Issue No. 33 (vol. 11, no. 3)
Autumn 2003
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 Announcement*** *(page 1)*
- ***Extensive Conference Reports*** *(starts page 6)*
- ***A Chat with Steve Richardson*** *(starts page 8)*
- ***MT Users' Desiderata*** *(starts page 19)*
- ***Upstart Data-Driven MT Companies, part 2*** *(starts page 20)*

Spotlight on the News

MT Summit IX

Fairmont Hotel, New Orleans, USA—September 23-28, 2003
Latest Information on Events, Sessions and Venue

The International Association for Machine Translation (IAMT) will hold the ninth Machine Translation Summit in New Orleans on September 23-28, 2003. MT Summit IX will provide a forum for everyone interested in using computers to help with language translation: developers, researchers, users, students, and 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

The Summit features an exciting technical program. There will be almost 60 papers on a variety of topics, from MT evaluation to user studies to studies of translation algorithms and data.

Pre-Conference Tutorials: Tuesday, September 23

Six tutorials will be presented, in morning and afternoon sessions.

Morning Session

Tutorial 1: **Computer Assisted Business Process Management for the Language Industry**
Adriane Rinsche, Language Technology Centre, Ltd.

The Language Technology Centre has developed a tool called LTC Organiser that has revolutionized business process management of many translation/localization companies as well as translation and localization departments in

Continued on page 2 ►

IAMT

MT News International

Issue No. 33 (vol. 11, no. 3)

Autumn 2003

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

MT Summit IX

...continued from previous page

multinational companies and international organizations. The application is designed to reduce the cost of managing documentation and translation projects, including DTP and printing processes, decrease the time to market, and maximize the benefits derived from human and technical resources. The tutorial will describe the most important aspects of the integrated solution.

Tutorial 2: **Finite-state Language Processing and Its Applications to MT**
Shuly Wintner, Department of Computer Science, University of Haifa

Finite-state technology is becoming an invaluable tool for various levels of language processing. The tutorial will provide an introduction to the technology and its many applications in natural language processing. Aimed at linguists and computer scientists alike, it starts with the very basics of finite-state devices and regular expressions and concludes with a sketch of how to design and implement a large-scale project.

Tutorial 3: **Thanks for the Memories: Translation Memory Demo**
Hans Fenstermacher, ArchiText Inc.

This tutorial focuses on demonstration of translation memory technology, and is designed to give participants a detailed overview of the Translation Memory (TM) process.

The session will explain in detail:

- what TM is
- what the different types of TM are
- how content is processed in TM
- what the advantages and disadvantages of TM are
- how TM generates cost savings, including a price quote simulation
- some tips and tricks for working with TM from the content developer's perspective
- how to prepare the source content better before it goes into TM

This demo is intended for a broad audience, including:

- Project/Product managers
- Product developers (engineers, web developers, etc.)
- Content developers (writers, editors, etc.)
- Anyone involved in localization or translation

Afternoon Session

Tutorial 4: **Information Architecture for Controlled Authoring and Translation**
Joerg Schuetz, Institute for Applied Information Sciences (IAI)

In the last decade, the idea of controlled authoring (CA) is discussed within several industries at various levels, but not often implemented, and if so then with quite different success stages. Common to most of the implemented CA applications is that they are technology add-ons to already existing processes. The technologies that are employed range from natural language processing utilities to machine translation.

In this tutorial, we analyze the situation and introduce the steps that are necessary for building a success story in controlled authoring and translation. Many examples are taken from our customers who use our product CLAT (Controlled Language Authoring Technology). The focus of the tutorial is not on our product, it is on the industry-level stages of definition, realization and deployment of the concept of CA.

Tutorial 5: **Introduction to Statistical Machine Translation**
Kevin Knight and Phillip Koehn, University of Southern California, Information Sciences Institute

Accurate translation requires a great deal of knowledge about the usage and meaning of words, the structure of phrases, the meaning of sentences, and which real-life situations are plausible.

Continued on page 16 ►

Products and Announcements for the MT Community

AMTA Selects New Focal Point and Headquarters

The AMTA has a new Focal Point, Priscilla Rasmussen. Priscilla has been ACL's (the Association for Computational Linguistics) Business Manager for many years. In that capacity, she's been present as an organizer at all of ACL's annual conferences.

As AMTA Focal Point, Priscilla's duties include maintaining contact with the membership, answering your questions, responding to inquiries from the public, and helping organize conferences and other ongoing activities. Here are Priscilla's coordinates:

Priscilla Rasmussen
AMTA Focal Point
3 Landmark Center
East Stroudsburg, PA 18301
phone: +1-570-476-8006
fax: +1-570-476-0860
email: ***focalpoint@amtaweb.org***

❑

MTNI Volunteers Strengthen Editorial Team

MTNI recently benefited from two additions to the editorial team: David Clements, the AMTA regional editor, has taken over desktop publishing and copy editing of each issue, in addition to contributing articles. Karin Spalink of Sony Ericsson has taken on some of the news reporting and editing.

Thanks very much to both! Readers have David to thank for bringing MTNI back on a more regular schedule.

❑

GTI Launches TranslateTV

[adapted from press release]

In January 2003, Global Translation, Inc. (GTI), a Columbus, Ohio-based provider of real-time translation solutions, announced the launch of *TranslateTV*, a first-of-its-kind product that enables television broadcasters, cable operators, advertisers, and program distributors to offer programming in up to eight languages. Based upon proprietary technology, *TranslateTV* provides instant live language translation of television closed-captions. WBNS-TV (Channel 10) in Columbus became the first station in the nation to roll-out the product by offering English-to-Spanish translation to its growing Hispanic audience. *TranslateTV* is supported technologically through advanced patent pending software, and sophisticated communications technology. GTI supplies, installs, and maintains a translation server at the broadcast site that instantly decodes closed captions, translates the text stream, and then re-encodes the results in unused caption fields such as CC2, CC3 and CC4. Although GTI's translation service is performed primarily by software, a professional team of lexicographers and linguistics engineers continually updates and customizes the translation software to address specific program material.

In addition to Spanish, *TranslateTV* is currently available in seven other languages including: Chinese, Korean, Japanese, French, German, Italian, and Portuguese.

In April, *TranslateTV* and VITAC, the national leader in captioning and multi-language subtitling; were awarded three NAB "Pick of Show" awards for their unique television technology solution shown the National Association of Broadcasters convention. VITAC and *TranslateTV's* exclusive partnership enables live, instantaneous translation of English to Spanish captions. As the Hispanic marketplace grows rapidly (now 13% of U.S. population), this state-of-the-art technology, allows television broadcasters and producers to reach this critical and burgeoning market for less time and money than is traditionally required to create Spanish captioning. *TranslateTV's* proprietary, patent-pending technology consists of rules-based linguistic and unique caption-processing software. The software can provide real-time translations through a range of mediums including broadcast TV, cable, satellite and Internet streaming video.

TranslateTV's CTO, Mary Flanagan, is known to many in the MT community as a hardworking and persistent pioneer in MT deployments. ❑

Cross Language to Focus on Business MT Implementation

By Karen Spalink

Jaap van der Meer, former CEO Alpnet (now part of SDL), together with Heidi Depraetere and Mike McMahon, have established a new company. Cross Language is a consulting business with a focus on the business viability of machine translation implementation.

Cross Language offers QuickScan, a case-based rapid assessment of the enterprise environment and the ROI calculation for machine translation. It covers all areas of translation applicability from Intranet and Extranet, to production and complete enterprise solutions.

As independent consultants they are basing their recommendations not on a particular system but on the needs of

Continued on page 13 ►

Phraselator — Not Quite MT but Very Successful

Engineers from outside the MT community proper have tried a variety of ways around the problem of full text translation to solve a variety of problems. While perfectionists continue to shoot for fully automatic MT, others have embraced a bundle of simplifying assumptions, and built some remarkably useful tools. One of these is the "Phraselator" and its relatives, a translation system that relies on reusable large chunks of text – phrases – that can be reused in a variety of contexts. Not quite example-based MT, the phrases are chosen to suit the translation needs in very specialized contexts – when apprehending criminal suspects, or POWs. Others have used the same idea to create multilingual communication capability in multiplayer computer games.

Phraselator: One-Way Speech to Speech communication

In 1995 or so, Dragon Systems together with some collaborators came up with the idea for what is now known as the "Phraselator."

Using the Phraselator, a source sentence might be composed of three chunks as follows: "Show me where" + "the soldiers" + "buried the mines." The long source language segments are much easier to recognize when spoken than an unpredictable sequence of words. The limited number of phrases also simplifies the translation problem. The phrases and combinations were deliberately designed to be answered with gestures or actions, avoiding the problem of trying to recognize and translate responses from untrained users. The original system was designed for the conflict of that time, to handle Serbo-Croatian. The project has continued with U.S. Government funding, under the leadership of one of the designers, Ace Sarich. Here are some details from the website of Marine Acoustics, and VoxTec, companies Mr. Sarich leads, and which develop, evolve, and market the Phraselator.

The DARPA One-way development is sponsored by Defense Advanced Research Projects Agency (DARPA) Information Technology Office (ITO) Human Language Systems. Originally developed as text-to-voice phrase translator by the Naval Operational Medical Institute (NOMI), speech recognition was later added to to enable voice-to-voice one-way communication. Designated the Multilingual Interview System (MIS), the system was deployed to Bosnia 1997.

In support of Maritime Intercept Operations (MIO), the DARPA One-way was deployed to the Arabian Gulf July 1998. The MIO specific DARPA One-way system consists of commercial-off-the-shelf (COTS) hardware and voice recognition and translation software. The language module consists of approximately five hundred phrases and words translated into the four most common languages used in the Gulf: Arabic, Farsi, Hindi and Urdu. A two pound, 5x8 inch Toshiba Libretto 100CT with a 166 MHz processor runs the software. A sensitive noise-canceling microphone is used for speech input, and a small speaker is used for translation output.

MAI (Marine Acoustics) was awarded a DARPA SBIR grant January 2001 to develop a handheld PTS called the Phraselator. After the 9/11 attack, the development was accelerated, and about 500 Phraselators were built and delivered to military units in support of operation Enduring Freedom. VoxTec is a new company, organized to commercialize and market the technology developed by Marine Acoustics, a high technology contractor to DARPA.

Focused Subject – Quick to Build

A typical module with 500 custom phrases can be built in less than two weeks. Force Protection (FP) and Medic module developed for the U. S. Army Pacific. Both FP and Medic are translated into Chinese (Mandarin), Korean, Cambodian, Thai, Russian and Tagalog. FP also translated into Polish, Dari (Eastern Farsi), Pashtu, Arabic (Gulf) and Urdu. Kosovo Refugee and Kosovo Medic modules were developed to support the processing of Albanian speaking refugees coming out of Kosovo. 800 phrases translated into English and Albanian. The 400 phrase Basic Medic module was developed to support the Fleet Battle Experiment/ Urban Warrior exercise in California, March 1999. It is translated into German, French, Spanish, Arabic, Korean, Turkish, and Albanian. Maritime Intercept Operations (MIO) module was used for boarding operations in the Arabian Gulf the Summer of 1998. Over 400 phrases translated into Arabic, Farsi, Urdu, and Hindi. Displaced Persons module used in the Strong Angel Exercise June 2000. 553 phrases are translated into Tagalog, Japanese, Spanish, Egyptian Arabic, Korean, Swahili, and French. A 350-phrase system translated into Mandarin and Spanish for evaluation by US Coast Guard personnel involved in Immigration and Naturalization Services boardings. The Medical Language Translator module prepared by the Naval Operational Medicine Institute. Over 3000 medical phrases and words translated into English, Chinese, Korean, Portuguese, Thai, Bengali, Singhalese, Arabic, French, Indonesian, Russian, Spanish, Japanese, and Persian-Farsi. Debriefing Aid module prepared by the Naval Operational Medicine Institute. Approximately 5000 intelligence debriefing phrases and words translated into English, Persian-Farsi, Singhalese, Haitian-Creole, Russian, Serbo-Croatian, Cambodian, Spanish, French, Korean, and Egyptian Arabic. Over 1500 useful tourist phrases and words translated into English and Spanish for a tourist module.

