Foreword

Machine Translation and the Information Soup!

Over the past fifty years, machine translation has grown from a tantalizing dream to a respectable and stable scientific-linguistic enterprise, with users, commercial systems, university research, and government participation. But until very recently, MT has been performed as a relatively distinct operation, somewhat isolated from other text processing.

Today, this situation is changing rapidly. The explosive growth of the Web has brought multilingual text into the reach of nearly everyone with a computer. We live in a soup of information, an increasingly multilingual bouillabaisse. And to partake of this soup, we can use MT systems together with more and more tools and language processing technologies—information retrieval engines, automated text summarizers, and multimodal and multilingual displays. Though some of them may still be rather experimental, and though they may not quite fit together well yet, it is clear that the future will offer text manipulation systems that contain all these functions, seamlessly interconnected in various ways.

What is the position today? What opportunities and challenges of multilinguality exist on the Web? How can we adapt existing technology to address all languages? This conference offers invited speakers, papers, panels, and plenty of time for professionals in machine translation to learn about the issues, think about them, and discuss them with colleagues and friends. All of this in a pleasant setting. We hope you enjoy it!

This year, as you see, the AMTA conference proceedings have a new format. We are very happy in having secured an agreement with Springer-Verlag, based in Heidelberg, to publish our conference proceedings in the book format from now on. This means that AMTA papers will be distributed on a far wider scale than before. We hope that the format pleases you.

AMTA is fortunate in having a diverse membership of lively, involved, and friendly people. Without them, conferences such as these would be impossible to organize and awful to attend. Great thanks are due to David Farwell of the Computing Research Laboratory (CRL) of New Mexico State University and Laurie Gerber of SYSTRAN Software Inc., who together assembled an extremely interesting and diverse program, containing more papers and more tutorials than any previous AMTA conference. Of special note are the sessions devoted to user studies, an area underrepresented in past years. Linda Fresques and Helen Fleishenhaar of CRL assisted in various ways. We thank the CRL and SYSTRAN in allowing their staff to devote some time to organizing the conference.

The program committee are to be thanked for their reviewing and the comments that helped authors improve all the papers accepted: William Albershardt (US Dept of Defense), Scott Bennett (Logos Corp.), David Clements (Globalink Inc.), Bonnie Dorr (University of Maryland), David Farwell (NMSU Computing Research Laboratory—co-chair), Pascale Fung (Hong Kong University of Science and Technology), Laurie Gerber (SYSTRAN Software Inc.—

co-chair), Kurt Godden (General Motors Corp.), Bonnie Glover Stalls (USC Information Sciences Institute), Viggo Hansen (Hofman-Bang A/S), Stephen Helmreich (NMSU Computing Research Laboratory), Eduard Hovy (USC Information Sciences Institute), Lori Levin (CMU Language Technologies Institute), Susann Luperfoy (then of MITRE Corporation), Elliott Macklovitch (Université de Montréal), Ingrid Meyer (University of Ottawa), Mary Ellen Okurowski (US Dept of Defense), Patricia O'Neill-Brown (US Dept of Commerce), Boyan Onyshkevych (US Dept of Defense), Fred Popowich (Simon Fraser University), Randall Sharp (Universidad Autónoma Nacional de México), Beth Sundheim (NCCOSC RDTE), Virginia Teller (Hunter College, CUNY), Jin Yang (SYSTRAN Software Inc.).

One of the hallmarks of a good conference is the memorability of its invited speeches. With the speakers and topics selected by the program chairs, this conference can't go wrong! Many thanks to the invited speakers, and also to the tutorial presenters and the panelists, who put a lot of work into preparing their presentations!

The local arrangements are in many ways the most harrowing and difficult aspects of organizing a conference. Martha Palmer, Jennifer MacDougall, and Elaine Benedetto, the Local Arrangements Committee, have done a tremendous job, and deserve special thanks. Their attention to the banquet and special events must be particularly recognized, for this is often what helps make a gathering memorable. AMTA gratefully recognizes your work!

