing what company literature calls "breakthrough technology" using elements of both example-based MT (EBMT) and Statistical MT. The company, founded by Israeli immigrant Eli Abir, has created a technology centering on Mr. Abir's theory of the "DNA of language," resulting in two patent-pending processes: the "automated cross-language database builder, and the "n-gram connector". The database builder forms the core of the system's learning component. The n-gram connector is responsible for generation of natural language output. In addition to these two established components, a third technology, AIMT (for Artificial Intelligence) is being developed which will reduce the reliance on fully bilingual text sources for training and may even eliminate a need for them entirely. Fluent Machines claims its advanced processes will provide "a complete and comprehensive solution for achieving human-quality MT." While the two patent-pending processes have been tested operationally using English-French, English-Spanish, and English-Hebrew, Fluent Machines does not yet have any deliverable products, and testing been done primarily on components. The company does not have translation samples to share or examples of translation tests. The Automated Cross-Language Database Builder is based on insights into natural language by Mr. Abir, and enables a computer to generate a database of translation pairs of n-grams without regard to the size of the n-gram . These translation pairs are automatically learned from previously translated written text. The system, through statistical analysis, looks at a complete document or combination of documents (e.g., books, articles, manuals, journals) in two or more languages and begins to distill the translation of all the component parts of the texts.

Although the Database Builder currently operates with "modest processing power and limited access to cross-language texts," Fluent Machines has begun to build cross-language databases that it anticipates, within a year, will be the largest in existence. Fluent Machines is also developing (but has not yet tested) a method that is not dependent on parallel text to glean n-gram translations. The method is called "AIMT" because it focuses on the actual meaning of an n-gram. This method uses large monolingual source and target corpora in combination with existing translation methods or word-for-word dictionaries. The method requires more processing power than the database builder, but will enable the system to learn broader coverage of the language pairs being translated because it doesn't rely on parallel text, which is much less available.

The second of Fluent Machines' patent-pending processes is the N-gram Connector, which connects contiguous, n-grams in a target language with (the company claims) "human-quality accuracy." N-grams will be connected only if the system knows with certainty that the connection will yield an accurate new word-string translation).

The company makes three interesting points about the N-gram Connector:

Each translation added to the database increases the number of word-strings that can be accurately translated in the future by a large multiple because all naturally connecting n-grams in the database can combine with the new database entry. This allows Fluent Machines to translate word-string combinations that the system has not yet encountered. The system can automatically build many new, longer word-strings each time a single new entry is added to the cross-language database.

The system's ability to lock word-strings together only at points where they organically fit is analogous to the process by which a strand of DNA replicates itself. It is this locking mechanism that allows the reproduction of an infinite number of variations from a finite set of building blocks.

Human editors can focus their review on the portion of the translated text highlighted by the system as "not approved" because the system can identify potentially incorrect portions of its translation.

The Database Builder is responsible for completeness, while the N-gram Connector is responsible for human-quality accuracy. Until a cross-language database is complete (or reaches critical mass), the system will continue to yield accurate, but not complete translations. That is, the N-gram Connector will produce translations only for the portions it is confident of. It may not produce a translation for the entire text."

At a time when many commercially available MT systems aim for mere "gisting," Fluent Machines maintains the goal of achieving human-quality MT, although company literature is careful to distinguish human-quality MT from "perfect translations." Fluent Machines has received the endorsement of Dr. Jaime Carbonell of Carnegie Mellon University. According to a 2002 report issued by Dr. Carbonell and distributed by the company, "The EliMT Method is clearly the most promising and theoretically important MT development in the past several years (and probably since the advent of MT itself). It is the one recent development with the greatest possibility of making a major advance in practical large-scale diffusion of MT technology." EliMT is a term coined by Dr. Carbonell to refer to Eli Abir, the inventor of the Fluent Machines technology. Dr. Carbonell now serves on the Fluent Machines Board of Directors.

According to Mike Steinbaum, the company's COO, Fluent Machines currently has 13 employees So-called "EliMT" is

"memory and processing power intensive," and remaining tasks, besides parallel corpus acquisition, include combining the two processes discussed in this article, as well as increasing computational speed and efficiency. In addition, the company will continue the development of its AIMT methods. The Fluent Machines system, he says, "knows what it knows," and will produce "incomplete translations, but not wrong translations." The company expects that as it refines its algorithms and build larger databases, its system will lessen the incompleteness of translations and approach ever closer to human quality, extending the range of the system beyond the common Euro-centric and English-centric pairs.

Fluent Machines is a subsidiary of Meaningful Machines, Inc., a technology development company. The sister company of Fluent Machines is Internet Driver, which has developed a patent-pending technology that "provides the Internet's missing piece of multi-lingual infrastructure, allowing users to access the entire *existing* Internet (domain names, e-mail addresses and subsites) using any language's character set."

For more information on Meaningful Machines and its subsidiaries, visit the Web site at www.meaningfulmachines.com, or contact Mike Steinbaum, COO, at mike@meaningfulmachines.com. □

A Chat with IAMT President, Eduard Hovy

IAMT President Eduard Hovy has been involved in the AMTA almost from the beginning. He was elected vice-president in 1994, served as AMTA President from 1994-2000, and as IAMT president from September 2001. As a natural organizer and visionary, he has played an influential role in both the AMTA and the Association for Computational Linguistics (ACL), in which he has also long been a member and officer – most recently as ACL president for 2001. Hovy is a perpetual motion machine, moving from the forefront of one technology trend to the next. A new direction is "Digital Government", a program of the National Science Foundation, which he is helping develop together with colleague Yigal Arens, also of USC/ISI. This program aims "to promote emergent information technologies by creating partnerships among government, industry, and academia, providing access to the information, partnerships and financial resources available to create the Digital Government of the 21st Century." MTNI caught up with Dr. Hovy in May 2002 to talk about IAMT.

MTNI: What are the accomplishments of the IAMT and its daughter associations? How do you see their role and importance?

EH: The IAMT and the regional associations have had a key role in establishing communication between users, researchers, and developers. Other research associations, such as ACL and SIGIR, have a strong academic focus, while users never participate. User/commercial forums, such as the magazines *Multilingual Computing and Technology* and *Speech Technology*, have no research/theoretical component. While the MT community is also focused on its own problem, the IAMT has done a better job of integrating across users, commercial enterprises, and researchers. Many productive collaborations were born in friendships formed at IAMT events such as the regional conferences and the MT Summit. Without such events, these connections wouldn't have been made; people simply wouldn't have met each other.

A second accomplishment is to provide a focus for MT: a natural distribution point for such products and services as the *MT Compendium*, a place where the non-research world can hear about new developments in research, and a channel for information about current events and funding opportunities.

MTNI: Where should the IAMT be in 5 or 10 years? How should the association evolve to stay relevant?

EH: There are many ways it could evolve, but what I hope will happen is that the various natural language processing (NLP) communities will flow together to form a broader community, one cutting across the various NLP technologies. Today, the communities around MT, Information Retrieval / web search, speech recognition and synthesis, question answering, text summarization, etc., all have somewhat separate identities, conferences, and (to some extent) technologies. Many of the professional groups, for example IAMT, SIGIR (Information Retrieval / web search) and ISCA (speech recognition), tend to be focused on their own particular problem. The ACL, which cuts across the NLP technologies, is very focused on research. I would love to see a future in which the subcommunities regularly meet in a single language technology conference. This will foster cross-cutting applications, such as speech-based translation, multilingual web search, etc., and will enable techniques, inventions, and funding to flow much more easily.

There has been some development along these lines already. It took 20 years before statistical techniques invented for speech recognition in the 1970s moved into MT in the early 1990s. Today there is much more awareness among researchers. The organizational infrastructure is evolving more slowly, though. Over the past 6 years I have worked within the ACL to help redefine its structure more toward the IAMT/AMTA approach of an umbrella organization and regional associations. In this regard the IAMT, though a smaller organization, is well conceived.

Over the next few years, the IAMT and its regional associations can take more of a leading role in reaching out toward other organizations, thereby enriching its technology base and commercial footprint. Multilingual web search, multilingual text summarization, speech translation, and similar applications need not be "something plus MT"; it is possible to integrate the processes more and search for quality improvements, processing speedup, and cost reduction.

Another aspect is rather important for IAMT and the regional associations. It should become much more active in providing information and educating the general public. Over the past decade, the Artificial Intelligence community has sponsored an immensely popular event at its conferences: the Robotics competitions have drawn students, high-school children, the general public, etc. This event has drawn a lot of media attention and has

Continued on page 22 ▶

Association News

IAMT: The International Association for MT

With this issue we conclude the 4-part series on the IAMT and its daughter associations. In addition to a description of the association, we have the history of the association, provided in accounts by the founders themselves. This issue marks the 10 year anniversary of MTNI, and starts its 11th year.

The Association Itself

The IAMT, incorporated in 1991, is the umbrella organization which links the three regional associations: AMTA (Association for Machine Translation in the Americas), EAMT (European Association for Machine Translation), and AAMT (Asia and Pacific Association for Machine Translation). For historical reasons, the IAMT is incorporated as a non-profit organization in the United States. The IAMT bylaws state "The purposes of the IAMT shall be exclusively nonprofit, scientific, and educational. It shall bring together users, developers, researchers, sponsors, and other individuals or institutional or corporate entities interested in machine translation..." All members of the regional associations are automatically members of IAMT.

Regular Events and Activities

MT Summit: Held in odd-numbered years, and rotating between the three regions: the Americas, Asia, and Europe. Although the MT Summits are coordinated by the IAMT, each Summit is organized by, and is the financial responsibility of, the regional division where it is held. The dates and locations of the past eight MT Summits can be found at: *www.amtaweb.org/activities.html*

Officers

The IAMT council is composed of representatives of the three regional associations, plus the editor of MTNI. The IAMT presidency rotates between the three regional divisions. The new president takes office at the MT Summit, and is from the region in which the next Summit will occur.

Publications

The IAMT publishes this newsletter, Machine Translation News International. In addition, the IAMT has underwritten and promoted the publication of the "Compendium of Translation Software", edited by John Hutchins.

General Assembly

Meetings are held at the MT Summit, and provide an opportunity for discussion and voting on association business with the participation of all regional members.

Awards

Beginning with MT Summit VI in 1997, the IAMT council established a "Award of Honor" award to acknowledge outstanding contributions to MT. This award is given at the General Assembly meeting during the MT Summit.

Bylaws

The Association bylaws are available online at:

www.amtaweb.org/IAMT-bylaws.html

The Beginnings of the IAMT: Four Stories

Following are accounts from four of the founders of IAMT and its regional associations: Muriel Vasconcellos, the first president of AMTA, and the one who handled the association paperwork, Makoto Nagao, the first AAMT and IAMT president, John Hutchins the first editor if MTNI (these three also happen to be the first three recipients of the IAMT Award of Honor), and Veronica Lawson who helped get things started...