Multiplayer Online Games

A very similar approach was taken by Japanese game designer Yuji Naka at Sega. In an effort to allow the increasingly multinational online gaming community to communicate during play, he came up with the "word select system" that is designed to let players converse and get to know each other. Users select common phrases and words to quickly compose sentences such as. "I like sailing." In addition, users have the option to use a selection of icons to represent their ideas.

For more information about the Phraselator and its relatives, see: ***www.voxtec.com****,* ***www.sarich.com/ translator****;* ***www.phraselator.com****.*

❑

Surprise Language Project Develops Hindi/English Translator

[adapted from press release]

In less than a month, during June 2003, researchers at USC's Information Sciences Institute (ISI) and collaborators nationwide built one of the world's best systems to translate Hindi text into English and query Hindi databases using English questions.

The effort was part of the "Surprise Language" project, a test of the computer science community's ability to quickly create translation tools for previously un-researched languages. The exercise was sponsored by the Defense Advance Research Project Agency (DARPA) and ended July 1.

"A month ago, we didn't even know what language we would be working on," explained Ulrich Germann, a computational linguist at ISI, part of the University of Southern California School of Engineering.

At 10:55 p.m. PDT on June 1, the manager for DARPA's TIDES (Translingual Information Detection, Extraction and Summarization) program fired the starting gun with an email: "Surprise Language is Hindi.... Good luck!"

Teams at 11 different sites across the US and one in the UK jumped into action. Twenty-nine days later, they can present an impressive array of information processing tools for Hindi.

"We succeeded in all aspects of the exercise," said Douglas W. Oard, an associate professor at the University of Maryland who is currently spending a sabbatical year at ISI. "A month ago, we had no information retrieval for Hindi, no machine translation, no named entity identification, no question answering. Now we have all of these."

In addition to USC/ISI, other participating institutions included the University of Maryland, College Park; the IBM Thomas J. Watson Research Laboratory, Carnegie-Mellon University; the University of California, Berkeley; New York University; the University of Massachusetts, Amherst; Johns Hopkins University; the University of Pennsylvania; the University of Sheffield (U.K.); the MITRE Corporation; BBN Technologies, and the Navy Space and Naval Warfare Systems Command (SPAWAR). Hindi was the first official language for the Surprise Language project. An earlier practice run in March, 2003 worked on Cebuano, a Philippine language.

❑

PC Magazine Italy Recommends LogoMedia

[adapted from press release]

LogoMedia was named "Recommended Internet Translation Service" by PC Magazine Italy. The recommendation appeared in an article titled "Poliglotti con il Web", in the April 2003 issue. The article reviewed and compared a number of online translation systems.

LogoMedia, which is based in Belmont, MA, USA, provides online translation services employing four different interfaces. Interface application is determined by the length and use of text to be translated.

TransIt is best suited for short texts. The translations can be automatically copied to applications like instant messaging, for example.

LogoTrans is geared towards longer texts.

Translation Mirror automatically translates the active window, updating the translation as you make changes in the active window, or translating a webpage as you browse.

FileTrans automatically translates entire files or folders consisting of any number of files.

The purpose-specific interfaces together with the large number of languages were some of the criteria that counted towards LogoMedia's high ranking. The company is expanding the set of languages offered to include Arabic, Turkish and Persian.

LogoMedia sells its services on a subscription basis. The fee schedule is tied to the translation volume.

See: ***www.logomedia.net****; Email:* ***info@logomedia.net***

❑

PROMT Announces Translation Quality Evaluation Tool

[adapted from press release]

PROMT has announced the release of new translation quality evaluation tool named CORVET. Multifunctional capabilities of the new product are provided with the accumulated experience of the PROMT company in the field of machine translation technologies. Corvet performs a comparison of machine translation results with the translated text treated with manual editing (what is called "ideal translation," e.g., in TRADOS Translation Memory format). Working with PROMT system (for example, filling up the dictionary in an interactive mode) makes it possible to see how a skillful adjustment of the PROMT system can allow translation quality to quickly approach the ideal.

The program also allows users to compare the quality of variants of translations made by different people - translators, or different translation systems. Comparison of Translation Memory segments before and after they have been corrected by human translators can help estimate the volume of editing work with a TM database.

"Actively working with our professional users, we have realized the necessity to create a tool which would help those who already work with automated translation tools (PROMT and TRADOS), objectively to estimate the

Continued on page 13 ►

Conference Reports

LangTech 2002

September 2002
Berlin, Germany

LangTech 2002 was the first in a new conference series that is designed to bring together the business community and emerging language technologies.

Overview

LangTech 2002 was attended by some 330 representatives from over 30 countries and across five continents. The actual program featured presentations from over 70 companies from 20 nations. Most importantly, nearly two-thirds of LangTech attendees came from industry or commercial concerns. These demographics naturally led to a balanced and comprehensive account of issues, business models and future opportunities for the speech and language technologies sector across the globe.

LangTech Program Highlights

Professor Hans Uszkoreit, LangTech Programme Chair, opened the conference by pointing out that the key current challenge to the speech and language technology sector was not so much bringing research concepts to market but dealing with the depressed business climate. There is a fairly advanced capacity to absorb innovation on the demand side, and despite current pessimism, the market is set to rebound strongly.

Key Strategic Points

A "user centric" drive toward "natural" communication and interfaces is widely regarded as the way forward. The European Commission's Sixth Framework Programme (EC FP6) appears to be addressing this explicitly in its "multimodal" roadmap, and many of the company pitches at the event had this concept at the centre of their business model.

Many voice and multilinguality-based technologies are now mature. As more and more applications are reaching the market, this process is set to gather greater momentum. However, several groups called for more EC support for translation technology efforts.

A key catalyst for market penetration is visibility at the board level. Marketing of language technologies must incorporate a greater effort to reach corporate decision-makers. Business consultants may emerge as an important champion for this cause.

There are usually almost no successful generic solutions in language technology; solutions have to be customized to a specific company, sector, task etc.

Language technology currently represents around 2% of the value added to software products.

Keynotes

Bill Dolan (Head of Natural Language Processing at Microsoft) reminded the audience that deployable language technologies have been expected 'in 5 years time' right from the beginning of machine translation (MT) in the 1950s. Yet we still, have not developed a feasible commercial model for rolling out the technologies to the mass market. Whilst Natural Language Processing smarts are gradually being integrated into consumer software, Dolan stressed that current user interfaces are far too clumsy: going forward, computers must now adapt to users rather than the opposite model that has driven the market. Microsoft is deploying NLP in the form of behind-the-scenes grammar checkers, smart tags and other morphological analyzers in consumer software products. He also showed how high quality MT tools can learn "automatically" from available bilingual texts in a specific domain, claiming that a single general purpose MT solution is probably not feasible. We are more likely to see thousands of specialized MT engines distributed over the web.

Professor Wolfgang Wahlster from the German research centre DFKI, focused on the use of language technologies in the mobile Internet environment, maintaining that the natural interface will indeed be multimodal. Mobile based UTMS and 3G devices will eventually provide access to all communication messages, information, entertainment and web based content, creating significant opportunities for the language technology sector. After introducing the revolutionary transportable interface concept, Smartkom, Professor Wahlster concluded by stressing that multimodal interfaces increase the robustness of user interaction and lead to more intuitive and efficient dialogues.

This theme was further supported by Giovanni Varile, from the IST Intelligent Interfaces & Surfaces Unit. Through the IST program, the EC has a vision of building a knowledge society for all, with user-focused interfaces in the foreground. This is evidenced in a research budget of over 3,600 million euro for Knowledge and Interface Technologies within the IST Framework. Mr. Varile identified the development of semantic-based and context-aware knowledge systems together with natural and adaptive multimodal interfaces as key EC objectives.

Guests at the LangTech evening reception on Thursday 26th September were addressed by Mr. Paul Hector, representing the Information Society Division of UNESCO. Mr. Hector stressed UNESCO's dedication to support measures that help preserve the right of individuals to participate in the information age through their native language. UNESCO is particularly concerned about the pace at which minority languages are disappearing. With this in mind, Mr. Hector outlined Initiative B@bel, ***www.unesco.or.kr/cyberlang/introframe.htm***, which seeks through policy, awareness raising, and the development of software applications and tools, to foster the development of information content and promote equitable access within a multilingual cyberspace.

Funding Innovation

The LangTech program featured a dedicated venture capital session, with a panel of four venture capitalists discussing some of their recent deals and their different approach to evaluating and selecting ventures for funding. Again it was stressed that the development of a natural user interface was a key interest area.

Marcus Jochim from Deutsche Telekom VC unit, T-Venture, pointed to the comparatively low level of investment intensity and investor confidence in the current market, but suggested strong future potential for voice based services. Jochim indicated some of the key success criteria for technology VC propositions as: quality of management, status of marketplace, “uniqueness” of technology, a flexible and open architecture, valuable business model, attractive expected ROI and potential synergies with the VC firm.

The ‘Elevator Pitch’ Competition

During the two days of LangTech, 23 companies from across the globe gave five minute “elevator pitch” presentations of their corporate project with a view to attracting venture capital interest, and of course, competing for the LangTech prize!

Voted by an international jury, prizes worth a total of 3,000 euro were awarded to the three best presentations. The jury - comprising technology and investment know-how - paid particular attention to the overall impact, degree of innovation/R&D capabilities of the organization, relevance of market scenario (size, development, competitors), company development potential (human resources), and appropriateness of investment required. With a large number of high-quality submissions, judging these entries proved to be challenging. But we are pleased to announce the following prize winners.

1st Prize (1,500 euro): Language and Computing, Belgium

2nd Prize (1,000 euro): Natural Speech Communication, Israel

3rd Prize (500 euro): The Language Technology Centre, UK

This post-conference report was compiled by the organizing committee: Bente Maegaard, Organisation Chair; Hans Uszkoreit, Programme Chair; Michael Huch, Local Chair:
organisation@lang-tech.org

❑

EAMT-CLAW 03

Dublin City University
Dublin, Ireland
May 2003

By Andy Way

The EAMT-CLAW 03 conference on Controlled Machine Translation combined the 8th European Association for Machine Translation Conference (EAMT) and the 4th Controlled Language Applications Conference (CLAW). EAMT-CLAW 2003 brought together two significant international events in the field of Translation Technology: the annual EAMT conference and the bi-annual CLAW Conference. Although both of these events deal with topics with significant overlap, no event had ever previously sought to unite researchers and practitioners from both fields.