The exhibits were managed by Kimberly Kellogg Belvin, who eagerly located exhibitors and arranged exhibition space, computers, and all the seemingly little things that together sum to a smooth and professional presence. Advertising, mailings, and early registration were ably handled by the AMTA Registrar, Debbie Becker, whose continued service to AMTA is appreciated very much.

Several corporations were kind enough to sponsor portions of the conference. Many thanks to SYSTRAN Inc., Logos Corp., Globalink Inc., and the University of Pennsylvania's Institute for Research in Cognitive Science for their generous donations.

May this conference help you enjoy the variety and spice of the Information Soup!

Eduard Hovy
Conference Chair
Marina del Rey, August 1998

Tutorial Descriptions

MT Evaluation

John S. White
Litton PRC, Fairfax, VA, USA

Evaluation has always been a fundamental part of the MT discourse. Yet many participants in the field will claim that there is no generally agreed upon method for evaluation. In part, this sense springs from the realization that there are different purposes for MT, different interests of the participants in the process, radically different theoretical approaches, and, of course, different languages. Also, MT evaluation has some unique difficulties over evaluations of other language systems; in particular, there is never a single 'right' translation, and therefore never a solid ground truth against which MT output may be compared. This tutorial faces all of these issues, discussing the different evaluation needs of different MT stakeholders, the problems of using subjective judgments for evaluation, and a variety of classic and new approaches to evaluation.

Survey of Methodological Approaches to MT

Harold Somers
UMIST, Manchester, England

This tutorial will present various approaches to the problem of Machine Translation taking us from early methodological approaches, through the 'classical' architectures of 1970s and 1980s MT systems to the latest ideas, ending with a consideration of some outstanding topics for MT researchers. On the way we will also consider how various external factors (use and users) affect MT system design.
The tutorial will be divided into six topics, as follows:

1. Historical perspective
2. 2nd generation: Transfer vs interlingua, rule-based systems
3. Making life easier I: sublanguage and controlled language systems
4. Making life easier II: Tools for translators
5. New paradigms: EBMT and statistical MT
6. Hard problems don't go away

Survey of (Second) Language Learning Technologies

Patricia O'Neill-Brown
U.S. Department of Commerce, Washington, DC, USA

Computer Assisted Language Learning (CALL) Programs come in all shapes and sizes these days. They are on the Web and they are on your software store shelf. In this tutorial, we will work through a variety of CALL programs and see what they have got to offer to you as the MT developer, whether you are in the beginning or advanced phases of system development.
The tutorial will be divided into 4 topics:

1. Identifying and locating CALL programs: Where do you find them?
2. What is out there?: Developing a taxonomy of CALL programs
3. Reviewing the reviews on CALL: Evaluating the evaluations and making the most out of them
4. The value of CALL to MT development: What can the MT developer get out of CALL that cannot be obtained elsewhere?

Ontological Semantics for Knowledge-Based MT

Sergei Nirenburg
Computing Research Lab, New Mexico State University, Las Cruces, NM, USA

The idea of representing a source text in a language-neutral format and then generating the target text off the latter is simple and well known. The devil, as always, is in the details. Ontological semantics is a computational-linguistic theory devoted to the issues of deriving, representing and manipulating meanings of concrete natural language texts. While the theory can serve as the basis for many information technology applications, in the area of machine translation it is buttressed by a detailed and tested development methodology developed in the framework of the Mikrokosmos R&D project. This tutorial will address the following topics:

1. Introduction to the concerns, assumptions, content and justification of ontological semantics. Comparison of ontological semantics with other semantic theories
2. The body of the theory
3. Methodological issues
4. Status of theory and methodology development
5. Applications of ontological semantics outside machine translation

Cross Language Information Retrieval

Gregory Grefenstette
Xerox Research Centre Europe (XRCE), Grenoble, France

Contrary to fears and beliefs of five years ago, the WWW will not be an English-only resource. Information is readily accessible in growing numbers of languages. Cross Language Information Retrieval (CLIR) supports the view that foreign language documents are sources of information and not just noise to be eliminated.