Muriel Vasconcellos

The Nuts and Bolts of IAMT

In late 2001, I met with Muriel Vasconcellos in preparation for the coverage in this issue on the history of IAMT. She generously took the time to share information about the history of the IAMT and AMTA, the legal status of the two associations, the dates and key events. I was moved by the drive and energy on the part of the early organizers that brought the associations into existence. I was also struck by the almost fragile nature of our status as a non-profit asso-

IAMT Council

President	Eduard Hovy	USC/ISI	AMTA
President-Elect	Junichi Tsujii	University of Tokyo	AAMT
AMTA President	Elliott Macklovich	Universite de Montreal	AMTA
EAMT Chair	John Hutchins	EAMT	EAMT
Secretary	John White	Northrop Grumman	AMTA
Treasurer	Claudia Gdaniec	IBM Research	AMTA
VP for Programs	Hozumi Tanaka	Tokyo Inst. Of Technology	AAMT
VP for Information	Bente Maegaard	Center for Sprogteknologi	EAMT
Other member	Viggo Hansen	Zacco	EAMT
Other member	Yuzo Murata	AAMT	AAMT
Other member	Key-sun Choi	KAIST	AAMT
MTNI Editor	Laurie Gerber	Language Weaver	AMTA

ciation, the value we get through non-profit status, and the fact that as an association, we need to be mindful of maintaining it. This is potentially a difficult issue for an association with a rotating volunteer executive committee and board of directors, where corporate memory can easily fade. The questions and answers toward the end of the interview focus on the legal status of the association and will probably not be of interest to the casual reader - in fact Muriel suggested that they not be included as part of the interview. But I have included them here to make sure that these points about the association's history are a matter of record, and easily accessible to present and future officers.—LG

MTNI: How did the discussions about an international association for MT get started? Who were the people? What made the timing right to do it?

MV: In 1989 the machine translation field was finally emerging from the long pall cast by the ALPAC Report and many players were entering the arena. The same zeal and that had been so necessary to win out over ALPAC was now fueling a scramble for scarce funds, and in some cases contributing to the formation of intensely committed and competitive cliques.

Although the subject of a worldwide association of people interested in machine translation may have come up at the MT Summit in 1987 (which I did not attend), certainly the first serious steps were taken at the International Forum for Translation Technology, held at Oiso, Japan, on April 26-28, 1989. The Forum had been convened by Professor Makoto Nagao to explore the practical implementation of MT, and more than 400 people attended. Professor Nagao chaired the event with great skill, sensitivity, and humor; his masterful "emceeing" of the lively discussions created a feeling of relaxed openness among the participants – an eagerness to cooperate – that I had never before seen in a meeting of this kind.

The meta-message of the Forum was clear: open communication and cooperation were crucial to the success of the MT field. Veronica Lawson and I began talking about how important it was to join forces and work together. She told me that she had been thinking about an MT

Continued on page 20 ►

Makoto Nagao

The Beginnings of IAMT

The first MT Summit was organized at Hakone, Japan in September 1987, where our main intention was to report on the final result of the Japanese Governmental MT Project which took place during 1982-86, conducted by myself. Many MT researchers from the U.S. and Europe participated in the event, and it was successful enough that a second Summit was scheduled in Munchen in 1999. We proposed to form the organization, IAMT, at Hakone, and a serious discussion was also held at the International Forum on Translation Technology which was held at Ohiso, Japan in April 1989, and continued at the Second MT Summit in Munchen, August 1989 (I was the responsible person for promoting the organization). In both of these meetings MurielVasconcellos, Veronica Lawson, Christian Rohrer, and many other people who actively participated in the estalishment of IAMT gathered together. The IAMT was formed in July 1991, on the occasion of the third MT Summit in Washington DC, as a joint association of regional associations in Asia, Europe and the American continents. The AAMT (Asia-Pacific Association for Machine Translation) was originally established as JAMT (Japan Association for Machine Translation) March 1991, and I remember EAMT and AMTA were established a month or two before IAMT, if my memory is correct. These regional associations were essential for the establishment of IAMT.

I think that many people expected the practical usability of MT systems just at the time of the formation of IAMT, but they were largely disappointed. Then in the middle of the 1990s, MT systems became very cheap, and available on PCs with more and more rich dictionaries, and the quality of MT was significantly improved. I think that the time is coming for the second boom of MT which will not disappoint users either in the form of MT on PC or in the form of on-line requests for MT service.

I would like to recommend that IAMT takes an initiative to persuade the governments of advanced countries to develop multilingual MT or MT for non-major languages, or at least the cooperative construction of MT dictionaries for these languages. □

John Hutchins

Editing MTNI 1992-1997

In 1991 at the third MT Summit in Washington I was approached by Makoto Nagao and Muriel Vasconcellos. I had heard that an international association was being planned, but I knew no details. It was therefore a considerable surprise for me to be asked to be the editor of a newsletter for the new association. Although I had written quite a lot about machine translation, I had no experience at all of editing a newsletter. After much hesitation and many second thoughts, however, I accepted the invitation to start one. What it should look like, where it should be printed, who should contribute, and how often it should appear were all matters left in my hands.

Virtually all that was agreed at this time was that the first issue should appear the next January (1992), that issues should be printed at three or four monthly intervals, and that each regional association should be responsible for the printing and distribution of the newsletter to its own members. I had six months to set the operation in motion.

My first step was to ask for advice from Geoffrey Kingscott, whom I had known for a number of years, and who was the editor and publisher of *Language Monthly* (later *Language International). We discussed content,* layout, getting contributions, printing, distribution, and much more. Geoffrey's company Praetorius agreed to undertake the designing of the newsletter and the printing of copies for EAMT. The intention was that the newsletter should have a uniform appearance in each of the regions. The other associations were to be sent electronic copies of the complete

Continued on next page ►

John Hutchins

...continued from previous page

newsletter issue, with the design and layout decided by Praetorius and myself, from which they could print and distribute copies for their members.

As for the name of the newsletter, I thought of various possibilities but decided in the end on something quite simple: *MT News International.* The name seems to suit its function well – at least nobody, to my knowledge, has yet suggested an alternative.

It may be noted that from the beginning we had considered an electronic newsletter – either instead of, or in addition to, a traditional printed one – but it was felt by all concerned (i.e. the founders, the IAMT Council, and the regional associations) that many members did not have the facilities to receive it in that form, and that most members probably preferred hard copy in any case. The possibility of electronic distribution was left for a later date.

What was the newsletter to include? Most obviously it had to carry news about the associations (IAMT, AMTA, EAMT, and JAMT – shortly to become AAMT), e.g. association bylaws and regulations, notices for members, reports of meetings and assemblies, and notices and announcements of conferences and workshops. Also, it should include reports of any conferences and workshops related to MT held by other organisations. There should also be a calendar of forthcoming events, meetings, workshops and conferences. For meetings directly devoted to MT there were of course more details, e.g. calls for papers, invitations to participate, and registration forms when appropriate.

What else? An important function of the newsletter was to provide information about new MT systems, new companies, new products, new research projects, and so forth. Each issue contained extracts from press releases, items from other journals, information downloaded from the internet, etc., reviews of recent publications – including substantial ones for the various JEIDA reports – and lists of articles related to MT appearing in magazines and journals. In addition, the newsletter included reprinted articles or extracts from other journals, e.g. the *AAMT Journal* (translated with help from various members), *Language International, and Language Industry Monitor.*

To assist me, regional editors were appointed – initially they were Tom Gerhardt for Europe, Joseph Pentheroudakis for North America and Hirosato Nomura in Japan. Later Jörg Schütz became the European representative and David Clements represented North America. Through them I made contact with possible reporters of conferences and reviewers of publications, and they sent me notices of new products and services and other items for inclusion.

As most editors of newsletters know, persuading busy people to write (and submit promptly) is a time consuming and often frustrating experience. Quite often, as far as conferences and new publications were concerned, I found myself doing reports and reviews shortly before deadlines. Most problematic of all was trying to persuade users of MT systems to write about their experiences. From the beginning there was plenty of news to cover the interests of researchers and vendors, but for the users it was always a struggle.

The early issues of MTNI did contain some more substantial items from time to time. There were Muriel Vasconcellos' survey of MT users (no.6, September 1993), Karin Spalink's proposals for an evaluation methodology (no.9, September 1994), the obituary of Paul Garvin by Christine Montgomery (also in no.9), the space devoted to recent patents (no.10, January 1995), and the major survey of MT users by Colin Brace, Muriel Vasconcellos and Chris Miller, with an annexe of current commercial MT systems (no.12, October 1995). And, believing that readers ought to be interested in the history of MT, I wrote a series of articles ("From the archives") marking MT events of thirty, forty or more years before..

Getting the first issue out took me longer than expected – in hindsight, not surprisingly – and the first issue was not distributed to members until March 1992. However, from that time onwards I was particularly concerned that all deadlines should be met. There is nothing worse than for a newsletter containing current information to be published and distributed when the news has become old or outdated. As I had a full time job, I did not feel that I could physically and mentally produce an issue four times a year, and so an early decision was made to publish three times a year. The decision has certainly caused confusion, since many people assume that such a publication ought to be quarterly and so they expect to receive four issues for their membership. However, I took the view that I would rather be on time with three issues than risk being persistently late. After the first issue, MTNI was timed to appear in January, May, and September.

The first two issues of MTNI were printed separately in Europe, the United States, and Japan. It was, however, soon concluded that separate printing would be too expensive. It was agreed that from no.3 onwards EAMT's copies should be printed together with the AMTA copies in the United States, and that the EAMT copies should be sent in bulk mail to the EAMT offices in Geneva, from where they would be sent out to members. This arrangement continues to this day. What this decision meant was that the responsibility for DTP and printing was transferred to the AMTA regional editor, Joseph Pentheroudakis. I supplied him with the contents, and with indications of what should appear where; but he prepared the publishing layout and supervised the printing – with some assistance from his colleagues at Microsoft. It is to the great credit of Joseph that he performed this job so well that publication of MTNI was rarely more than six weeks after I had put together each issue.

In the first three years, MTNI gradually increased in size, from about 20-25 pages in 1992 to some 30-40 pages during 1995. However, it was becoming evident that the income of IAMT from membership subscriptions was not keeping pace, and from this date onwards the size of issues had to decrease. It meant that coverage could not be as full as hoped; items less central to MT were dropped completely, fewer details were included about conferences, about new products and about current research projects, and the more substantial articles ceased.