Sponsors and Venue

DCU was considered to be a very suitable location for this since it is home to two major research centres working in translation technology, the National Centre for Language Technology (***www.computing.dcu.ie/research/nclt/***) and the Centre for Translation and Textual Studies (**webpages.dcu.ie/~studiest/content.html**), as well as undergraduate and postgraduate degrees in Applied Computational Linguistics and Translation Studies (run by the School of Computing and the School of Applied Language and Intercultural Studies).

Scheduled Papers

The three-day event scheduled papers dealing primarily with MT, Controlled Translation, and Controlled Language Technology. Two speakers were invited to address the topic of Controlled MT: Steven Krauwer, lecturer at Utrecht University and Chair of the Executive Board of ELSNET, the European Network of Excellence in Human Language Technologies, and Lou Cremers, translation technology manager at the Dutch firm Océ Technologies. There was also one panel session on the middle day where the panel was composed of leading industrial practitioners and academics.

The conference was officially opened on the Thursday, May 15, by Andy Way (DCU), and by John Hutchins (EAMT President) and Arendse Bernth (CLAW representative). The first keynote address was given by Lou Cremers on “Controlled Language in an Automated Localisation Environment.” Eight papers on the themes of the conference completed the day. The program on Friday the 16th comprised 8 more individual papers, followed by a panel session on Controlled Translation. This was chaired by Enrique Torrejon of IBM, Spain, and included panelists from academia and industry. A very lively session ensued, with many contributions from the floor. The final day on Saturday, May 17, included 7 more papers, as well as the second keynote address from Steven Krauwer, on “(Towards) a Roadmap for Controlled Translation.” Like the other invited talk, this was a very thought-provoking speech, and gave many participants a view of what the future may hold for our field.

Who Attended

There were 96 participants at the conference, from 16 different countries (13 in Europe, plus the US, Japan and Australia), and our 23 speakers were spread across 11 different nations. Excluding Ireland, most participants came from Europe (56%). 32% came from Ireland, with 7% coming from the US, and the remaining 5% from Japan and Australia. In the climate in which the conference took place (post-Iraqi war, SARS pandemic), we were not surprised that few Japanese and US residents were able to travel. Indeed, one conference speaker was prevented from coming as his US company had cancelled all non-essential travel. Given the serious nature of these world events, at one point we were quite worried that we might be unable to attract 50 participants, so the fact that nearly double this number attended was extremely satisfying.

Furthermore, 57% of the conference attendees were from industry, and 43% from academia. We consider this to be an

Continued on page 18 ►

Special Feature: Speaking of MT

A Chat with Steve Richardson of Microsoft

Microsoft began using its own hybrid machine translation system to translate technical support documents into Spanish in early April. Production versions of the Microsoft support knowledge base are located at ***support.microsoft.com****. Click on "international support" and choose "Spain" as your country. Choose the first option "****Busque artículos de ayuda en nuestra base de datos****", and then enter a Spanish query term, such as "equipo" in the window labeled "Buscar". The articles marked with a "gears" icon have been machine-translated. —ed.*

MTNI: How is the Microsoft deployment of MT for tech support going?

SR: It is still going strong with English->Spanish tech support. Customer service conducted a pilot survey – asking the question, "Did the article help answer your question?" Respondents answered on a scale of 1-9. Answers above 5 were taken as "yes". The results showed that for users of the original English documentation, 53% of respondents gave an answer of 5 or above, and 49.7% of users of the machine-translated Spanish gave an answer of 5 or above, so the MT output is perceived to be nearly as useful as the English.

MTNI: How was this deployed?

SR: There are about 140,000 articles in the knowledge base. There are a total of 50-60 million words, with hundreds of articles updated per week. Only 7,000 of those had ever been translated into Spanish. The 133,000 or so articles that hadn't been translated were machine translated and cached to be searchable in the target language. Once a week, all of the articles that have been updated or added are machine translated and the knowledge base is updated. The site tracks the frequency of access of the machine translated articles. Frequently accessed articles may get priority for human translation. The 7,000 articles were a core set of documents that were important to make available. Once the core set were identified, they were human translated over the last 2-3 years.

MTNI: What else is in the works?

SR: English->Japanese is in testing, with a pilot set for July. Japanese already had more articles human translated. With a bigger budget, they had about 30,000 articles in Japanese. But the Japanese audience is much tougher. Although similar levels of quality have been achieved according to metrics, the Japanese staff was not as satisfied.

English-French and English-German deployments are to be ready by August. These systems use a learned generation component, in contrast to Japanese and Spanish which have manually-developed generation components (all systems use the same English parser, etc.) There are papers out on this. The training data for each language pair is over 1M sentence pairs. We were able to collect TMs for many product areas to build and train the system.

MTNI: How and when did you get started in Machine Translation?

SR: I got involved in an MT project at Brigham Young University (BYU) as a student. Eldon Lytle was the main professor in linguistics, and he had a theory called Junction Grammar. He was Anti-Chomsky, and had developed his own linguistic theory to enable translation. I did my (Mormon) mission in Brazil and became fluent in Portuguese, and really enamored with the language. When I got back I was looking for a job and Professor Lytle's project was looking for a Portuguese Lexicographer. I joined it in the spring of 1975. It was an ongoing project with 20 people. They hired students to create dictionaries and transfer rules. The project was called the "BYU Interactive Translation System" (ITS) and it was focused on translating from English into Chinese, French, German, Portuguese and Spanish. Alan Melby was also a part of the project and has written about it. The idea was to have the user disambiguate the source text through fairly heavy interaction with the analysis phase, and then produce perfect output. As time went on, I became the Portuguese generation person, and was later in charge of Portuguese transfer, and after that helped with the transfer module in general, as well as overseeing other research projects. This was going on for the rest of my junior-senior year and masters program. (BS in Computer Science, Linguistics and Portuguese in 1977; MS in Linguistics with a computer science minor in 1980.) After completing my graduate coursework, I started full time, and became a university staff researcher. In 1980 the project lost funding. There were a couple of problems. The translators who would use the system felt a bit threatened, and they didn't like the translation quality for editing. Also, the system was running on IBM mainframes. When they looked at the numbers, the cost of the mainframes was much more than they could realistically save through increased productivity. When the project at BYU ended, almost everyone from the project went to ALPS, a company formed by Eldon Lytle in Provo.

ALPS continued on for many years, becoming a translation services company. They stopped their MT research later in the 1980s and became ALPNET. ALPS had come up with the notion of translation memory and created an early TM product as part of their MT system. It provided a whole environment for translation work. The TM component was called a "repetition file." The weakness of the

Continued on page 12 ►

An Interview with Gregor Thurmair of Comprendium

Gregor Thurmair is the determined captain of a team of developers who have worked with the Metal machine translation technology for 17 years. He has sailed the stormy seas of commercial language technology, through financial and organizational trials. MTNI spoke to him in July 2003. —ed.

Background: The Metal MT technology, whose development was originally funded by Siemens, was sold in 1995 to a new company called GMS, formed to continue development and commercialization of the technology. GMS licensed Metal translation systems to Langenscheidt, the famous dictionary publisher, who sold the system as retail software under the name "T1". GMS was subsequently acquired by L&H in 1997. Well before the L&H debacle, the Metal group split off again as SAIL Labs, a subsidiary of L&H. Following the L&H collapse. however, Sail was not able to make a go of it financially. Following a brief period of bankruptcy, the development group and technology were picked up by a content management company that reformed itself with the Sail group as a new company, Comprendium, headquartered in Munich Germany.

MTNI: What happened next?

GT: When L&H disappeared, our main sponsor disappeared. Sail labs continued to exist but went into insolvency in March 2002. The technology was bought by a company, and later called Comprendium. The current Comprendium company has a document management system with 2 main products: The Infostore system – the target customer is mid-sized companies. The other product area is enterprise content mgt systems – which can be enriched with multilingual technology. The core focus is on document management. There is a big customer base for Infostore. Comprendium has 2 or 3 big contracts, primarily with insurance companies. They wanted to combine multilinguality with their existing content management software. We resumed our activities in May or June 2002. The team was reduced – it had been over 120 people at L&H. Now it is 20-30 people after joining Comprendium.

MTNI: The same group has been part of many different organizations.

GT: Through the various business arrangements, there has been continuity in the technological development. There are 10 people who have been part of the group since the Siemens days.

Concerning organizational ups and downs, there have been mistakes in marketing. Siemens was famous for choosing the wrong hardware platform. Then we focused too long on the translation market. We moved to the PC market with Langenscheidt and set up workgroup solutions for small translation agencies.

MTNI: How do you see MT being used commercially now?

GT: Now we're focusing on the translation service centers within corporations. People wanted more control over terminology etc., and we came to offer server-based solutions with pattern matching, pre-editing, etc. The corporate line—focuses on network servers—mainly intranet.

The best applications of MT involve customization – tuning the terminology so that there is better acceptance of the output. Sometimes we get hired to do the customization, other times the client does the customization themselves. For example for CLS (Corporate Language Services in Switzerland) we added 60,000-70,000 terms in the financial domain. We are getting good feedback on that. Banking is a good application because of security issues. Many of the documents to be translated are so sensitive, they cannot just be sent out to translation services. .

Translation service departments may offer MT at a significant discount. At Daimler Chrysler, the language services department offers MT at a discount in their internal accounting. People use it for information gathering. The cost to maintain the system is divided up among departments. It is also used for communication. It is good for this type of work that would never be sent to human translators anyway; there is too much volume and not certain enough value.

MTNI: I understood that L&H wanted to combine the best parts of all of the MT technologies they had acquired, but that seemed very ambitious.

GT: Sail labs—which was part of L&H—had the goal to provide the next generation of technology for L&H. The idea was that it would be a new version of the Globalink technology. L&H Sponsored SAIL labs to do the technical development. The new version never materialized due to L&H's bankruptcy. However, we did do development work to be able to recombine and facilitate new language pair development.

L&H tried to buy revenue – to buy companies that had their own technology. They ended up with many different platforms and left the staff to consolidate them. They set up an architecture team, experts on each of the platforms. There were meetings to present the insides of each system to the rest of the group and decide on the best features. The Neocor system was determined to be spaghetti. Metal T1 was the best engineered. Globalink used tree-to-tree mapping. Apptek was unification-based, but not as well engineered. For the next step we had some ideas. We wanted to enrich T1 with components of the other systems. When L&H set up SAIL Labs, they kept Apptek and Globalink inside L&H proper. Then the groups started charging each other for time. This introduced organizational obstacles to evolving and unifying the systems. There were no more planning meetings or schedules for the release of new technology,

Continued on page 18 ►

Feature Article: From the Garage to the Attic

An Insider's View of Entrepreneurial MT

By David Clements

The Lernout & Hauspie saga first brought welcome attention, and then unwelcome attention, to the machine translation and language technology world. This installment of David Clements's story continues a first person account of the early years of MicroTac and Globalink that began in MTNI 33. David Clements, a veteran MT developer, is also the AMTA regional editor of MTNI. —ed.

Part 2: Nuts and Bolts

This product is copyrighted by MicroTac, but may be freely copied and shared…. The registered user's version is memory-resident and has a much more interesting HELP system, plus other goodies. Support ShareWare authors--keep us off the streets at night! (The nine people who registered in the first year were enough to make me live up to the promises I made in the original version!) [From MicroTac Foreign Language Assistant "read me" file, 1988]

The first phase of the new operation was to get the product reproduced and into sensible packaging. Also, the name took on its current familiar form: Spanish Assistant, French Assistant, German Assistant and Italian Assistant. Although people today refer to these as "Language Assistant," there was never a product called "Language Assistant," but rather the "Language Assistant Seriesä," consisting of the four products mentioned above. Versions up through 3.0 came in plastic boxes, with cheap paper covers. The first one was about the size of the original 5.25" floppy disks. One version even had credits on the back of the box, to all Tac's friends who helped out. Garet too, eliminated this: no lingering amateurism was allowed.

The Internet's Long Memory

Amazingly, a recent (January 2003) search through Google found links (though they all seemed dead) to the 1988 Spanish Language Assistant. One Web page has the following description: "A really useful program which helps with the conjugation of Spanish verbs in all fourteen tenses. It has an indexed verb-search feature in case you're not exactly sure of the Spanish spelling!" Another "ancient" link, from ISSCO, says, "The Language Assistant series … are integrated packages aimed primarily at people writing in foreign languages. They contain bilingual dictionaries licensed from Random House, conjugation generators, and a bidirectional batch translation mode for sentences. Over two-hundred thousand copies have been sold at US 79.00. Spanish, French, German, and Italian versions of Language Assistant are available…"

Conjugator and Grammar Help

The goal now was to move beyond the simple verb conjugator principle. What was it that would make a good "foreign language" product? There weren't many models in the software world to chose from, then. So, Tac and Garet looked at educational books, such as the Schaum's Outline Series. Within this series were basic grammars of the four languages of interest, with lessons and other help for students. With this in mind, we set about to add to the verb conjugator in two ways: add paradigms for nouns, adjectives, pronouns, etc., and add grammar "help topics." This became a writing project for me, as I assembled and wrote up the French help topics, and later helped out on German (and a year or so later, Italian). The paradigm tables, which we named "conj tables" or "decl tables," were modified for the new forms. Tac changed the software so that everything would display correctly, with help topics corresponding to each form popping up on demand.

With this progress, the company started to grow, and the "office" moved from Tac's home to a former doctor's office suite in Downtown San Diego. Ironically, the building was across the street from Planned Parenthood. Not only was Tac an ardent social activist in liberal causes, but he also remained a devout Catholic on moral issues. The first time I drove up to the new offices, I was struck by its paradoxical location.

This was still a time before I formally became an employee of MicroTac. Still finishing my dissertation, I consulted with the company to work on the grammar files and help topics. The downtown office started to grow, with help from some of Tac's other friends. Although my parents were always skeptical, one church friend of theirs who liked to do venture capital investments lent Tac a sizeable amount of money to fuel the company's growth. Still, MicroTac was an "S Corporation," and was essentially Tac's home business, grown large. He still drove his tiny white Ford Fiesta around San Diego and Tijuana, with the "No Nukes" sticker proudly affixed to the rear bumper.

Beyond the Shareware Model

The company's distribution was now breaking out of the "shareware" model and some employees were hired to do packaging, office work and phone sales. Technically, one of the last upgrades to the feature list was the addition of Random House bilingual dictionaries. Now the system could not only provide inflection information, but served as a powerful bilingual dictionary look-up tool. Unsure about your English word's best translation? Look it up and insert the correct inflected form directly into your document. Sales mounted steadily.

Continued growth led to another office move, this time from downtown to Cass Street in the San Diego coastal neighborhood of Pacific Beach. The Cass Street office was in a mid-70s style, three-story stucco building, with wide windows overlooking the busy streets of the beach community. In the distant west, on a good day, you could see the Pacific Ocean gleaming in the distance, as shiny and promising as the future that lay before the company.

"Thank you for calling MicroTac Software. How may I help you?" was the mantra new employees learned for answering the telephone. In September 1990, after graduating from UCSD, I formally joined the MicroTac family as an employee. I was already the fifth employee by this time. There were Tac, Garet, two students who did part-time phone and sales work, and a secretary/office manager. Already, Tac was becoming a shrewd and successful businessman. He drove some hard deals with his associates and employees.

"Yo Ser Hambrioso"

The feature that, to our surprise, really made the "Language Assistant Series" start to take off was something that was really an add-on. As part of the dictionary look-up program, there was an "auto-insert" function. Users could select this option to do an automatic "word-for-word replace" in their documents. It was great fun, and fascinating to customers, to watch Spanish Assistant's cursor magically dance through sentence after sentence, inserting literal translations for every word it encountered. For example, if French Assistant encountered the sentence, "I am happy," it would produce a translation of "Je être heureux." Even though idioms were present in the Random House dictionaries, the system couldn't handle them. "I am hungry" would translate as something akin to "Yo ser hambrioso."

"This Thing Is Great!"

This was very primitive, but on some level showed how ready the general public was for an automatic translation tool, by the early 1990s. Since I handled a lot of the incoming phone calls, I often talked to customers who would say, "This thing is great! If only you could translate idioms, and make verbs agree…." It was this customer demand, rather than any "grand design," that led MicroTac into an exciting, but perilous odyssey in the translation and MT industries.

❑

How Should We Define Data-Driven MT?

In April 2003, when we were working on the article surveying emerging data driven MT products, (the second installment of which appears in this issue), we found that few of the emerging or updated MT systems fit our previous notions of Statistial MT or Example-Based MT (the two familiar approaches to data driven machine translation). For example:

1) Many developers claim to include a statistical component. What does the MT community think is necessary to make such a claim? What is necessary for a system to claim to be "statistical MT"?

2) A number of developers describe their systems as example-based, but all of the examples or patterns are hand-built by linguists with apparently no automated learning component. In addition, patterns may be abstracted to phrase-structure rules. At some point, it starts to look a lot like rule-based MT.

3) Given the fuzziness in the two categories, can there be any useful definition of "hybrid" systems?

We submitted the following questions to the larger community via the MT-List *(see **www.eamt.org/mt-list.html** to subscribe, browse archives, etc.):* "Are there any iron-clad, authoritative definitions of Statistical Machine Translation (SMT), Example-based Machine Translation (EBMT) and/or Data-driven MT that would provide sufficient and necessary conditions for MT systems to claim membership in any of the above categories? How about working definitions?"

The concern at the time was that we had no clear means to evaluate vendor claims about what type of data-driven MT system they were offering. The question sparked off a lively online discussion. A compilation of the responses has been assembled, and is being edited for publication at a future date (possibly in the Machine Translation journal), however, we wanted to give a small sampling from the discussion that highlights the issues. In this sample, we include only a few of the efforts at broad, clear definitions. The full debate got quite heated for stakeholders in various data-driven MT paradigms, sometimes coming down to fine points of definitions, and even community lore, such as whether the IBM Candide project ever used any hand-built rules!

Michael Carl

EBMT and SMT are both different instantiations of Data-driven MT. While SMT systems are rooted in the IBM models, EBMT is based on analogical reasoning.

1) The fact that an MT system uses a statistical component does not make it a statistical one, in the same way a system does not become rule-based if it uses a (set of) rule(s).

2) In a paper by Davide Turcato and Fred Popowich What is Example-Based Machine Translation? ***(http://www.eamt.org/summitVIII/workshop-papers.html)*** the authors also find it difficult to distinguish MT paradigms by just looking at the resources used.

3) Hybrid systems integrate different (computational) paradigms where the author(s) want to stress that the paradigms are equally important in solving the task.

Bob Frederking

I use the notion of a "space" of MT systems. There are certain points in that space that are clearly "Example-based," "Statistical," and "Rule-based," but many (perhaps most) real systems fall into grey areas in between the obvious examples. It would make sense to talk about whether a system is "closer to" the pure EBMT or pure Statistical or pure Rule-based point in that space. I also make some definitions:

1) an EBMT system is one that uses the parallel corpus at run-time, as opposed to a model trained in advance from the corpus (whether or not it uses careful mathematical justifications)

2) a Statistical system is one that

Continued on page 17 ►

Richardson Interview

...continued from page 8

MT system was that it always required interaction on the front end.

Utah was quite a hotbed of activity around that time. In addition to the BYU project, Weidner also started up in late 1970s. Weidner is sometimes said to have come out of the BYU project, but that is not true, although some of its people did. Around the same time ECS (Executive Communication Systems, which produced the "ECS Toolkit" for building MT systems based on LFG) was started in the early 1980s by Larry Gibson, a former Weidner VP. It's not surprising - there were lots of people in the area with a strong language background.

I went to IBM in 1980 when the BYU project disbanded, and was a programmer at IBM Endicott Lab. In the winter of 1983 I went to IBM TJ Watson Research Center (near to New York City) where there was a group working on language technology. They needed s systems programmer with linguistic experience.

George Heidorn was the group manager. He had done his PhD at Yale, and is the creator of PLNLP (Programming Language for NLP). Karen Jensen, a linguist there had written a syntactic grammar of English and they had started to use it for grammar checking. *(Coincidentally, Michael McCord arrived at TJ Watson the same year as Steve Richardson. McCord wanted to focus on MT and had developed "slot grammar" which he wanted to exploit. McCord still heads one of the MT research groups at IBM, and MT systems based on his slot grammar form the core of IBM's WebSphere MT offerings. However, in the early 1980s there was a little competitiveness between the two groups. George Heidorn's group gave their PLNLP parser to the IBM Tokyo group, which used it to build the first SHALT MT system. — ed.)*

In 1986-87, we started building parsers & grammar checkers in other languages. By 1988 the technology was ready to put into products. Around this time, we hooked up with an IBM group in Bethesda, and all three of us moved there. We worked with the development group there to bring the grammar checker into a product called "ProcessMaster," which worked in VM/CMS. It was an enterprise-level publishing system. The whole division in Bethesda group worked on "Office Vision," IBM's be-all end-all office solution. The team then began to ready the grammar checker to be part of a word processor in Office Vision. Development of Office Vision involved thousands of developers at 10 locations worldwide, and ultimately it crumbled under its own weight. In early 1990, after great support at IBM, things went downhill. It was obvious that the Office Vision product wasn't materializing, and the market was already owned by other products. IBM was trying to play catch-up with WordPerfect and Word. The group decided to try license the grammar checker to other software vendors, and got permission to do so. We pitched it to Microsoft and WordPerfect. Halfway through the year – IBM retracted permission for promoting the technology. They were afraid to give any leading edge technology to the competitors with whom they were trying to catch up.

By February 1991, the group was very frustrated. Finally, Karen just called up Microsoft. Bill Gates & Nathan Myrhvold had just decided to start Microsoft Research, and NLP was one of the areas they wanted to work on. By April, all three of us were hired. When we decided to leave, we were still hoping that we could continue to collaborate with our colleagues at IBM. But unknown to us, IBM and Microsoft were going through an ugly divorce at exactly that moment in time. IBM started trying to lay people off. We had planned to take the early leave incentive, but ultimately we had to sacrifice the resignation bonus because we refused to sign a non-compete agreement. The story actually appeared on the front page of the New York Times business section. May 21, 1991, reporting that three researchers from IBM leaving to form Microsoft research. IBM stock actually went down, and Microsoft went up following the announcement. It was the only time that I have personally affected the stock market! It was the first quarter that IBM announced a quarterly loss too.