There had to be other changes. It was becoming more difficult for Joseph to continue with the publishing side, and in

1994, Muriel Vasconcellos proposed that the DTP and printing should be taken over by Jane Zorrilla, who had performed great services for the *ATA Chronicle*. *Jane* refashioned the MTNI masthead and designed a new IAMT logo, and she introduced other improvements in layout and format – with the results much as they are today. In addition, in order to reflect more realistically the actual dates of publication and distribution, it was decided to change the dates given on the masthead: first, by shifting a month to February, June and October, and then as February/March, June/July, etc. But, for various and multiple reasons, delays in printing and distribution continued – which was most disappointing for everyone.

By this time, I had been made president of EAMT in 1995, and I was due to be President-elect of IAMT after the MT Summit in 1997. I decided that I could not to continue to be MTNI editor – in my view, nobody ought to have more than one official duty in the association – and at the meeting of the IAMT Council in October 1997, I resigned and proposed that Muriel Vasconcellos (then coming to the end of her period as IAMT president) as my successor. She had already been active, unofficially, in the backroom processes with Jane Zorrilla, and it seemed appropriate that she should take over fully. Muriel continued as editor until early in 2001, since when the editorship of MTNI has been in the capable hands of Laurie Gerber.

I enjoyed my period as editor; it was hard work but most rewarding, even if frustrating from time to time. From a personal point of view, it kept me in touch with the rapid developments in MT and I enjoyed contacts with all the leading characters and personages of what continues to be a fascinating and challenging part of the 'language industry'. My performance as editor I leave to the judgement of others.

Readers may like to know that electronic versions (HTML) of back issues of MTNI are being made available on the EAMT website. So far, issues 1-18 have been uploaded.

See: www.eamt.org/mtni.html □

Veronica Lawson

The Seeds of IAMT

MTNI Spoke to Veronica Lawson by email earlier this year.

MTNI: The MT Summit conference series started in 1987, several years before the IAMT was started. How did the idea for an international association get started?

VL: I proposed it while giving a paper at a conference on translation technology in Oiso, Japan, in late April 1989. Muriel had been thinking along very similar lines. Professor Nagao, who was responsible for inviting us to that conference, welcomed the proposal. He really picked it up and ran with it. As I remember it, he suggested that the principle be put to the coming MT Summit conference in Munich in August 1989, so that detailed proposals could be worked up and placed before the following Summit for approval two years later.

You can always tell the pioneers by the arrows in their backs.

MTNI: How long did it take for the idea to become a reality?

VL: That long: two years and four months or so, to the MT Summit in Washington D.C. in 1991. For the first few months it was essentially the three of us faxing early draft constitutions between Kyoto, Washington and London, establishing the basic principles: a worldwide umbrella covering three big regional groupings (which became Asia-Pacific, the Americas and Europe), each including everyone in the field (users, developers, universities, etc).

MTNI: What was your role in all this - what were you doing at the time career-wise, and what made it seem like a good/timely idea?

VL: I had turned from pure human translation to MT consultancy in 1978. Since 1977 I had also been active in the International Federation of Translators (FIT), of which I was then a vice-president. FIT was a non-governmental organisation with the coveted Grade A consultative status at the UN Economic and Social Commission and at Unesco. Grade A status meant that those bodies had to consult FIT on any issues relevant to translators. FIT was respectable, and people listened to it. In addition, FIT held a world congress every three years, covering every imaginable aspect of translation, and quite a few more. It was in fact one of these congresses, in 1977, which whetted my appetite for machine translation. Unlike FIT, the MT field was deeply unrespectable. You can always tell the pioneers by the arrows in their backs. I reckoned MT needed that respectability, that voice, and the fertilising influence of an all-embracing forum.

MTNI: Who were some of the other supporters for such an association in Europe?

VL: At the very beginning it was a question of hammering out some sort of proposal to float to possible supporters later. So I was working alone. I had to drop out in late 1989 for a few years, while Muriel and Professor Nagao steamed ahead with admirable efficiency. Both, in their different ways, had been a joy to work with. By April 1993 I was became active in the association again and was voted onto the EAMT Committee for a couple of years.

The European Commission believed in pure users' associations, not associations combining users with developers. So Loll Rolling, who had done so much for MT in Europe, did not support the founding of the EAMT.

MTNI: Who was the first EAMT president?

VL: Professor Nagao, happily, asked Maggie to establish the EAMT. She was its first president.

□

Conference Announcements

EAMT-7/CLAW-4 Joint Conference

Dublin City University, Ireland
May 15-17, 2003

Call For Papers

The theme of the 2003 joint conference of the European Association for Machine Translation (EAMT) and the Controlled Language Applications Workshop (CLAW) is "controlled translation". Papers addressing this theme will be featured on the second day of the conference, with the first day devoted to general papers on machine translation (MT), and the final day dedicated to other papers focusing more on controlled language issues.

Over the years, there have been many conferences on MT, involving rule-based approaches, statistical and example-based approaches, hybrid and multi-engine approaches as well as those limited to particular sublanguage domains. In addition, there has been an increased level of interest in controlled languages, culminating in the series of workshops on controlled language applications (CLAW). These have given impetus to both monolingual and multilingual guidelines and applications for using controlled language for many different languages.

Controlled languages are subsets of natural languages whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity. Traditionally, controlled languages fall into two major categories: those that improve readability for human readers, particularly non-native speakers, and those that improve computational processing of the text. It is often claimed that machine-oriented controlled

language should be of particular benefit when it comes to the use of translation tools (including machine translation, translation memory, multilingual terminology tools etc.).

Experience has shown that high quality MT systems can be designed for specialized domains (e.g. METEO). However, the area of controlled translation has remained relatively unaddressed. This is rather strange given its undoubted importance. Such examples that exist use rule-based MT (RBMT) systems to translate controlled language documentation, e.g. Caterpillar's CTE and CMU's KANT system, and General Motors CASL and LantMark, etc. However, fine-tuning general systems designed for use with unrestricted texts to derive specific, restricted applications is complex and expensive.

There are several examples of using Translation Memory (TM) tools in a controlled language workflow, yet these have been primarily for combining TM and MT tools. Very few attempts have been made where Example-based MT (EBMT) systems have been designed specifically for controlled language applications and use. This is even harder to fathom: using traditional RBMT systems leads to the well-known `knowledge acquisition bottleneck', which can be overcome by using corpus-based MT technology. Furthermore, the quality of EBMT (and Translation Memory) systems depends on the quality of the reference translations in the system database; the more these are controlled, the better the expected quality of translation output by the system.

Conference Goals

The aim of this conference is to provide a forum in which the problems accompanying controlled language translation may be outlined, possible solutions proposed, and in general to bring together developers, implementers, researchers and end-users from the publications, authoring, translation and localization fields to discuss how ideas from both the authoring and translation camps might be integrated in this common area.

Categories for Submission

Controlled translation: What is controlled translation; RBMT and controlled translation; TM/EBMT and controlled translation; Influence and interplay of controlled language upon both source-language parsing and target-language generation in an MT system; Role of the lexicon in controlled translation; Can we expect better controlled translations from a hybrid approach? Or from a multi-engine approach? Towards a Roadmap for controlled translation - the way ahead?

Machine Translation: MT for the Web; Practical MT systems; Methodologies for MT; Speech and dialogue translation; Text and speech corpora for MT and knowledge extraction from corpora; MT evaluation techniques and evaluation results; MT postediting.

Controlled Language: Examples of controlled languages: their definition, by whom, and intended usage; Consequences for technical authors and implications for Natural Language Processing; Practical experiences of teaching and using controlled languages; Application of controlled languages in speech systems.

System Demonstration: Developers should outline the design of their system and provide sufficient details to allow the evaluation of its validity, quality, and relevance to controlled language.

Invited Speakers

We are pleased to announce that invited speakers for the conference will include Steven Krauwer, University of Utrecht and Coordinator of ELSNET, and Lou Cremers, Océ Technologies.

The conference co-chairs are John Hutchins <WJHutchins@compuserve.com> on behalf of the EAMT, and Arendse Bernth <arendse@us.ibm.com>, on behalf of CLAW. For details see www.eamt.org/eamt-claw03/ □

EAMT Important Dates	
Abstract deadline	January 10, 2003
Reviews due	February 14, 2003
Notification to authors	February 28, 2003
Final papers due	March 31, 2003

MT Summit IX

New Orleans, Louisiana, USA
September 23-28, 2003

Call for Papers

The ninth Machine Translation Summit will take place in New Orleans, the birthplace of jazz. As in previous Summits, the ninth will provide a forum for everyone interested in using computers to help with language translation: developers, researchers, users, students, people who love languages. The program will be packed with invited talks, research presentations, demonstrations, panels, and an exhibition fair that showcases established companies, side-by-side with new MT startups.

A lively social agenda will include a reception and a surprise banquet that promises a very enjoyable evening. The conference hotel, the Fairmont New Orleans, offers a stunning environment for a conference and is within walking distance from the French Quarter. Other accommodations are within two blocks.

Call for Papers

MT Summit IX will feature a comprehensive programme that will include research papers, reports on users' experiences, discussions of policy issues, invited talks, panels, exhibits, tutorials, and workshops. We define machine translation in the broadest possible sense, to include not just fully automatic MT but tools for translation support and multilingual text processing as well.

We invite all those with an interest in translation automation-researchers, developers, translation service providers, users, or managers-to participate in the conference.

MT Summit IX hereby invites original submissions on all aspects machine and machine-aided translation. The submissions must be in English and fall into one of three categories:

• research (or theoretical) papers: maximum length 8 pages;

• user's studies (including manager's experiences): maximum length 8 pages;

•system presentations (with optional demos): maximum length 4 pages.

Proposals for Panels and Special Sessions

Proposals are also invited for panel and/or special sessions on issues of general importance to machine and machine-aided translation, which should be the subject of public debate at MT Summit IX. Please send your proposals to the Program Chair and include a description of the session theme, a justification of its importance, and the names of some suggested speakers or panelists.

Organizing committee

Conference Chair: Eduard Hovy, USC/ ISI <hovy@isi.edu>

Program Chair: Elliott Macklovich, University of Montreal <macklovi@iro.umontreal.ca>

Local Arrangements Chair: Flo Reeder, Mitre <freeder@mitre.org>

Exhibit and Tutorial Co-chairs: Laurie Gerber <gerbl@pacbell.net> and Keith Miller <keith@mitre.org>

Announcements will be posted on the conference website: www.mt-summit.org

Summit IX Important Dates

Submission deadline	May 11, 2003
Notification to authors	June 30, 2003
Final papers due	July 31, 2003

Client Side News

Globalization and Technology ROI Expo

Aspen, Colorado
February 20-21, 2003

This new event series is by and for the users and buyers of globalization and localization technology and services. Participation by vendor organizations is severely limited in order to foster an atmosphere of open exchange among localization professionals responsible for purchasing, managing or developing international products. CSN is the only client-focused organization empowering clients to drive industry solutions.

See: www.csnevents.com

LISA Global Strategies Summit

Foster City, California
March 3-6, 2003

Understanding Customer Requirements: The Localization Industry Standards Association's 45th international conference will address: How are clients succeeding in today's global environment? What is their ROI? What role do Open Standards-based systems play? The result will be 4 days of dedicated focus on the needs of clients in today's internet world, with special emphasis on how the most successful companies are making global business a priority while integrating business processes across the company and with key partners, suppliers and customers.

See: www.lisa.org/events/2003usa

HLT-NAACL 2003

Edmonton, Canada
May 27-June 1, 2003

HLT-NAACL 2003 combines the HLT (Human Language Technology) and NAACL (North American Chapter of the Association for Computational Linguistics) conference series. The conference especially encourages submissions that discuss synergistic combinations of language technologies, and new approaches to and uses of unsupervised and lightly supervised learning techniques. **Deadline for late-breaking short papers and posters: March 7, 2003.**

See: www.hlt-naacl03.org

41st ACL

Sapporo, Japan—July 7-12, 2003

Conference organizers: Junichi Tsujii, Univ. of Tokyo, (general chair); Erhard Hinrichs, Univ. of Tuebingen, and Dan Roth, Univ. of Illinois (pgm. co-chairs); and Kenji Araki, Hokkaido Univ. (local organization chair). **Paper registration deadline: Feb. 21, 2003, submission deadline Feb. 26.**

See: www.ec-inc.co.jp/ACL2003/

Book Review:

Empirical Methods for Exploiting Parallel Texts

I. Dan Melamed

MIT Press, 2000

Reviewed by Graham Russel
RALI, DIRO
Universite de Montreal, Canada
<russell@iro.umontreal.ca>

Parallel texts (i.e. text-pairs comprising an original and its translation) are packed with implicit information reflecting their translators' expertise. For those of us concerned with mechanical simulation of translators, therefore, such texts provide a potentially rich source of data, one whose exploitation has been a central theme of research for some fifteen years. The most newsworthy application has been statistical MT, beginning with the work of Peter Brown and his colleagues at IBM in the 1980s and continuing with the Verbmobil project in Germany, for example. Applications of data extracted from parallel texts go far beyond MT proper, however: they extend to the fields of bilingual terminology and lexicography, localization, cross-linguistic information retrieval and extraction, and aids for human translators in the form of indexed archives, specialized editors or consistency checkers. The only problem is how to make the valuable implicit information accessible. Melamed's book, a revision of his 1998 University of Pennsylvania PhD thesis, presents a unified collection of techniques for doing just this. It is organized in three sections, of which the first describes the fundamental technology developed by the author, the second discusses what it means for two words to co-occur in the source and target texts, and the third applies these techniques to several tasks connected with translation modeling and bilingual lexicon construction.

The first problem is how to identify which passages within the text are mutual translations; this is complicated by the fact that in general translation preserves neither the number nor the order of textual units. One standard approach is known as `alignment' and involves regarding each of the two texts as a sequence of segments (usually sentences). Pairs of subsequences of segments are then grouped into blocks or regions, each having the property that none of the target text lying outside a given block translates the source-text content of that block (and, in so far as the alignment has been carried out according to symmetrical criteria, vice versa).

For practical reasons, the alignment approach embodies a number of simplifying assumptions. In particular, inversion can be handled only crudely, by placing all of the displaced material within a single, possibly very large, alignment block. Melamed's preferred approach, the so-called bitext map, is not subject to this drawback. The basic idea is simple, and can be visualized in terms of a two-dimensional graph whose horizontal and vertical axes represent positions of tokens (words, punctuation, etc.) in the source and target texts respectively. Each point in the graph thus represents a pair of tokens, the first from the source and the second from the target. If we now mark all pairs that are potential mutual translations, we should observe a clustering of points in areas of the graph that correspond to translationally equivalent areas of the texts -- typically concentrated along a diagonal from bottom left (the start of the texts) to top right. Somewhere within this region lies a line that best characterizes the true translation relation.

This account begs three questions: how to identify the tokens in the source and target texts, how to recognize potential mutual translations, and how to find the correct translation path. The first of these is in fact a more ambitious and demanding process than that normally thought of as tokenization, since certain complex words must be given special treatment so that e.g. the German compound `Kindergarten' is analyzed not only as a unit but also as the pair `Kinder' and `Garten'. Candidate translations are proposed by a matching predicate specific to the pair of languages in question. This might involve some combination of a bilingual lexicon and a simpler notion of `cognatehood', based on string similarity.

To solve the searching problem, Melamed proposes an algorithm named the Smooth Injective Map Recognizer (SIMR -- `smooth' because it aims to produce a smooth path and `injective' because it assumes that no source or target token can occur in more than one valid translation-pair). It performs a constrained search within an area surrounding the main diagonal, subject to parameters including length of a recognized subchain and maximum permitted local deviation from the angle of the diagonal. The algorithm is evaluated with respect to a variety of text-pairs and performs well. Its complexity characteristics are referred to only briefly and informally, however.

The third chapter shows how a conventional alignment may be derived from the bitext map produced by SIMR. Given Melamed's mention of the limitations of the alignment format, it would have been interesting to see how some more general correspondence scheme, allowing crossing dependencies, could be generated. The results given are superior to those achieved by some more standard alignment programs; this effect may in part be due to sentence segmentation errors in the test data, to which SIMR is likely to be less sensitive than aligners which base their decisions directly on segmentation. Of course, even if this is the case, the consequent robustness of Melamed's approach is a strong point in its favor.

Chapter 4 closes the first section by describing the application of the bitext map to the problem of detecting omissions in translated texts. Again, the basic idea is a simple one: a passage in the source text that has not been translated will be reflected by a horizontal discontinuity in the broadly diagonal path representing true points of translational equivalence. Complications arise from the fact that non-literal translations and errors in the bitext map itself also generate such discontinuities, and methods are proposed for dealing with these.

Chapter 5 contains an important discussion of token co-occurrence in the context of the bitext-map approach presented earlier. Here, a pair of words may be considered to co-occur if the point representing them lies within a given distance of the translation path produce by SIMR. Different methods of counting multiple co-occurrences are also given, and related to assumptions

about typical word translation patterns (one-to-one, one-to-many, etc.) More empirical work is needed in order to predict which of these will be best suited to a particular set of circumstances. This is followed by a chapter dealing with manual annotation of texts to indicate translation relations between words. A sample data set is available from the author's web site (currently http://www.cs.nyu.edu/~melamed/) for researchers who wish to evaluate their own techniques.

The third section begins with a chapter dealing with the central question of estimating the parameters of a statistical word-to-word translation model, namely the probability for each source-target word pair (x,y) that x and y realize the same concept. Three methods of increasing sophistication are covered, all falling within the same general framework. The basic method is an iterative procedure which estimates the probability by assigning each word pair a score reflecting the likelihood that the words are translations. This score is used to guide a linking stage, in which x and y are considered to be linked if no other pair involving x or y has received a higher score. The translation probability is derived by normalizing the link count, and the scores are then reassigned to take into account the current estimated translation probability.

The second method incorporates a "noise model": additional parameters are used to encode the probability that a word-pair is linked given that they are (or are not) mutual translations. The rationale for this step is that a source-language word will tend not to be translated in different ways within a single text, and conversely that a given target-language word will tend not to be the translation of more than one source-language word. While it is easy to think of cases where this is not true (French prose norms being more resistant to repetition than English, for example), it seems a reasonable general assumption. Clearly, however, some caution is needed in applying the notion "single text" to the case of large collections.

The third method further specializes the additional parameters in order to model the behavior of different token classes. The experiments reported here made use of such classes as function words, content word s, several varieties of punctuation (end-of-sentence, end-of-phrase, quotes and parentheses) and other symbols.

The chapter continues with an extended discussion of the performance of the various models, comparing them with the well-known Model 1 of Brown et al.; those incorporating additional parameters are generally superior to either Melamed's basic method or Model 1. The treatment of evaluation here and elsewhere in the book is unusually thorough. Some useful final remarks are made on the use of these probabilistic models in constructing lexicons for MT systems.

In Chapter 8, Melamed turns to a question that has been touched on obliquely several times in the preceding text: how to account for the fact that standard typographically-based notions of "word" (a string of characters delimited by spaces or punctuation, say) do not always yield useful translation units. "Non-compositional compounds" (NCCs) are complex expressions whose translation is not a simple function of their components: French "poste de radio" translates as "radio set" rather than "radio post", for example. Melamed detects these by comparing a plain word-to-word translation model with one in which the candidate expression is treated as a unit; if the latter is better able to predict some parallel corpus then the candidate is accepted as a NCC. The situation is complicated by the need to identify many NCCs and the fact that a text contains many more potential NCCs than single words. To avoid creating a new translation model for each candidate when the vast majority will be rejected, Melamed employs a predictive heuristic to select the more promising cases, and tests many non-overlapping candidates (i.e. expressions having no word in common) in parallel. Once again, while the algorithm is stated to be efficient, its complexity properties are not studied.

The final chapter (apart from a brief "summary and future prospects") refines the assumption made earlier that words tend to translate consistently within a text, Melamed notes that, since different senses of a word will typically receive different translations, it is senses rather than words which an adequate translation model must take into account. He presents a method for determining which sense distinctions are relevant in a given text, based on the presence of cue words. Among the examples given is English `close', which has translations including `fermer' and `étroit'. Each of these is assigned to a different class using contextual cues: the `fermer' sense is associated with the `close down' context, `étroit' with `close consultation', and so on. The sense clustering algorithm has the advantage that no predefined inventory of senses is required; the number of senses identified for a word is adapted to the text being analyzed. How far this restricts the portability of a translation model is not discussed, but valid cues would clearly tend to recur in a variety of texts. The technique is evaluated by comparing the accuracy of translation models with and without sense clustering; a modest improvement is found for the clustered model.

Melamed's book is a slim volume, with a generous provision of graphs and diagrams, but no padding. The writing is clear and the production of high quality. Its mathematical content is accessible to readers with a basic knowledge of probability (a suitable refresher can be found in the recent textbook by Manning and Schütze, which also contains relevant material on alignment and probabilistic translation models). The algorithms are presented in sufficient detail for a competent programmer to reproduce them, but are not spelled out explicitly in pseudocode.