In 1993 we published our collected papers from the IBM years in a book: Jensen K., Heidorn G., and Richardson S. Natural Language Processing: The PLNLP Approach, Kluwer Academic Publishers, 1993. IBM objected at first, claiming we were releasing proprietary information, but all the papers had been previously published in conference proceedings, so they gave up.

We started the Microsoft NLP group building syntactic parsers and dictionary technology. In 1995 we started working with multiple languages. At the time, Microsoft was licensing grammar checkers from elsewhere. But by Word '97, the grammar checker was internally-developed, based on our technology. By 1999 we had parsers and other components for a lot of languages, but we hadn't done translation, so we decided to try it.

Around 1994 a natural language development group started within our research group. Shortly thereafter, they became their own entity, working on moving the grammar checker into Word. Now they focus on supplying the rest of Microsoft with various forms of natural language technology, while the research group has been focusing for the last 4 years on machine translation.

MTNI: You had a remarkably long collaboration with Karen Jensen and George Heidorn, what kept you together as a team?

SR: A good friendship, passion for NLP, and common philosophy of trying to create something that can actually be used by lots of people. Karen Jensen, George Heidorn, and I spent 8 years together at IBM and 11 at Microsoft. By the way, they both retired from Microsoft early last year after making great contributions to both IBM and Microsoft, and to NLP in general.

MTNI: Are you responsible for the "that" vs. "which" distinction in the Word grammar checker?

SR: No, but I do admit that I understand the distinction, and that because of my work with grammar checking, it often triggers an "error" in my head whenever I read text that incorrectly uses those words.

MTNI: Didn't you do your PhD at CUNY?

SR: When I came to IBM I took classes at SUNY Binghamton. From IBM TJ Watson, I took classes at CUNY, as there was a work/study program offered for IBMers, and started a PhD in computer science. Early on in 1988, I started to work on aligning parse structures (a precursor to our MT work over a decade later). I had finished up my coursework before we moved to Bethesda, and continued the research on the side. At Microsoft, I started working on MindNet, and then made that the focus of my dissertation. I finished in 1996 – the outer limit of time, and got my PhD in February 1997. My dissertation is available on the Microsoft Website, along with other publications from the NLP group at Microsoft Research: ***http://research.microsoft.com/nlp/nlppubs.aspx***.

MTNI: Machine Translation can be a somewhat discouraging business, what keeps you interested?

SR: I've been working on MT or MT-related technology for 28 years now. I've always loved languages, enjoyed computers, and been fascinated by the challenge posed by MT. Another way to say it is that it simply gets into your blood. I jokingly tell my friends that it's the ultimate "job security" job – it will yet be many, many years before we have something approaching general, high quality MT.

MTNI: Where do you think that MT is going or should go?

SR: First, data-driven MT opens up great possibilities that were never there before. We can build and train a system using resources that are constantly being created in our company (and the same is true for hundreds of other companies). We have millions of sentence pairs that we can use for training, and in fact we retrain the system every night, and do regression testing. This capability will lower barriers to machine translation.

What has impeded growth of MT so far is that: 1) standalone MT is not useful. Integration is something that researchers never want to bother with. But for users, it is critical, and it usually turns out to be an unexpected and huge cost. 2) Customization issues are prohibitive. We've been stuck with broad coverage low quality MT on the Internet or high quality, high cost, customized MT. There hasn't been anything in between.

Lower cost customized MT will open up lots of opportunities. Lots of people would use MT if they had ever had the chance to use customized MT….

Lower cost customized MT will open up lots of opportunities. Lots of people would use MT if they had ever had the chance to use customized MT, but the current manual customization approach is not only cost, but time prohibitive.

What we have with Babelfish (the free MT service using Systran on AltaVista) is a monolithic system that tries to cover hundreds of thousands of terms, but never has the context that one needs. The answer is in the Internet. If we could develop MT systems that were easily customized to any domain and then made them available on the Internet, we could have a huge collection of networked MT systems. Then, when you wanted a translation, rather than sending it to a monolithic MT system, it could be routed to the right engine (one of perhaps thousands of customized engines). This collective "brain" on the Internet could provide the coverage we need. Data-driven MT is not just specialized, but it is easily and readily specialized.

MTNI: Computational linguists often have trouble explaining their work to their families. Do your relatives understand what you do?

SR: I say I work to make computers *seem* like they know something about human languages. It's always easiest, though, to talk about it in terms of specific applications, like grammar checking or machine translation. People generally know what those are.

❑

Cross Language

...continued from page 3

the business as per their assessment. They also assist with implementation and integration. To help their potential and existing customers familiarize themselves with the obstacles and benefits of machine translation, they offer a one-day workshop.

Cross Language is based in Gent, Belgium.
See: ***www.crosslang.com***
Tel: +32 (0)9 267 64 73;
info@crosslang.com

❑

PROMPT

...continued from page 5

efficiency of applying this software to reduce the volume of routine work, and also expected and real increase of productivity of all participants in the translation process. For those who make decisions about what solution to apply, Corvet will surely become fine and objective 'advisor,'" - *Svetlana* Svetova said. Svetova is the director of linguistic technologies of PROMT.

Free online test of PROMT XT is available at
www.Online-Translator.com.
Alexander Andreev; Marketing Manager; Tel: +7 (812) 327-4425; E-mail:
Alexander.Andreev@promt.ru

❑

Write for MTNI

Got an idea for MTNI? We need editorials, letters, news and features related to the MT community. If you've got a news item or a story that you'd like to see published in MTNI, just contact Laurie Gerber (***mtni@eamt.org)***, David Clements (***dclemen1@san.rr.com***) or one of the regional editors.

We want to hear from you!

❑

Conferences and Events

RANLP 2003

September 10-12, 2003
Borovets, Bulgaria

RANLP-2003 (Recent Advances in Natural Language Processing) is the fourth RANLP conference. The RANLP-events have always been a meeting venue of scientists coming from all parts of the world, facilitating contact between researchers from Central and Eastern Europe with their colleagues from Western Europe and America. Both sides greatly benefit from talks and exchange of ideas and experience. The acceptance rate for papers submitted to RANLP is relatively low, so the scientific level of RANLP events is internationally recognized as very high. The full list of accepted papers is available on the conference website: ***lml.bas.bg/ranlp2003***.

Preconference Tutorials

September 7: Dan Cristea, University of Iasi, "Discourse theories and technologies" and Piek Vossen, Irion Technologies BV, "Wordnet, EuroWordNet and Global Wordnet";

September 8: Hamish Cunningham, Sheffield University, "Name Entity Recognition" and John Prager, IBM T.J. Watson Research Center, "Question Answering";

September 9: Ido Dagan, Bar Ilan University, "Machine Learning in NLP" and Inderjeet Mani, Georgetown University, "Automatic Summarization."

Keynote Speakers

Branimir Boguraev (IBM), Shalom Lappin (King's College), Inderjeet Mani (MITRE/ Georgetown University), Stephen Pulman (Oxford University), Hans Uszkoreit (University of Saarland), Yorick Wilks (Sheffield University)

Organizing Committee

Prof. Ruslan Mitkov, University of Wolverhampton (UK), Program Committee Chair ; Central laboratory for Parallel Processing(CLPP), Bulgarian Academy of Sciences (BAS), Local Arrangements

See: ***lml.bas.bg/ranlp2003/.*** ❑

LangTech 2003

November 24-25, 2003
Paris, France

LangTech 2003 will feature keynotes from leading players, presentations from a wide range of developers and solution providers, panel discussions of key issues affecting the market in Europe and beyond, and an exhibition of applications, products, services and research prototypes. Special sessions will enable start-up companies to promote and pitch their products and services and explore funding possibilities.

Demonstrations of applications, products, services and research prototypes will be featured in the exhibition, and the forum will provide ample opportunities for face-to-face meetings with potential users, providers, partners and investors. LangTech 2003 will also offer pre-conference tutorials on new methods and hot technology developments.

Topic Areas

Technologies: existing speech and language technologies ready for deployment.

Solutions: new solutions ready or close to market.

Transfer: case studies showcasing successful technology transfer.

Exploitation: best practice reports on exploitation of speech and language technologies.

Marketing: success stories on the marketing of speech and language technologies.

Financing: venture capital for companies in the HLT sector.

Trends: new trends in research and future market opportunities.

Targeted Technologies

Speech technologies and applications: voice-controlled products and services, speech recognition an synthesis, etc.

Semantic Web and knowledge management: content management systems, text mining, authoring and search environments, taxonomies, etc.

Multilinguality: applications and solutions in localization, machine translation systems, cross-lingual information retrieval, speech-to-speech translation, etc.

Target Audience

LangTech is targeted at developers, integrators, entrepreneurs, researchers, facilitators, investors, users of language technology, as well as media representatives and technology information providers.

LangTech 2003 is organised by ELDA, the Evaluations and Language resources Distribution Agency, with the collaboration of several European organisations.
See: ***www.lang-tech.org***. ❑

ALTW2003

December 10, 2003
Melbourne, Australia

A one-day workshop on Natural Language Technology will be organized by the Australasian Language Technology Association (ALTA). The workshop will be held in conjunction with the Australasian Language Technology Summer School in Melbourne: ***www.cs.mu.oz.au/research/lt/ALTSS2003/***.

Workshop Goals

The goals of the workshop are: to bring together the growing Language Technology (LT) community in Australia and New Zealand; to encourage interactions between this community and the international LT community; to provide an opportunity for the broader artificial intelligence community to become aware of local LT research; to provide a forum for discussion of new research; to foster interaction between academic and industrial research. Our hope is to get as many Australasian LTers together as possible. We also encourage non-Australasian LTers to submit papers, and to participate in the workshop.

Topics

Topics include, but are not limited to: speech understanding and generation; phonology, morphology, syntax, semantics, pragmatics, and discourse; interpreting and generating spoken and written language; linguistic, mathematical, and psychological models of language; language-oriented information extraction and retrieval; corpus-based and statistical language modeling; machine translation and translation aids; natural language interfaces and dialogue systems; message and narrative understanding systems; computational lexicography.

Send in Your Submissions

We particularly encourage submissions that broaden the scope of our community through the consideration of practical LT applications. We especially invite people from industry working on LT to send us their submissions and offer an opportunity to discuss and demonstrate their latest applications in front of an informed audience.

Program co-chairs are: Alistair Knott, University of Otago (NZ); and Dominique Estival, DSTO (AU)

See the conference website: ***www.cs.otago.ac.nz/research/ai/ALTW2003****; You can contact the workshop organisers for further information:* ***altw-info@cs.otago.ac.nz.*** ❑

ALTW2003 Important Dates

Submission deadline	August 30, 2003
Notification to authors	Sept. 22, 2003
Camera-ready copy	October 20, 2003
Workshop	December 10, 2003

LREC 2004

May 24-30, 2004
Lisbon, Portugal

LREC 2004 is the fourth in the biannual Language Resources and Evaluation Conference series, organized by ELRA, the European Language Resources Association. The aim of this conference is to provide an overview of the state-of-the-art, discuss problems and opportunities, exchange information regarding language resources (LRs), their applications, ongoing and planned activities, industrial uses and needs, both with respect to policy issues and to technological and organizational ones. Corpora and Lexica

Examples of LRs are written or spoken corpora and lexica, which may be annotated or not, multimodal resources, grammars, terminology or domain specific databases and dictionaries, ontologies, multimedia databases, etc. LRs also cover basic software tools for the acquisition, preparation, collection, management, customization and use of the above mentioned examples.