The book betrays its academic origins in several ways. There are brief but comprehensive surveys of relevant previous work, serious attention is paid to evaluation, and no pretence is made of having settled all of the questions raised; there are many signposts to future work, and any graduate student in need of ideas for a research project will find plenty here. Finally, it is worth noting that Melamed's web site also has links to a large number of computational tools that will allow readers of his book to make a start in implementing the techniques described therein.

References

P. Brown et. al. *A Statistical Approach to Machine Translation*, Computational Linguistics, Vol. 16, No. 2, pp. 79 85 (1990).

C. Manning and H. Schütze "Foundations of Statistical Natural Language Processing", MIT Press, 1999.

Muriel Vasconcellos

...continued from page 13

association for some time, but her main interest was MT users, whereas I was excited about the idea of bringing together researchers, commercial developers, and users in a common endeavor. As we talked, the latter concept gained ground. Once the idea took shape, Veronica introduced it on the floor of the Forum. Momentum gathered, and Professor Nagao brought up the subject again at the banquet. By the close of the meeting Professor Nagao made certain that we all went home with the commitment to create an international body that would be open to all people interested in machine translation.

The first MT Summit, also organized by Professor Nagao, had taken place two years earlier on September 16-18, 1987, in Hakone, Japan, and that meeting created the structure that was ultimately to be blended with the proposal for IAMT. European MT-ites, led by Christian Rohrer and Tom Schneider, followed up on the first Summit by convening MT Summit II in Munich on August 16-18, 1989, and thus a tradition was born.

At Munich Professor Nagao showed Veronica and me a handwritten outline he had prepared (I believe he wrote it there), which spelled out the objective, proposed activities, and organizational structure of IAMT. The newsletter was a key component, and there was provision for a broad range of educational activities, including conferences and training programs. From the outset, the basic idea was to bring together all three interest groups, but the regional approach did not emerge until later.

Back home in D.C., I put the outline into prose and fleshed it out with details, then faxed it to Professor Nagao and Veronica for their comments. We continued to exchange our thoughts by fax over the next few months. As the text – and our thoughts – became more refined, we began to realize the complications that would be involved in managing an association with members all over the world, especially when it came to mailing out the newsletter and collecting dues. So Professor Nagao came up with the regional concept, and we incorporated it. His solution took care of all the loose ends that had been bothering us. Now we were ready to present our plan to the global MT community.

The occasion was the Symposium on Japanese-to-English Machine Translation, convened by the U.S. National Research Council and held at the august headquarters of the National Academy of Sciences in Washington, D.C., on December 7, 1989. As the participants took their seats, they each found a copy of the IAMT prospectus in their chair, and both Professor Nagao and I took advantage of our places on the program to promote our proposal. Tom Seal, president of ALPS, reinforced our efforts by speaking to the subject as well.

MT Summit III was scheduled to be held in Washington, D.C., on July 1-4, 1991. It became our goal to have all the pieces in place so that IAMT and the three regional associations would come into being at the Summit. The timing was crucial, as it would link IAMT to the Summit series. And we certainly couldn't wait two more years for the next Summit.

I should back up to mention that at a break during the December 1989 Symposium, one of the participants took me aside and pointed out that nothing would happen until we got a *lawyer involved* in the initiative to work out the details of making it happen. That casual comment marked the beginning of my awareness. There was so much to do – so much to know! Of course, we couldn't afford a lawyer – we had no money at all, because we didn't exist. Since we couldn't hire a lawyer, it fell to me to spend much of my spare time over the next two years researching and figuring out what needed to be done at each step of the way. The trickiest part was dividing the responsibilities between the international and the regional associations, testing out the formula with myriad if-then scenarios to make sure everything would work. I relied on faxes to elicit input from interested people all over the world. All that we do so easily today by e-mail, I did by fax, working evenings and week-ends from home and footing the bills myself. It became an obsession for me.

This process went on through the course of 1990. Professor Nagao used the intervening time to establish the Japan Association for Machine Translation, thus ensuring structure and leadership for the Asia-Pacific region. For the Americas, I had a tripartite team of committed individuals who had been receiving and responding to my faxes over the past months. Veronica had health problems and was unable to carry the ball for Europe. At one point that year I had a visit from Loll Rolling of the European Commission, at which he offered to give us a budget if we would form an association for MT users only, presumably to be based in Luxembourg – he was not interested in including commercial developers or academic researchers. I turned him down and instead asked Maghi King if she would be willing to organize the European component along the lines we had been proposing. She graciously accepted. So Maghi and I proceeded to prepare our respective agendas for EAMT and AMTA, and at Summit III we convened our organizational meetings. Maghi's group chose her to head EAMT, and I was genuinely surprised when I was elected president of AMTA. I had been totally focused on getting that far and assumed that my work was over. I was looking forward to taking a rest!

In keeping with our plan, the regional officers automatically became officers on the international Council. In addition to the three of us, starting with Professor Nagao as the first IAMT president, Roberta Merchant was elected treasurer, Scott Bennett was chosen to be secretary, and John Hutchins agreed to serve as editor of the newsletter (he later created the name *MT News International).* The last session of Summit III was turned into an open town meeting at which we reported the steps that had been taken, introduced the officers for IAMT and the three regions, and encouraged everyone's involvement.

But that was only the beginning. We still weren't "legit"! I enlisted a volunteer attorney, Robert Carswell, who made it legal for us to set up bank accounts for AMTA and IAMT so that we could start collecting money. However, not only did we need money in order to operate, we needed to be incorporated, and in order to incorporate we needed to have bylaws. Based on a process of consensus, and with considerable input from Scott Bennett, I hammered out bylaws for both AMTA and IAMT, as well as the verbiage needed for the Articles of Incorporation. My patient mentor was Deanna Hammond, who had re-

cently stepped down as president of the American Translators Association and was extremely knowledgeable about the ins and outs of bylaws and the incorporation process. She gave unstintingly of her time, and her help was invaluable.

On Deanna's advice, we decided to seek 501(c)(3) nonprofit status for both AMTA and IAMT so that we could receive grants and take advantage of steep discounts on postage. Roberta prepared the application. At first we were turned down, but she went back to the drawing-board and persevered. It was necessary to introduce some changes in the bylaws in order for us to qualify (*we are an educational nonprofit, and the modifications needed were to explicitly outline the educational component, and omit "promotion" of MT, which was seen as too commercial),* and this was done. Between those applications and the many others that we had to file, Roberta must have filled out hundreds of pages of forms. She was also a very stern treasurer and authorized nothing but bare-bones spending. We owe much of the success of the two associations to Roberta's dedication in the early years and her careful husbanding of our meager resources.

I was still hoping to take a rest when one day Bill Fry, a good friend of IAMT, remarked "Until you start having activities in between the Summits, your association will be nothing but a paper tiger." So it was decided to hold our first IAMT workshop, "MT Evaluation: Basis for Future Directions." Thanks to our newly won 501(c)(3) status, we qualified to apply for a grant from the National Science Foundation. Sergei Nirenburg and I put together and signed the application, Roberta filled out reams of forms, and we got enough money to pay for some twenty speakers from all over the world. The meeting was held in San Diego on November 2-3, 1992 and was a huge success.

MTNI: Although a Charter for the IAMT was approved at a Council meeting (presumably at the Summit) in July 1991, there is a lot of correspondence and revision of the IAMT bylaws through July 1993 (the parties involved there seem to have been you, Veronica, Maghi King, and John Hutchins mainly on MTNI.), and of the Articles of Incorporation for AMTA through August 1994 (your correspondence shows that the parties were you, Deanna Hammond and Scott Bennett.) Were these revisions mainly to satisfy the requirements of incorporation as a non-profit?

MV: Yes, all this was done to satisfy the requirements of incorporation as a nonprofit organization.

MTNI: Do both IAMT and AMTA have the same 501(c)(3) status as educational non-profits? Or is IAMT different?

MV: Initially there was no need for for 501(c)(3) status for IAMT (since all "operations" were handled at the regional level), but, in order to enable us to receive grants, we voted to go for it at the meeting of the IAMT Council in 1995 (Kobe). As I recall, Roberta handled that one alone, based on our experience with AMTA. By that time I had moved to California.

MTNI: You once told me a story of having to first file as a corporation with bylaws (in DC), and then file for non-profit status, and having to repeat the process each time the IRS (right) quibbled with something in the bylaws. About how many times did this happen? Can you remind me of some of the issues that required the repeated revisions and reapplications?

MV: This happened only once. We filed and were rejected, so we had to re-phrase the purpose so that it did not include the word "promote," and we had to eliminate some items from the charter, and sprinkle the word "educational" all over the place. When Roberta presented our revised text to the IRS, she ended up speaking directly with the person there and continued to tweak the text until it was acceptable to them. Once that was done and we were approved, there was no need to repeat the process. What did happen, is that we were AUDITED, and we had to answer a lot of questions and clean up our act. The thrust of the questions was to satisfy the IRS that we were truly educational and to reassure them that we had broad representation of the MT community in our publications, including MTNI. They seemed to be especially interested in evidence showing that the officers were not an inbred "clique."

MTNI: Does our tax status preclude advertisements in MTNI?

MV: The short answer is ABSOLUTELY. There are a few teeny exceptions, but they become moot when we apply for non-profit status with the Post Office, which is another process. I worked with another nonprofit organization, and our lawyer was ADAMANT about not having any ads in our publications.

MTNI: Were other categories of tax-exempt or not-for-profit statuses considered?

MV: Yes. ATA (the American Translators Association), for example, has non-profit status under another section of the code, but they are subject to a number of restrictions and are prevented from receiving any grants -- which is the main reason why we applied at all. The status that is coveted is definitely 501(c)(3). ATA hired a lawyer to try to change their status, but they spent a lot of money and were ultimately rejected. Four types of organizations may be considered for 501 (c)(3) status: charitable, religious, educational, and another that's something like medical or health-related. □

GILT

...continued from page 3

The Professional Association for Localization

www.pal10n.net: PAL was founded in 2001 as a more affordable alternative to LISA, aimed at individuals rather than corporate members. PAL is more of a vehicle for sharing information and resources than a club. They don't hold regular meetings or conferences, but they do offer a wealth of handy information on their website. For example, a very long table at www.pal10n.net/en/TrainingSpreadsheet.html lists dozens of books, consultants, and events that offer training in globalization, internationalization, and localization.

The Globalization and Localization Association

www.gala-global.org: GALA was founded in April 2002 as forum just for vendors of software and services. The idea is to provide a forum where vendors can share information and ideas and network in an environment without the pressure to compete in front of clients.

Client Side News

www.clientsidenews.com: CSN was founded in 2002 with a magazine, informational website, and a new conference just for clients – the buyers of localization tools and services. The idea is to provide an environment where clients can network and learn from each other without sales pressure. □

Hovy on IAMT

...continued from page 11

helped educate the public and the Government about the feasibility of robotics. With a little thought, as language processing technology matures, we might be able to do the same for MT. Imagine: given a toolkit, build your own MT system in three days, and then evaluate it in a public trial…

MTNI: Recently there has been some discussion about forming some new regional associations. India seems motivated to start a new regional group. What is happening there?

EH: It is natural and necessary for more regional groups to form as people become active in the field. And naturally, they will want to have associations that focus on shared issues. However, financial, social, and technological differences create research alliances that often do not align with geopolitical boundaries. While North America and western Europe are somewhat distinct entities, Asia is not, and so the AAMT may eventually exist as an umbrella organization for several regional ones, including for example for India and environs, Japan, China, Korea, and the smaller countries of South-East Asia. Similarly, the Middle East and eastern Europe might form their own multinational associations.

MTNI: It sounds like you prefer subgroups to having new chapters of IAMT.

EH: It is not clear to me whether the IAMT can be of better service to a larger set of smaller associations or a smaller set of larger ones. To the extent that the problems of an Indian MT association are similar to those of a Southern African and South American one, IAMT can help communication between them. But since neither researchers nor MT companies pay attention to formal organizational structures when they seek exciting problems or new markets, I don't think IAMT can provide much additional service as a sort of United Nations. Regional associations can more easily adapt to local conditions (language families, funding, technology distribution, etc.).

Besides, adding many new chapters to IAMT would become complex. The IAMT does not have a significant infrastructure beyond enabling the coordination of the regional associations. It is not prepared to handle a proliferation of regional associations, whereas the regional associations are strong (and have paid administrative staffs).

MTNI: What do you see as significant recent progress in MT technology research?

EH: Concerning the substantive topics related to progress and developments in MT, I'm hoping we'll hear more about example-based and data-driven MT. Various people have recently pointed out that (at the single word or two-word level) anything that can be said has already been said, *and probably translated*, and is probably even available on the Web. The problem is finding these translated words and phrases, and then weaving them together again. This is of course simply the Translation Memory model, taken to the extreme. It offers a whole new set of opportunities and problems, most of which I believe are easier to handle than the complex issues of statistics-based MT. So I hope that we'll hear more about the potential of the Web, the growth of parallel and semi-parallel corpora, and their use for real MT applications. This is where we will find the best potential for bringing up new MT systems with small amounts of data, to deliver linguistically higher-quality output.

MTNI: What about commercial MT?

EH: After the near-total ingestion of MT by L&H, and their subsequent implosion, it is heartening to see the resurgence of new MT companies. I hope they will develop new markets as well as transfer the new statistical MT technology developed on the research side. I wish them the very best in the current economic climate.

MTNI: What can we look forward to at MT Summit IX?

EH: The next summit will be held in New Orleans in September 2003. New Orleans is a great place for a conference. It doesn't have quite the history of Santiago de Compostela (site of MT Summit VIII), but it has the American equivalent! The conference site will be right next to the historic French Quarter. New Orleans has great nightlife, it is the birthplace of jazz, there is a lot to do, and I think everyone will have a great time. □

TMI Conference Summary

...continued from page 7

the future of computer-to-computer communication. In a context in which multiple computers interact to solve problems, a consistent, specific ontology is needed. In future, he suggests, a shift to the use of natural language for this purpose seems like a reasonable direction, since natural language is used for human-machine interaction anyway. MT could then play the role of allowing computers with different "native languages" to communicate. In later discussion, participants expressed a concern that human languages are redundant, and so, not efficient. But it was pointed out that computer-computer languages also have a great deal of redundancy built in for robustness, and that natural language is also quite compact (because of ellipsis).

Hitoshi Iida (Sony CSL) also speculated on appropriate MT applications, listing MT products such as documents, speech translation, and remote instruction; MT combined with some processing as in summarization, and quick exchange; and MT embedded in other systems such as the web, information retrieval and chat. He suggested that MT experience in the use of interlingua could inform appropriate language formatting for cyberspace; intra-language applications could include polishing documents, sign-to-verbal translation and editing notations, and he even envisioned possibilities for inter-species communication between humans and dolphins.

The discussion after these presentations explored whether MT could ever replace foreign language learning, with a number of participants suggesting that in fact, MT could be used as a powerful phrase book to enable language learners to learn more effectively. In fact, simple MT systems now available can be used in language learning contexts much the way calculators have been introduced into math classes.

Making MT More Efficient

Other presentations focused on ways to reduce the time and energy invested in building MT systems. A recurring

observation in these papers was that these methods perform as well as current, more labor-intensive, methods.

Paul Davis and co-author Chris Brew suggested in "Stone Soup Translation," that instead of using transducers in statistical approaches to MT, linked automata be used. They explained that, though linked automata are much simpler, if supplemented with techniques for approximation and generalization, they can perform as well as current methods.

One major issue for Japanese machine translation is ellipsis resolution. Shigeko Nariyama suggested the use of information from three different levels of "Grammar for Ellipsis Resolution in Japanese." She demonstrated that an algorithm integrating grammatical information at the levels of sentence, predicate, and discourse devices can achieve a high level of accuracy in ellipsis resolution.

Teruko Mitamura and her colleagues Eric Nyberg, Enrique Torrejon, Dave Svoboda, Annelen Brunner, and Kathryn Baker are attempting to make MT work efficiently without the need for large aligned corpora or deep analysis. They also addressed the problem of "Pronominal Anaphora Resolution in the KANTOO Multilingual Machine Translation System," and used syntactic structure information from a full parse of a text to resolve anaphora rapidly while preserving the accuracy of the translation.

Simple MT systems now available can be used in language learning contexts much the way calculators have been introduced into math classes.

Dictionaries for MT also received attention in a number of papers. Timothy Baldwin and Francis Bond proposed "Alternation-based lexicon reconstruction" in order to derive a hierarchical dictionary from the more usual flat dictionaries of MT. They automatically extract and link alternations in the Goi-Taikei dictionary (Ikehara et al., 1997) to enhance maintainability and consistency. These links represent otherwise uncaptured lexical generalizations.

Mikel Forcada described another method, developed with (first author) Alicia Garrido-Alenda and Rafael Carrasco, for efficient maintenance of lexical resources in an MT system by use of the "Incremental Construction and Maintenance of Morphological Analysers based on Augmented Letter Transducers." They describe how to add and remove lexical entries easily using deterministic augmented letter transducers, while keeping the transducers minimal, and thus keeping dictionaries compact.

Further suggestions for the efficient expansion of dictionaries were made by Sanae Fujita, who described "A Method of Adding New Entries to a Valency Dictionary by Exploiting Existing Lexical Resources," work done with Francis Bond. Japanese words without a valency pattern in the dictionary, can be assigned a valency pattern based on the word's similarity in meaning to a word in the dictionary which does have a valency pattern. The increased valency coverage in a dictionary so enhanced can improve translation quality.

Jesse Pinkham and Martine Smets explored the possibility of achieving good "Machine Translation without a Bilingual Dictionary." They combined the use of automatically derived transfer patterns with the use of a lexicon consisting only of function words, and tested their system against a benchmark translation system. Their system's performance was comparable to the benchmark system, including some successful examples in which the bilingual dictionary of the benchmark system blocked the learning of good transfer patterns.

Word alignment in large bilingual corpora can also be problematic. Ralf Brown suggested a method to improve word alignment automatically by "Corpus-drive Splitting of Compound Words." Cognates in a bilingual corpus are scored for similarity, and boundaries for compound words are hypothesized. The use of a corpus thus split into its compound components can improve performance in an example-based MT system.

The highly inflected nature of the Polish language was exploited for the automatic construction of an MT and summarization lexicon in a paper by Barbara Gawronska, Björn Erlendsson, and Hanna Duczak. The authors used morphological forms associated with animacy as cues for "Extracting Semantic Classes and Morphosyntactic Features for English-Polish Machine Translation." They concluded that it is possible to partially automate lexical acquisition in highly inflected languages without sacrificing quality.

Hiroshi Kanayama addressed the problem of the translation of Japanese postpositions by proposing "An Iterative Algorithm for Translation Acquisition of Adpositions." His algorithm extracts VPN-tuples from a monolingual English corpus, and then automatically restricts that set to those which can be used as translations for Japanese verb phrases using the postposition de. The method allows refinement of the selection of prepositions in translation.

Taro Watanabe, Kenji Imamura and Eiichiro Sumita suggested ways to avoid some of the difficulties inherent in statistical approaches based on word-by-word aligned bilingual corpora by doing "Statistical Machine Translation Based on Hierarchical Phrase Alignment." By aligning corpora phrase-by-phrase, the authors were able to achieve improvement in translation, even when the parsing that formed the basis for the phrase alignment was not optimal.

Kenji Imamura, continuing the move away from word-by-word alignment, proposed the "Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT." Hierarchical phrase alignment extracts equivalent phrases from a bilingual corpora; the results can then be used in a pattern-based MT system. When the pattern extraction is manually "cleaned," the performance of such an automatic system is comparable to the performance of systems with manually constructed patterns.

Another approach to machine translation, namely, the derivation of translation rules from controlled sentences elicited from native informants, was discussed by Katharina Probst and Lori Levin in a paper delivered by Ralf Brown. In order to make this approach effective, it is desirable to automate it; however, the authors note that there are numerous "Challenges

Continued on next page ►

TMI Summary

...continued from previous page

in Automated Elicitation of a Controlled Bilingual Corpus." These range from human issues surrounding the use of informants to language issues involving the automation of elicitation of wide varieties of linguistic devices, to computational issues of alignment and adequate coverage for learning.

Addressing a similar problem in writing rule-based grammars, Alicia Tribble reported on work done with her colleagues Alon Lavie and Lori Levin on "Rapid Adaptive Development of Semantic Analysis Grammars." In mature MT systems, analysis and generation components may be fine-grained and quite complex. This leads to problems in adding new languages; grammar writers new to the system may not have the depth of understanding acquired by long experience with the system. These authors propose to alleviate such difficulties by isolating and automating tasks done by grammar writers and those done by native speaker informants.

As the use of bilingual, tagged corpora for machine translation becomes more prevalent, there is a greater need for accurately tagged corpora. Working with a Japanese-English corpus in which tense, aspect and modality are tagged, Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara explored the "Correction of Errors in a Modality Corpus Used for Machine Translation Using Machine-learning." By using estimated probabilities of tags in context, they could assign confidence values to those tags and identify those that may be in need of correction.