Integration of LRs

The Conference targets the integration of different types of LRs (spoken, written and other modalities) and of the respective communities. To this end, LREC encourages submissions covering issues which are common to different types of Language Technologies, such as dialog strategy, written and spoken translation, domain-specific data, multimodal communication or multimedia document processing, and will organize, in addition to the usual tracks, common sessions encompassing the different areas of LRs. ❑

LREC 2004 Important Dates

Panel/workshop proposals	Oct. 20, 2003
Papers/posters/Demos	Oct. 31, 2003
Panel/workshop notification	Nov. 14, 2003
Papers/posters notification	Jan. 23, 2004
Camera-ready copy	March 1, 2004

EURALEX 2004

July 6-10, 2004
Lorient, France

The EURALEX Congresses bring together professional lexicographers, publishers, researchers, scholars, and others interested in dictionaries of all types. The program will include plenary lectures, parallel sessions, software demonstrations, pre-congress tutorials and specialized workshops, a book and software exhibition, and social events for participants and their guests.

EURALEX 2004 is organized by the Faculté de Lettres et Sciences Humaines of the Université de Bretagne Sud, Lorient.

Contact Geoffrey Williams, ***geoffrey.williams@wanadoo.fr***
See: ***www.univ-ubs.fr/euralex2004****.* ❑

EURALEX Important Dates

Submission deadline	October 23, 2003
Notification to authors	February 1, 2004
Final papers due	March 15, 2004

COLING-20

August 23-27, 2004
Geneva, Switzerland

The COLING conference series is an approximately biannual conference, organized by the International Committee on Computational Linguistics. Information about submission, registration and venue will be posted in due time on the Web site listed below.

Organizing Committee Chair: Prof. Margaret King, University of Geneva, ***Margaret.King@issco.unige.ch***.

Program Committee Chair: Prof. Sergei Nirenburg, University of

Continued on page 22 ►

MT Summit IX

...continued from page 2

Recently, there has been a fair amount of research into extracting translation-relevant knowledge automatically from human-built bilingual texts. Over the past years, several statistical MT projects have appeared in North America, Europe, and Asia, and the literature is growing substantially. We'll overview this progress.

Tutorial 6: **MT Customization**
Remi Zajac, SYSTRAN Software, Inc.

MT customization is becoming the preferred option for deploying high-quality machine translation systems for specific applications. This tutorial will give a detailed description of the process and tools for customizing MT systems with examples. Topics include why to customize an MT system, how to evaluate the costs and the potential benefits, and how to test and evaluate the customized system.

Workshops

A number of workshops of interest and impact for MT researchers, developers, vendors or users of MT technologies will take place towards the end of the MT Summit. Each workshop has its own web site including a Call for Papers and other details.

Pre-Conference Workshops: Tuesday, September 23

Workshop 1: **AMTA SIG-IL Sixth Workshop on Interlinguals**
Organizer: Stephen Helmreich (NMSU)
http://crl.nmsu.edu/Events/FWOI/SixthWorkshop/call.html

The Fourth and Fifth IL Workshops have featured active participation by workshop members in the substance of the workshop: they have been workshops in the literal sense of the word. The Fifth IL workshop, in particular, asked participants to code thematic roles prior to the workshop and then to make a short presentation about their activity.

This workshop will continue in that tradition. Instead of focusing on thematic roles, workshop participants will be asked to identify and mark up events, states, and objects in three texts: one English text, a translation of that text into another language, and a re-translation back into English of the second text. Active participants will also provide a short paper, discussing the markup task. In the afternoon, the combined results of the coding experiment will be discussed.

Workshop 2: **Machine Translation for Semitic Languages**
Organizers: Violetta Cavalli-Sforza (Carnegie Mellon), Alon Lavie (Carnegie Mellon), Nizar Habash (Univ. of Maryland)
http://www-2.cs.cmu.edu/~alavie/semitic-MT-wshp.html

Over the past decade there has been some progress on the computational processing of Semitic languages. Several workshops in recent years - both regional and affiliated with international conferences - have addressed the spectrum of issues relating to the processing of Arabic and other Semitic languages. The progress of recent years has opened the door to advanced computational applications such as MT. Research on MT of Semitic languages is, however, still in its early stages. Accurate translation of Arabic, Hebrew and other Semitic languages requires treatment of unique linguistic characteristics, some of which are common to all Semitic languages, others specific to each of these individual languages and their dialects.

Post-Conference Workshops: Saturday, September 27

Workshop 3: **Teaching Translation Technologies and Tools**
Organizers: Mikel Forcada (Universitat d'Alacant), Harold Somers (UMIST), Andy Way (Dublin City University)
http://www.dlsi.ua.es/~mlf/t4/

In view of the success of the preceding workshops on Teaching MT, the first held as part of the the last MT Summit in Santiago de Compostela in September 2001, and the second the 6th EAMT Workshop held in Manchester in November 2002, we propose a third workshop with an expanded scope which will not only address MT but also computer-aided translation technologies and tools. The workshop will provide an opportunity for MT and CAT instructors to exchange their experience by presenting papers or demonstrations describing the tools and techniques they use in the classroom or in the laboratory.

Workshop 4: **Towards Systematizing MT Evaluation**
Organizers: Leslie Barrett (Transclick, Inc., New York, NY), Maghi King (ISSCO/TIM/ETI, University of Geneva), Keith Miller (MITRE Corp), Andrei Popescu-Belis (ISSCO/TIM/ETI, University of Geneva)
http://www.issco.unige.ch/projects/isle/MTE-at-MTS9.html

Estimating the quality of any MT system accurately is only possible if the evaluation methodology is robust and systematic. The NSF and EU-funded ISLE project has created a taxonomy that relates situations and measures for a variety of MT applications. The "Framework for MT Evaluation in ISLE" (FEMTI) is now available online. The effort of matching these measures correctly with their appropriate evaluation tasks, however, is an area that needs further attention.

MT Summit Invited Speakers

Pierre Isabelle
Area Manager, Content Analysis,
Xerox Research Centre Europe, Grenoble
Multilingual Document Processing at XRCE

Akitoshi Okumura, NEC, Tokyo
Senior Manager, Human Language Technology Group
NEC Corporation, Japan
Development of Speech Translation for Hand-held Devices

Donald Barabé
Director, Business Development, Translation Bureau, Public Works and Government Services

Canada

Soaring Demand, Shrinking Supply in Translation: How We Plan to Make Ends Meet

Product Exhibition

The last couple of years have brought some exciting developments in many areas: evolutionary advances in MT research, lots of forward-thinking deployments of MT, significant improvements in existing commercial products, and a whole cohort of startup companies commercializing newer approaches to the core problems of machine translation.

The product exhibition will commence with the opening reception of the evening of Tuesday, September 23, and continue for the duration of the main conference.

Exhibitors

- ArchiText Translations
- Beetext
- Basis Technology
- Ciyasoft
- Corporate Language Services
- LanA Consulting
- Language Weaver, Inc.
- IBM
- Language Technology Centre
- Multilingual Computing
- Pan American Health Organization (PAHO)
- Systran Software

Get More Information

The Summit Web site has full program, date and venue information. Look for it at ***www.mt-summit.org***.

Plus: Panel discussions, live system demonstrations, free reception, sumptuous banquet in an unusual setting - all within walking distance of the French Quarter! A lively social agenda will include a reception and a surprise banquet that promises a very enjoyable evening.

Online registration is now open and, as a member of any IAMT regional association (AAMT, AMTA, EAMT), you are of course eligible for a discounted registration rate. This Summit promises to be a landmark conference.

Hotel: Special room rates of $149/night in the historical and elegant Fairmont New Orleans.

Social Events

Welcome Reception: Tuesday, September 23, 6:30-8:00

Banquet: Thursday Evening, September 25

About New Orleans

For more information on things to do and see while in New Orleans, check out New Orleans Online: ***http://www.neworleansonline.com/***.

❑

MT Summit IX Important Dates

Final papers	July 31, 2003
Late Registration	September 15, 2003
Conference	September 23-28, 2003

DDMT Defined

...continued from page 11

uses careful statistical justifications (these have so far usually used a statistical model built during training, and *not* the corpus, at runtime, but this has started to change recently)

3) a Rule-based system uses a set of discrete rules. (If they are built in advance, the distinction from EBMT is clear; if a system automatically builds rules from a corpus at runtime, the distinction from EBMT gets fuzzy.)

You can easily create hypothetical systems that straddle each of these boundaries, and in fact a number of the statistical systems in the current DARPA MT evaluations build a phrase-based component at runtime from the corpus, which I think makes them *both* SMT and EBMT at the same time.

Jeff Allen

We might want to consider that Translation Memory (TM) tools are more or less the translation job production-based representatives of EBMT systems. Note: I want to avoid the word commercial-based here because some TM tools are in fact industrial in-house built production tools rather than commercial off-the-shelf (COTS) products. There are TM tools ranging from those that are more purely example-based (some in-house built TM systems), to those like TRADOS Workbench and other commercial tools that have fuzzy matching threshold levels that can be set by the user, to those that specifically allow for the Hybrid approach of using TM/EBMT + MT.

Ed Hovy

It seems interesting to differentiate between the processing (translation) stage and the core knowledge resource(s) used to perform the translation (the data/rule gathering). All MT systems use knowledge that specifies transformations of source into target, usually via a series of steps. Whether this knowledge is encoded as (traditional) rules, as EBMT-style patterns, or as probabilistic tables, it seems increasingly the case that people use either manual or automated (learning) methods to acquire the knowledge. You can build these resources manually or using a learning/counting program. At "runtime," doing the translation, the system can use this knowledge in ways that look more "statistical" or more "rule-based."

Is this such a big difference in paradigm? It seems so today, because we are new to these procedures, and because statistical MT was introduced to us in a rather colourful and perhaps overly combative way. But I suspect that as research in statistical MT continues, it will narrow the gap between statistical and rule-based systems, and then IBM's "pure" models 1 and 2 will be seen for what they are: a new way of implementing the very oldest form of MT, namely direct replacement.

❑

Thurmair

...continued from page 9

or for migrating.

SAIL Labs got out in time to (avoiding entanglement in the scandal) but L&H had been the main sponsor and SAIL didn't have its own sales force. As soon as L&H collapsed, we had to look for sponsorship. Somehow, we always find someone who is interested.

MTNI: What are the goals and vision for the translation technology now?

GT: To try to support a client server and web access infrastructure; More statistical technology in terminology extraction to shorten the time to set up terminology customization. What causes the worst response to machine translation is bad translations of terms. We also need to handle untranslated terms, proper names etc.

We want to work in the context of Compren-dium's content technology, for example Cross Lingual Information Retrieval (CLIR). We got a couple of projects in law enforcement. We did CLIR where you could query in any European language, the search would be done in English, and the results were translated back into the source language, either by full translation or by key terms.

We work on named entity recognition. We work on classification – subject area identification – which works well with the content management, for example to distinguish medical from insurance from clinical reports. We also have news classification – economics vs. sports. This can also be used for email routing. We work on Internet protection technology to kill pornographic sites. Our technology is quite adaptable. There is a natural link to content management.

MTNI: MT is generally a discouraging business, and you have had many ups and downs. How do you (and your colleagues) keep up your spirits and motivation through so many changes of organizations? Or more simply, why do you stay in this field? (Or if you don't find it to be a discouraging business, why not?)

GT: I can only speak for myself here. I think there are two reasons:

1. I come from the language side originally (studied literature, linguistics etc.). I simply want to know how far you can get in machine-supported treatment of language. The point is that language reflects human mind, which (I am convinced of this) can NEVER be covered by machine. However, I am also convinced that machines can do better than they do now. And the challenge is to reach this limit, and this we have not reached yet. So it is a bit like trying out how far you can come in reaching the impossible.

2. I find many other activities not really challenging. Programming bank terminals, or programs to monitor the system behavior, is difficult in terms of complexity but is more like putting together simple pieces into a building but it assimilates to the mechanics of what the machines can do.

> ***However… machines can do better than they do now. ... it is a bit like trying out how far you can come in reaching the impossible.***

You see commercial success is not on top of my list, and the success and personal motivation is more on the technical side: We had the best MT system (well, WE believe), we had information extraction and cross-lingual retrieval systems with the best language coverage I know, we had a super term extraction system for 11 European languages, and a lot of other things. Only the companies I worked for did not want to, or did not know how to make commercial use of it. Also, as long as we do not improve the QUALITY of our language technology, commercial success will be very difficult to achieve: Just assume you had a translator with 100% correctness: This would be the hit in the market for sure: Because it solves a PROBLEM that people really HAVE. ❑

EAMT/CLAW 03

...continued from page 7

extremely healthy mix: one of the aims of the conference was to provide scientists with an opportunity to make contacts with other industrial and academic research bodies and to stimulate cooperation with these bodies. All too often academic conferences are unable to attract large numbers of industrial participants.

A Prudent Investment

Given the obvious relevance of the conference theme to the language and localization industries both in Ireland and abroad, we were able to buck this trend by attracting many more attendees from industry than from academic institutions. This is all the more extraordinary given the current economic climate in which we are all operating: despite cutbacks in the language and computing areas given the downturn in the IT industry, many companies thought it a prudent investment of valuable resources to send employees to our conference—if additional proof were needed of the importance to the global economy of the field of translation and the ever increasing reliance on computer-assisted translation tools to meet the demands of translation, this is surely it.

Copies of the Proceedings (hard copy and CD-ROM) are available from the EAMT website: ***www.eamt.org****.*

Dr. Andy Way is a Senior Lecturer at the School of Computing, Dublin City University;
Email: ***away@computing.dcu.ie****;*
Tel: +353-1-7005644

❑

MT Users' Desiderata Part II

by Jackie Murgida

What do users really want from machine translation? What would they buy and actually use, if it existed for their language pair/ direction? This is the second installment in a series of articles addressing these questions. It continues the list of features, begun in Part 1 (MTNI #30, March 2002), that translators would want in an MT system. A future installment will address the requirements of non-language specialists, such as researchers and analysts, and individuals using the Internet and Web, as well as users in enterprises like multinational businesses and international organizations.

Part 1 said that translators want: high-quality raw output; a comprehensive lexicon with multiple, stackable domains; flagging and correct translation of proper nouns; and diagnostic linguistic information from the system that would help a translator during post editing. In this installment, I explore other features that would increase MT usability for translators

On-Line References

Translators want a suite of relevant references with information that isn't already incorporated in the MT dictionary for their languages and domain. These should be accessed easily during the post editing process. Examples are monolingual source- and target-language dictionaries and thesauri and bilingual dictionaries, all covering as many domains as possible, as well as other sources frequently used by translators, such as encyclopedias, target-language style manuals, grammar and usage books, and concordances.

MT lexicons are usually not sufficient for a translator who has to produce a publication-quality product. Translators need the kind of information given in traditional lexical resources, such as examples and grammatical information, presented in a translator-friendly way -- not written in computer code. It's always possible to consult the hardcopy sources, but electronic versions would make post editing more efficient.

While I think most people would agree that "the more, the better" applies here, realistically the rule of thumb for electronic references would be to put the most dog-eared hardcopy resources for a language pair/direction and domain—or their equivalent—in electronic form. One colleague who translates financial and legal material from several languages into English told me she would like to have Black's law dictionary and her Herbst's English-German-French dictionary of commerce, finance, and law accessible while postediting or translating. That's, of course, in addition to the standard, general dictionaries for her languages.

This applies to Web-based resources, as well. Translators can and should search online for terms and consult Websites that offer such aids as specialized dictionaries, encyclopedias, and Islamic-Gregorian date converters. However, the ones they use most often should be integrated into the MT program, or be easily accessible from it.

Installation, Interface, and Care and Feeding Must Be Easy

Remember, we're talking, for the most part, about people who use computers as a tool, not as an end in itself. The majority studied foreign languages, even medieval poetry, not C++, Java, and Perl. They want to insert the CD, follow some prompts, and have the program work on their own PC or Mac. As one colleague put it, the software should be able to make itself at home on any platform.

The user wants options on the look and feel of the interface, how things are arranged on the screen: source text above or below, or to the left or right of the target text, for example. Type size and fonts should be easily adjustable too.

Make dictionary update easy. Translators want both to enter new terms on the fly while post editing, and to import whole glossaries without having a degree in computational linguistics. In addition, good terminology management tools should be integrated into the system, with the latest facilities for terminology exchange and for adding terms from the terminology tool to the MT lexicon.

Electronic references should be accessible while post editing. The user should be able to drag and drop the desired term into the translated text or otherwise select it and have it appear where the cursor is in the target text, replacing anything highlighted in the target text.

Post editing should be exactly the same as word processing. That is, it shouldn't be in a different editor that is more primitive to use than the latest version of Microsoft Word. In fact, MT developers should consider offering special macros for revising translations, such as reversing the order of two words.

One Program for Multiple Languages

For translators who work in more than one language pair/direction, it's very desirable to have them all in one product. They don't want to switch from one product to another and confuse their brains with different procedures and operations for each one.

Clear, Well-organized Documentation and Online Help

This is a tough one. Good translators are good at words. They're impatient with what a friend of mine calls "helpless help." You want to do something, but the help file or the hardcopy manual is arranged in a baffling way. Or you find the task or topic you're looking for and are overwhelmed with too many ways to do the same thing, only it's not clear if they really are for the same thing. Maybe translators should write the documentation?

Improvability

It would be great if the system could

Continued on page 22 ►

Data-Driven MT Grows Up

By LG

This is the second part of an article on "Upstart Data-Driven MT Companies" that appeared in MTNI 32. In this article, we continue to catalogue the ever-growing number of new and existing companies that are developing systems using "empirical" approaches to MT.

Company: Behavior Design Corporation

Behavior Design has internal testing underway on a new, Corpus-Based Statistics-Oriented MT system. The system's core approach is Corpus-Based Statistics-Oriented Two-Way Training. The languages of interest are English-Chinese.

Company: Hua Jian

The Research Center of Computer & Language Information Engineering, Chinese Academy of Sciences has been engaged in MT researches for more than 10 years and has made some achievements.

They started research on Intelligent MT in 1986 and research on hybrid strategy MT in 1998. So far, we have implemented multilingual MT systems and specialty MT systems, including: English-Chinese, Chinese-English, German-Chinese, Russian-Chinese, Japanese-Chinese, Chinese-Japanese, Chinese-Spanish, and Chinese-French. On the basis of all these, a series of product-related systems have been developed, for example, large-scale Integrated Network Information Translation Processing System, PC-oriented Multi-lingual Machine Translation Software, Embedded Machine Translation System, Machine Translation Aided Processing System and Service-oriented MT Engine. They are widely used by well-known enterprises in China and abroad such as IBM, Compaq, Toshiba, NEC, Legend, Founder, etc.

Hua Jian first sold one of their hybrid MT systems, English-Chinese Intelligent Aided Translation (IAT) in June 2002.

Hua Jian's core approach uses integrated hybrid translation strategies, including a rule-based approach, an example-based heuristic analogy approach and a statistical method.

All 8 language pairs are hybrid systems. Regarding the English-Chinese, Chinese-English and Russian-Chinese systems, they have already been put into commercial use and developed into products, such as Huajian IAT, Huajian Easytrans, etc. Although Hua Jian didn't develop products with Japanese-Chinese and Chinese-Japanese systems, they have licensed their core technology to Japanese Logovista Co. German-Chinese, Chinese-Spanish, and Chinese-French systems will be commercially available in October this year.

Hua Jian
Beijing, China
+86-10-62333660.
www.hjtek.com

Company: Language Weaver

Language Weaver was founded to commercialize 20 person-years of research conducted at USC/ISI. The group has been very successful in advancing the state of the art in statistical MT, and felt that the technology was mature enough to compete favorably against the existing state of the practice.

Language Weaver's core approach is Statistical MT.

What sets your system apart?

- Quick, easy, automatic customization to new subject areas and text types.
- Ongoing technology transfer agreement to commercialize advances made by the statistical MT research group at USC/ISI
- Deep commitment to pushing up the quality ceiling for both specialized and general purpose translation.

4640 Admiralty Way, Suite 423
Marina del Rey, CA 90292
Tel: 310-437-7300
www.languageweaver.com

Company: Microsoft

The NLP Research group was started in June, 1991, as the first group in Microsoft Research, with George Heidorn, Karen Jensen, and Steve Richardson, who together started building the basic components (parser, grammar, dictionary) from which various applications, including our MT system, have emerged. The work at Microsoft had its roots in earlier work at IBM Research, which originated from Heidorn's Yale dissertation. The application to MT, and in particular the development of the example-based transfer component, was led by Richardson but has involved many members of the group.

Currently the MT system is only being deployed inside Microsoft, for use in translating the online Support Knowledge Base, assisting in product localization, and aiding communication between support personnel and customers. After a successful Spanish pilot in 2002, permanent deployment of the Spanish KB is targeted for the end of the first quarter of 2003. Japanese, German, and French pilot deployments are expected in 2003

MSR-MT is a data-driven MT system that combines rule-based and statistical techniques with example-based transfer. Microsoft believes this hybrid system to be the first practical large-scale MT system capable of learning all its knowledge of lexical and phrasal translations directly from data.