Setsuo Yamada, Kenji Imamura and Kazuhide Yamamoto advocated capitalizing on the efficiency of statistical approaches to MT and the accuracy of human rule writing by doing "Corpus-Assisted Expansion of Manual MT Knowledge." In their approach, source sentences are extracted from a corpus using a source pattern as the retrieval key, and translations are provided. Human rule writers judge whether the sentences match the pattern and whether the target phrases are correct or not. If necessary, modified target phrases are integrated into the system and the process repeated. This approach is a more efficient means for expanding MT resources than human rule writing alone.

Reflections on the Main Conference

A preponderance of papers at the conference were concerned with acquiring knowledge in MT systems, and some with advances in statistical MT. For the most part, the papers represented incremental improvements in current methods rather than qualitative changes in approach.

Hindsight allows an interesting observation. In the Roadmap Workshop that followed the main conference (see companion article), reference was made a number of times to the dissatisfaction users express with the quality of machine translation. Yet virtually none of the papers in the main conference appeared to have quality of translation as their primary concern. Instead, exploiting MT in web-based or other applications, and developing methods to make system-building more efficient were the prominent themes. In the latter case, performance of the system was subordinate to its efficiency. Only one paper in this group reported clear improvement in performance; four reported slight improvement; two reported performance levels comparable to previous systems (another type of "no improvement"); three simply gave percentages of coverage; and six gave no evaluation results.

Conferences provide a great opportunity to gauge the current trends in a field and to ask questions about priorities and future directions. The questions I was left with are these: Is improvement in translation quality so hard to achieve? Are we saving so much time and manpower by improving our methods for developing systems? If so, what are we doing with these savings?

References

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. Goi-Taikei—A Japanese Lexicon. Iwanami Shoten, Tokyo. Five volumes.

TMI Roadmap Workshop

...continued from page 7

between meaning (i.e., perceptions and experience) and structure (i.e., language). In order for automatic processing of human language to resolve the inevitable ambiguities, the system requires information, derived either from a deep analysis of the expression or from some external source. Since the latter approach involves the application of indeterminate amounts and types of world knowledge, Ikehara restricted the goal of semantic analysis to the former.

"In the Roadmap Workshop, reference was made a number of times to the dissatisfaction users express with the quality of machine translation. Yet virtually none of the papers in the main conference appeared to have quality of translation as their primary concern."

Ikehara then outlined the design concepts for a semantic dictionary. The first problem is the granularity of the analysis. Because the system he is working with is a Japanese to English translation system, the correct level of granularity can be considered to be the level at which the system gets the English correct. Furthermore, since the computer is unbiased with respect to symbol set, English symbols for semantic concepts can be used.

The current status of semantic dictionaries is represented by the Goi-Taikei (Ikehara et al., 1997). It consists of a semantic attribute ontology, a word dictionary, and a structure dictionary (including valency patterns). However, much more work needs to be done to expand the coverage of the dictionary to include complex sentences and noun phrases. Ikehara estimates that 100,000 semantic patterns could provide adequate coverage for the Japanese language, and points out that the existence of this kind of dictionary would change the face of natural language processing. A project entitled Analogical Machine Translation Method based on Typological Semantic Patterns, supported by the

Japan Science and Technology Corporation, is currently underway to address this goal.

Roadmap Workshop Papers

The three presentations which followed all addressed very different ways to realize progress in MT.

Francis Bond examined the nine steps recommended for humans to follow in making good translations (Nida, 1964) and drew analogies to the MT field in "Toward a Science of Machine Translation." His resultant proposal is Multi-Pass MT, in which texts are parsed several times for a variety of purposes, and models are retrained on the fly. In addition, MPMT would integrate rules and stochastic rankings, taking advantage of the benefits of both. Now that processing power is increasing, the improvement of translation by multiple processing of texts is becoming more feasible.

Mikel Forcada proposed "Using Multilingual Content on the Web to Build Fast Finite-state Direct Translation Systems." The existence on the web of large bitext (bilingual text) resources could be exploited to supplement constructed bitexts now used in MT, especially for languages for which such constructed bitexts are not available. The web can be mined for bitexts and appropriate methods used to align them at sentential and subsentential levels. The resulting resources could be used with translation-memory-like technology to provide translation units. Well-developed finite-state transducer techniques could alleviate the problem of intensive translation unit look-up. Careful integration of current technologies could result in a web-based translation service that would act as a fast pre-translator, and provide translated browsing to users and a translation unit lookup for professional translators. Users could also contribute bitexts to the server for integration.

Nigel Ward took current MT to task for showcasing technology without considering the real needs of possible users of MT in his talk entitled "Machine Translation in the Mobile and Wearable Age." He advocates trying to help people do what they already do, but better, by setting as the goal for MT "overcoming language barriers," a broader goal than translation. He described a wearable system which allows users to pre-select utterances needed for finding directions or ordering food in a foreign language, from a built-in menu. This is an easier and more efficient method than speech or keyboard input. The system then "speaks" the utterances to an interactant. The system is "heads-up," so that the user can supplement the translation it provides with facial and hand gestures. It uses an eyeglass display and pre-recorded speech. In an experimental situation, the system was found to be faster than using a phrase book, though heavy. At present it uses no MT at all; Ward suggests that focusing on further usage scenarios will suggest appropriate insertion points for MT in future.

Summary of Presentations

Ikehara's and Bond's presentations were concerned with two different approaches to MT, but both outlined ambitious programs of research for improving the quality of MT. Forcada addressed the problem of the acquisition of enough resources to make MT possible not only for well-supported languages, but also for "minority" languages. Ward 's presentation did not actually deal with MT at all; instead he focused on a topic that received some further attention in later discussions, namely the needs of MT users.

"Where do we stand? MT Summit and TMI Report"

Steven Krauwer then returned to the podium to give a (European) assessment of the current status of MT. He noted two trends in particular:

• larger projects with shared goals, tools, resources and technologies that tended to be inflexible and involve little or no basic research, and

• the integration of NLP into other interfaces, systems and services, resulting in reduced responsibility for NLP, but greater problems of integration.

At recent MT conferences, he noted the following tendencies:

• from "one solution for all situations," to "one solution for each situation"

• statistical-based approaches becoming acquisition tools instead of the sole approach to MT

• moving from strings to hierarchies in order to make better use of linguistic knowledge

• more attention to "minority" languages, i.e., languages that haven't had a lot of attention before

Krauwer noted interesting work at theses conferences, but no real revolutions, and lots of islands of research with few connections. While intellectual property is a huge issue, it remains a problem unsolvable at such workshops as these.

There are generally three different types of beneficiaries from a roadmap workshop such as this one. Researchers usually want to find "the truth," whatever the cost. Technology and service providers want to sell their products, and users want MT done cheaper/faster/better/more. This workshop needs to speak to all three points of view.

Interactive Session

To this end, participants were asked to involve themselves in the following exercise: Imagine you had $100 million and five years. What would you do to move the field forward? The answer could only be one challenge with a well-defined output, and would be categorized as a "pure" research, R&D, or user goal.

As a preliminary to this exercise, participants listed goals that had been met in the last five years:

• Keeping html tags in translated text from web

• Automatic, free translation of web pages

• Commercially available speech-to-speech translation

• Use of speech recognition for captioning television programs

The table on the next page lists the suggestions made.

Participants also tried to identify topics which relied on one another or were connected in some way; the categories that are cross-indexed were so identified.

One other suggestion that did not readily lend itself to categorization was made, namely, not to spend a large sum of money on one large topic, but rather, to distribute it amongst a number of

TMI Roadmap Workshop

...continued from previous page

smaller topics. This spreads support for a variety of efforts.

Invited Speaker

Harold Somers (UMIST) presented an invited talk asking "What are we celebrating today?" He gave a humorous overview of the progress made in MT over the last fifty years, marking the advent of a number of different MT projects such as Systran and Eurotra. He implied that choices about languages addressed in MT systems are often politically motivated and that progress is sometimes measured simply in the number of languages accommodated by a system. While he left us all chuckling, he also left us thinking about the directions MT has taken and will take in the near future.

Closing Discussion

The day closed with continued discussion about such topics. One participant noted current users' frustration with lack of good quality MT products. Bond likened current MT products to early models of cars: not entirely refined and workable, but available, nonetheless. And, like changes/improvements in automotive design over the years, we can expect incremental improvements in MT products in the future.

Somers picked up this point, noting that MT products have only recently become widely available. Users were at first amazed, then disgusted, then pragmatic, and finally, those who have stuck with using MT products have become wise in how to make the best use of such products. One useful activity now would be for the MT community to take a hard look at such users and see what they actually do with MT products. Have they just gotten used to and put up with the inadequacies of the products? Or will they pressure the industry to improve the products in particular ways?

Another problem broached by a participant was the fact that developers can't make their money back on R&D focussed on "minority" languages. Yet, there is widespread agreement that we need to add new languages to the MT repertoire quickly. It was suggested that components from MT for one language could be usable in systems for other languages—a kind of horizontal portability. How to do that effectively, in terms of both research and development, is an issue that is in urgent need of addressing.

More information about ELSNET is available at www.elsnet.org. Another summary of this workshop is available at some point from Steven Krauwer's web server: www-sk.let.uu.nl.

References

Nida, Eugene. 1964. Toward a Science of Translating. E. J. Brill, The Netherlands.

Researchers	R&D	Users
Develop a theory of translation	Exploit existing mark-ups for MT	Language plug-ins
Develop a semantic theory	Use language learning materials to develop MT	#"Consumer Guide" for MT, i.e., large-scale evaluation
*Develop knowledge database	%Give a boost to "minority" languages	Glasses to translate text-to-text
Use translation memories in rule-based systems	Develop an n-text, aligned and massively annotated corpus	Extended, controlled language menu for small screen interactions
Develop new linguistic theories for MT implementation (so, cross-train linguists and computer scientists) [this could be cross-indexed with many topics]	Develop mobile phone plug-ins for language translation	#Better user models to predict needs and evaluate uses
*&Achieve robust (audio-visual) speech recognition meaning		Web translation engines for cross-linguistic IR
%Set up open resources for all to use		&Automatic stenography with MT
Create a theory of cross-lingual communication aids		
Make a formal description of spoken language/communication		

Table: Answers by Roadmap Workshop participants to the question: What would you do to move the field of MT forward if you had $100 million and five years? Special marks indicate cross-indexing between categories.

AMTA Board Update

AMTA officer elections were announced at the AMTA general membership meeting at AMTA-2002 in Tiburon, California on October 10, 2002. The executive board, consisting of president, vice president, secretary and treasurer, all were re-elected to a second two-year term. Three new directors were elected for two year terms. In addition to the election of officers, two referenda were approved in the election. First, the executive board role of "Councilor" was made permanent, and is to be occupied by the immediate past president, currently Eduard Hovy. Second, the role of Webmaster was added to the executive board. The Webmaster is an appointed role, rather than elected (like the MTNI editor and regional editors). The first AMTA Webmaster is Jin Yang.