The central feature of the system's training mode is an automatic logical form (LF) alignment procedure, which creates the system's translation example base from sentence-aligned bilingual corpora. During training, statistical word association techniques supply translation pair candidates for alignment and identify certain multi-word terms. This information is used in conjunction with information about the sentences' LFs, provided by robust, broad-coverage syntactic parsers,

Continued on page 22 ►

Data-Driven MT Companies at a Glance

Company: Behavior Design Corporation

Founded: 1988
Inventor: Keh-Yih Su
CEO: Keh-Yih Su
President: Steel Su
Customer/Investor contact: Keh-Yih Su. ***kysu@bdc.com.tw***
First Product deployment: BehaviorTrans (1989)

Company: Hua Jian

Founded: June 1997
Inventor: Research Center of Computer & Language Information Engineering, Chinese Academy of Sciences, Beijing
President: Dr. Huang Heyan, ***heyan.huang@hjtek.com***
Customer/Investor contact: Mr. Zhou Ding, ***ding.zhou@hjtek.com***
First product deployment: June, 2002

Company: Language Weaver

Founded: January 2002
Inventors: Dr. Kevin Knight and Dr. Daniel Marcu and research associates at University of Southern California, Information Science Institute (USC/ISI)
President/CEO: Mr. Bryce Benjamin
Customer Contact: Ms. Laurie Gerber, ***lgerber@languageweaver.com***
Investor Contact: Mr. Bryce Benjamin ***cbryceb@languageweaver.com***
Company Size: 20
First product deployment: July 2003

Company: Microsoft

Founded: 1975
Inventors: NLP Research Group, started in June 1991 (see article for full details)
CEO/President: Bill Gates, of course
Customers/investors contact: n/a
NLP Research group size: about 30 people
First deployment: Currently the MT system is only being deployed inside Microsoft (see article for full details)

Company: MorphoLogic

Founded: 1991
Inventors: founders, Laszlo Tihanyi, Miklos Pal and Gabor Proszeky, as well as other staff members.
CEO: Gabor Proszeky.
Customer/investor contact: Mr. Szabolcs Kincse, manager of international relations, ***kincse@morphologic.hu***
Company size: 22 in R+D (plus 11 in the MorphoLogic Localisation department)
First product sold and deployed: 1991 (Hungarian spell-checker and other proofing tools)

Company: Oki Electric Industry Co. Ltd.

When did Oki start developing empirically based MT systems: 1997 (the NLP R&D group started in 1983.)
Inventor of the empirical technology that is being commercialized: Toshiki Murata, leader of the MT development group/NLP R&D group at Oki
Investor/Customer contact: Toshiki Murata, ***mura@kansai.oki.co.jp***
Size of NLP R&D group: 10 people
Deployment of data-driven MT: The pattern-based machine translation engine is not sold yet, but is used in a pilot deployment called “Yakushite Net.”

More Data-Driven Companies at a Glance on Page 23!

Desiderata

...continued from page 19

keep track of the most frequent corrections a user makes and then the user could change that in the system, without, of course, ruining the rest of the algorithms. (I didn't say all of the desiderata would be implemented easily or within the next 50 years!)

Translation Memory Integrated with MT

When I consulted translators about their MT wish list, more than one asked for a translation memory (TM) tool that is fully integrated with the MT. To be most useful for languages with greatly divergent sentence patterns and lengths, this TM should be based on phrases, not sentences. And for translators who don't have a significant volume of source and target text in electronic form for the legacy data, the TM should have its own base of parallel corpora. Translators could then use the TM immediately.

OCR

Speaking of not having source text in electronic form, many translators do not receive their assignments electronically. They can't use MT without an optical character recognition program. It has to be good, and easily used with the MT system, or already integrated into it.

Dictation Capability

This would be very nice: a voice interface for the post editor. Position the cursor and dictate the corrections.

Formatting and Other Annoying Matters

Then there's the matter of formatting, punctuation, tables, handling of currencies and dates, and a host of related things that are always mentioned in this field. Translators want all of these to be dealt with without being bothered, themselves. They don't want the numbers to appear in the wrong order when translating to and from Hebrew or Arabic-script languages. They don't want to spend hours dealing with weird glitches. They're the kind of people who think that if you can send several people from different countries into orbit in a space station and that if the New England Patriots can win a Super bowl, then you can do the formulaic manipulations of number, date and currency conversions automatically and accurately. This is a much less difficult task than full text translation -- why is it so often fumbled or neglected?

Wouldn't It Be Nice If...

That's the wish list so far. A more general desire is that regardless of which wishes are implemented from our list, the user could pick and choose what to have available and could use the program in the most suitable way for the particular project and end user. For instance, some people don't need as many bells and whistles, as long as the basic translation engine is adequate, because the client is a domain expert and can use raw output with minimal post editing to make it more readable.

Others might not care about the OCR or seeing alternative parses. At the same time, users who produce publication-quality translations want all possible resources for composing and polishing the target-language text. A suite of modules controlled by the user is the ideal, regardless of the type of user.

Jackie Murgida is Director of Cross Language Processes, JTG, Inc., Alexandria, Va. She can be reached at ***jmurg@ttlc.net****.*

❑

COLING

...continued from page 15

Maryland, Baltimore County,
sergei@cs.umbc.edu
See: ***www.issco.unige.ch/coling2004/*** .

❑

COLING Important Dates

Workshop/tutorial proposals	Dec. 15, 2003
Workshop/tutorial notification	Jan. 16, 2004
Submission deadline	March 26, 2004
Notification to authors	May 14, 2004

Data-Driven MT

...continued from page 20

to identify phrasal transfer patterns.

At run-time, these same syntactic parsers are used to produce an LF for the input string. The goal of the transfer component is thus to identify translations for pieces of this input LF, and to stitch these matched pieces into a target language LF, which can serve as input to generation. The example-based transfer component is augmented by decision trees that make probabilistic decisions about the relative plausibility of competing transfer mappings in a given target context.

Target market or application: Creating tuned MT systems for domains for which there are existing aligned bilingual corpora. Specifically, hundreds of thousands of source/target translation pairs, extracted from translation memories that were created while localizing Microsoft's products over the past few years. It is anticipated that any multinational company with a similar resource could exploit our system to produce reasonable translations for their domain.

What sets your system apart from other MT systems?

Previous example-based work using dependency structures (like our LFs) has never before been scaled to a production level, to our knowledge. We combine mature, linguistic technology (parsers used previously in Microsoft's grammar checkers and elsewhere) with statistical and example-based techniques. While other data-driven systems have been researched, undergone evaluations (e.g. DARPA), and are under development for commercial use, we are also not aware of any that are currently deployed in a commercial environment.

Company: MorphoLogic

MetaMorpho is a combination of example-based (but not statistical) and rule-based methods. The target market for MetaMorpho MoBiCAT is everyone who needs to understand English in general. For the coming intelligent translation memory-based MetaMorpho the target is the community of professional translators.

What sets your system apart from other MT systems?

- The linguistic knowledge is formulated in patterns: short patterns are lexical items, long patterns are idioms, underspecified patterns are rules in other systems. All these patterns are treated in a uniform way in MetaMorpho.
- Another important difference is that the target structure is being built while parsing, so we do not have a separate transfer phase, but it is not direct translation in the original word-to-word sense.
- In the MoBiCAT version we use the interface introduced for MoBiMouse (our EU IST Prize winning technology). You do not need to do anything but leave your mouse cursor over the sentence or expression you want to translate, and the translation of the sentence comes automatically in a bubble. If you move your cursor, it disappears.

MorphoLogic
Késmárki u. 8.
Budapest 1118, Hungary
Tel: +36-1-361-4721
www.morphologic.hu

Company: Oki Electric Industry Co. Ltd.

The pattern-based machine translation engine is not sold yet, though it is used in a pilot deployment called "Yakushite Net." Yakushite Net is being used by a community of beta testers. See ***www.yakushite.net***. The system is scheduled to be available commercially in October 2003. Oki's rule-based MT system (PENSEE) has been sold since 1986. Limitations on the quality of translation motivated the development on a new pattern-based MT system in 1997. Oki's core approach to machine translation is pattern-based. See ***http://www.oki.com/jp/RDG/English/pensee/pr971217.html*** .

Grammatical patterns are created by the development team (rather than being learned automatically from text).

Currently, in Yakushite Net can users may add words or terms. Later, they will be able to add patterns.

Learning from postediting: In Yakushite Net, users can postedit the translation result of Web page. Yakushite Net stores the pairs of original sentence and target sentence. Yakushite Net will use the stored target sentence when (exactly) same original sentence is appeared.

Oki reports that they are still "thinking about" their proposed target market.

What sets Yakushite Net apart from other MT systems?

Users can make the quality of the new MT better by themselves. The new pattern-based system will also be called Pensee.

See: ***www.oki.com***.

Company: Sehda Incorporated

Sehda reports their core approach to machine translation as hybrid example-based and statistical.

Sehda Incorporated
465 N. Fairchild Drive Suite 123
Mountain View, CA 94043
Tel: 650-864-9900
www.sehda.com

Company: Verbalis

Dr John Laffling started the academic research project which led to the current Verbalis technology 11 years ago. Since then he has led the development and implementation of the present system. He continues this role as Director of R&D at Verbalis

Verbalis has an example and analogy-based system in commercial operation, where it has already been deployed on large software documentation translation. However, Verbalis is service-oriented rather than product-oriented.

Currently we offer a German to English service and will offer English to German from the end of 2003. The system can then be easily extended to any other language pair and we will continue to expand our services.

The target market is corporate and other large volume users of translation services, specifically in software and technical documentation.

What sets your system apart from other MT systems?

Knowledge base of examples; Example-guided disambiguation; Analogical reasoning; Ability to exploit partial matches in the selection process; Template-guided placement.

Verbalis Ltd
Stadium House Business Centre
Alderstone Road
Livingston EH54 7DN
Tel. +44 (0) 1506 602542
www.verbalis.com

❑

Data-Driven MT Companies at a Glance

(Continued from Page 21)

Company: Sehda Incorporated

Founded: 1998
Inventor: Various people
CEO/President: Farzad Ehsani
Customer/investor contact: Farzad Ehsani, ***farzad@alchemy.sehda.com***
Company size: 12
First product deployment: 2004 (projected)

Company: Verbalis

Founded: 1999
Inventor: Dr John Laffling
Customer contact: Andy Crofts, CEO, ***andy.crofts@verbalis.com***
Investor contact: Adrian Smith, ***adrian.smith@verbalis.com***

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: _______ Postal code: ___________

E-mail: ______________________ Fax: ______________

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
3 Landmark Center
East Stroudsburg, PA 18301

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)
3 Landmark Center
East Stroudsburg, PA 18301

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)
3 Landmark Center
East Stroudsburg, PA 18301

Publications Order Form

Please return this form with payment or credit card information, to:
International Association for Machine Translation (IAMT c/o AMTA)
3 Landmark Center
East Stroudsburg, PA 18301

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.
[2] Member of AAMT, AMTA, EAMT.

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.

Today's date: __/__/__

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:

Mizuho Bank
Ochanomizu Branch
Tokyo, Japan
A/C No. 1737479
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.

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
3 Landmark Center
East Stroudsburg, PA 18301

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*

Representative: ______________________________

❒ *Profit-making institution* — *SFr 350*

Representative: ______________________________

Last name(s): ______________________________ *First name(s):*___________________________ *Title:* ________

Address: ____

____

Home tel.: ________________________ *Work tel.:* ________________________ *Fax:* ________________________

E-mail: ______________________________________ *Website:* ______________________________________

Institution / organzation: ____

Area of specialization:

❒ *MT User* ❒ *MT Developer* ❒ *MT Researcher* ❒ *Translator* ❒ *Manager* ❒ *Other* ___________________

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 be borne by the applicant

❒ *Credit card (please provide information at right ➔)*

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

Card number ________________________________

Expiration date: ___/ ___