AMTA Executive Board			
President	Elliott Macklovitch	RALI - U. of Montreal	macklovi@iro.umontreal.ca
Vice President	Laurie Gerber	LTB	gerbl@pacbell.net
Councilor	Eduard Hovy	USC/ISI	hovy@isi.edu
Secretary	Karen Spalink	Ericsson	
Treasurer	Stephen Helmreich	NMSU/CRL	shelmrei@nmsu.edu
AMTA Regional Newsletter Ed.	David Clements		dclemen1@san.rr.com
Webmaster	Jin Yang	SYSTRAN Software Inc.	webmaster@amtaweb.org
Directors			
Developer (until 2003)	Michael C. McCord	IBM Research	mcmccord@us.ibm.com
Researcher (until 2003)	Violetta Cavalli-Sforza	Computer Science Dept., SFSU	vcs@sfsu.edu
User (until 2003)	Christine Kamprath	Caterpillar Inc. Prod. Support	Kamprath_Christine_K@cat.com
Developer (until 2004)	Michael S. Blekhman	Lingvistica	ling98@canada.com
Researcher (until 2004)	Steve Richardson	Microsoft Research	steveri@microsoft.com
User (until 2004)	Tag Young Moon	CAS	tmoon@cas.org
Administration	Debbie Becker	2168 Greenskeeper Ct. Reston VA 20191 tel: 703-716-0912 fax: 703-716-0912	AMTAInfo@att.net

AAMT Board Update

In March, we published a list of board members for the AAMT, and committee and workshop chairs (Asian-Pacific Association for Machine Translation). In August, some adjustments were made to the board. The table at right reflects the current AAMT board of directors

AAMT Board		
President	TSUJII, Jun-ichi	Professor, University of Tokyo
Vice Presidents	KOTANI, Taizo KAWAMURA, Shinsuke	President, Inter Group Corporation Vice President, Toshiba Corporation
Directors	NAGAO, Makoto TANAKA , Hozumi ISHIZAKI, Shun YOKOYAMA, Shoichi IIDA, Hitoshi ISAHARA, Hitoshi CHOI, Key-Sun SORNLERTLAMVANICH, Virach MAEYAMA, Junji WASHIZUKA, Isamu GOTO, Satoshi KUWAHARA, Hiromi KUSHIKI, Yoshiaki USHIO, Shintaro TANAKA, Tatsuo	President, Kyoto University Prof., Tokyo Institute of Technology Prof., Keio University Prof., Yamagata University Prof., Tokyo University of Technology Group Leader, Communications Lab. Prof. KAIST, Seoul, Korea Division Director, NECTEC, Thailand Senior Vice President, Fujitsu Ltd. Sr. Exec. VP, Sharp Corporation VP, NEC Corporation Corporate Officer, Hitachi, Ltd. Board Member, Matsushita Electric Co. Adviser, Oki Electric Industry Co., Ltd. President, JEITA
Auditors	TODA, Motoyoshi KATSUTA, Mihoko	Executive Vice President, JEITA CEO, Toin Corporation

MT News International

Subscription Order Form
for non-members of IAMT Associations

*Subscription to **MTNI** is a benefit of membership in any of the three regional IAMT Associations. Non-members may also subscribe. This form should be sent to the appropriate region together with a remittance in the currency specified. The fee covers a one-year airmail subscription (individual or institutional) for three issues, starting in the spring of the current year.*

I/we wish to receive a one-year subscription to MT News International:

Name: _____

Organization: _____

Address: _____

City: _____State: _____ Post al code: _____

For individuals or institutions located in the Asia-Pacific region, please return this form with payment of ¥4,000 (bank draft or international money order) to:

Association for Machine Translation in the Americas
PMB 300
1201 Pennsylvania Avenue, N.W. , Suite 300
Washington, DC 2004 USA

For individuals or institutions located in Europe, the Middle East, or Africa, please return this form with payment of Sw.fr. 70.00 (check or money order) to:

Association for Machine Translation in the Americas (AMTA)
PMB 300
1201 Pennsylvania Avenue, N.W. , Suite 300
Washington, DC 2004 USA

For individuals or institutions located in North, South, or Central America, please return this form with payment of US$ 75.00(check, money order, or credit card) to:

Association for Machine Translation in the Americas (AMTA)
PMB 300
1201 Pennsylvania Avenue, N.W. , Suite 300
Washington, DC 2004 USA

Publications Order Form

Please return this form with payment or credit card information, to:
International Association for Machine Translation (IAMT c/o AMTA)
PMB 300, 1201 Pennsylvania Avenue, N.W. , Suite 300
Washington, DC 2004 USA

Please send the items marked at the right to:

Name: _____

Organization: _____

Address: _____

City: _____State: _____ Post al code: _____

E-mail: _____ Fax: _____

	Title	Price (in U.S. dollars)[1] Member[2]	Non-member
☐	***Compendium of Translation Software*** (on-line version)	FREE	$20.00
☐	***Proceedings of MT Summit VI***	$40.00	$60.00
☐	***Proceedings of AMTA-96***	$40.00	$60.00
☐	***Proceedings of AMTA-94***	$40.00	$60.00
☐	***Proceedings of Workshop on MT Evaluation*** (1992)	$55.00	$55.00

[1] Prices include shipping and handling.

The proceedings of AMTA-98 and AMTA-2000 appeared as #1529 and #1934 in the Springer series Lecture Notes in Artificial Intelligence. To order, contact the publisher at www.springer.de.

Method of Payment

☐ Check or M.O. ☐ Visa ☐ MasterCard ☐ American Express

Card number _____ Exp. Date: ___/___

Asia-Pacific Association for Machine Translation

CORPORATE / INDIVIDUAL APPLICATION FORM

Please fill out the appropriate form in both your native language (NAT) and English (ENG) and send it to the address below.

CORPORATE

Corporate name: Seal:

NAT: _____

ENG: _____

Capital: _____ Date established : _____ # employees: ____

Name of company president:: Seal:

NAT: _____

ENG: _____

Name of person responsible for this application: Seal:

NAT: _____

ENG: _____

His/her office/department:

NAT: _____

ENG: _____

Mailing address:

NAT: _____

_____

ENG: _____

_____

Tel: _____ Ext.: _____ Fax:_____

E-mail: _____

Business category:

- ☐ Government agency ☐ Manufacturer
- ☐ Service industry ☐ Translation business
- ☐ Other (please specify):_____

All corporate application forms should include a company prospectus.

INDIVIDUAL

Name: Seal:

NAT: _____

ENG: _____

Mailing address:

NAT: _____

_____

ENG: _____

_____

Tel: _____ Fax: _____

E-mail: _____

Occupation: _____

Company name: _____

Company address: _____

- ☐ Translator ☐ Other: _____
- ☐ University/Institution/Researcher

Specialty: _____

Individual Membership Fees

Initiation fee for individual members	¥ 1,000
Annual dues for individual members	¥ 5,000
Total payment	**¥ 6,000**

The foregoing amount will be paid by ___/___/___ , or within one month from the date this form is mailed.

Corporate Membership Fees

Initiation fee for corporate members
(1 unit = ¥10,000; minimum 1 unit) ___ units = ¥ _____

Annual dues for corporate members
(1 unit = ¥50,000) ___ units = ¥ _____

Please check appropriate box below:

- ☐ MT system developer (minimum 10 units)
- ☐ Other, capital over ¥10 million (minimum 2 units)
- ☐ Other, capital up to ¥10 million (minimum 1 unit)

Total payment (initiation fee + annual dues) **¥** _____

The foregoing amount will be paid by ___/___/___ , or within one month from the date this form is mailed.

Method of Payment

- ☐ Wire transfer to:

  Bank of Tokyo,
  Mitsubishi Bank / Roppongi Branch
  Tokyo, Japan

  A/C No. 1091515 (ordinary account)

  For: Asia-Pacific Association for Machine Translation (AAMT)
- ☐ International postal money order in Japanese yen to AAMT at ▸

Appplicants are responsible for all bank charges. Please retain your copy of bank draft or money order as proof of payment.

Today's date: __/__/__

Please send this form to:

Asia-Pacific Association for Machine Translation (AAMT)
c/o Japan Electronics and Information Technology Industries Association (JEITA)
Mitsui Kaijo Bekkan Building, 3F
3-11, Kanda-Surugadai, Chiyoda-ku
Tokyo 101-0062, Japan
Fax: +81 (03) 3518-6472

Association for Machine Translation in the Americas

MEMBERSHIP APPLICATION / RENEWAL FORM

Type of member and membership fee per calendar year:

- ☐ *Individual* — *US$ 60*
- ☐ *Institutional (nonprofit)* — *US$ 200*
  Representative: _____
- ☐ *Corporate* — *US$ 400*
  Representative: _____

Please return this form, together with your payment or credit card information, to:

Association for Machine Translation in the Americas
PMB 300
1201 Pennsylvania Avenue, N.W. , Suite 300
Washington, DC 2004 USA

Last name(s): _____ *First name(s):* _____ *Title:* _____

Address: _____

_____

Home tel.: _____ *Work tel.:* _____ *Fax:* _____

E-mail: _____ *Website:* _____

Affiliation: _____

Professional associations: _____

Area of specialization:

☐ *MT User* ☐ *MT Developer* ☐ *MT Researcher* ☐ *Translator* ☐ *Manager* ☐ *Other* _____

Method of Payment

- ☐ *Check enclosed*
- ☐ *Credit card*

Type of credit card: ☐ *Visa* ☐ *MasterCard* ☐ *American Express*

Card number _____ *Exp. Date:* ___/___

European Association for Machine Translation

APPLICATION FOR MEMBERSHIP

Please return this form, together with your payment or credit card information, to:

EAMT Secretariat, c/o TIM / ISSCO
Université de Genève
École de Traduction et d'Interprétation
40, blvd du Pont-d'Arve
CH-1211 Geneva 4, Switzerland

Type of member and membership fee per calendar year:

☐ *Individual*

SFr 50

☐ *Non-profit-making institution*

SFr 175

Last name(s): _____ *First name(s):* _____ *Title:* _____

Address: _____

_____

Home tel.: _____ *Work tel.:* _____ *Fax:* _____

E-mail: _____ *Website:* _____

Method of Payment

☐ *Cheque payable to EAMT, enclosed*

☐ *Banker's draft (copy enclosed) to account no. 351.091.40L*
Union Bank of Switzerland
Bahnhofstrasse 45
CH-8021 Zürich, Switzerland
Please note: All bank charges must

Type of credit card: ☐ *Visa* ☐ *Eurocard*

Card number

_____