This tutorial will present CLIR techniques and recent experiments attacking the problems raised trying to access documents written in one language by a query expressed in another.
The tutorial will be divided into four parts, as follows:

1. The science of Information Retrieval
2. Why CLIR is not information retrieval and is not machine translation
3. Linguistic techniques for CLIR
4. Experiments and results

Speech to Speech Machine Translation

Monika Woszczyna
Carnegie Mellon University, Pittsburgh, PA, USA

With the introduction of commercial text dictation systems (such as Dragon Naturally speaking) and text translation systems (e.g., IBM Personal Translator), two important technologies in natural language processing have become available to the general public. However, speech translation cannot be reduced to just speech recognition and text translation. Utterances that occur in unrehearsed, spoken dialogs are very different from written text. For such spontaneous input, speech recognition systems have a higher error rate and conventional text translation systems often fail due to ungrammaticalities, missing punctuation, and recognition errors. To make full use of all information present in the speech signal, a more integrated approach to speech translation is required.

The main section of the tutorial will cover common approaches and problems in speech recognition and speech translation. Some algorithms used in speech translation systems will be explained in more detail to provide a better understanding of the problems and possibilities. A description of past, present, and future speech translation systems with video and/or live demonstrations will round off the tutorial.

Multilingual Text Summarization

Eduard Hovy and Daniel Marcu
USC/Information Sciences Institute, Marina del Rey, CA, USA

After lying dormant for over two decades, automated text summarization has experienced a tremendous resurgence of interest in the past few years. This tutorial reviews the state of the art in automatic summarization, with particular emphasis on multilinguality, to the extent this has been addressed to date. The tutorial begins by outlining the major types of summary, then describes the typical decomposition of summarization into three stages, and explains in detail the major approaches to each stage. Next, discussion turns to the difficult issue of evaluation—measuring how good a summary is. Finally, we will outline the major open problems and research challenges that remain to be solved.

Panel Descriptions

A Seal of Approval for MT Systems

Eduard Hovy (moderator)

MT may not yet be a household term, but it is rapidly moving in that direction. Thanks to the World Wide Web, the need for occasional translation in the home or the monolingual workplace is increasing. The presence of translation systems on the Web—especially when partnered with Web access engines—makes it increasingly easy for the general public to access them.

Unfortunately, the occasional user is hardly likely to understand what MT is all about, and is in no position to distinguish the good MT from the bad. And it is quite likely that bad MT—not just low-quality output, but dishonorable business practice such as misleading advertising and worse—may proliferate.

The traditional hurdles that weeded out armchair MT and ensured that only serious MT practitioners eventually made it in the business are falling away. This will happen even more if current research in natural language processing succeeds in making it relatively easy to build a quick-and-dirty MT system with little effort, using web-based resources, in only a few months or even weeks.

What can we, as a community, do?

One response is to create a Seal of Approval, a tangible marker that indicates to the public at large that any software so marked has been recognized by AMTA or the IAMT. Though such a seal may not have any legal force, it could be registered with consumer watchdog agencies such as Consumer Reports and Better Business Bureaux, and could set an example for other software products in general.

The experts invited to this panel will discuss the feasibility of a Seal of Approval. Many perplexing questions must be solved, including:

- What are the criteria for approval? Do they apply only to the software, or also to the company creating and/or selling it?
- Who administers the tests? How frequently?
- Should there be only one seal, or various levels or ratings within the seal? How would a multi-level rating scheme differ from a normal evaluation?
- How do we prevent unscrupulous companies from simply using the seal?

Panelists (at time of printing):
Eduard Hovy (USC/ISI), president, AMTA (chair)
John Hutchins (University of East Anglia), president, EAMT
Bente Maegaard (CfS, Copenhagen), MT evaluation expert
L. Chris Miller (MCS, Washington), PC MT expert
Reba Rosenbluth (SYSTRAN Inc.), MT seller
Hozumi Tanaka (TITech), representing president IAMT and AAMT
Muriel Vasconcellos, past president, IAMT and AMTA
John White (Litton PRC), MT evaluation expert

The Forgotten Majority: Neglected Languages

Laurie Gerber (moderator)

The number of the world's languages spoken today is estimated to be between 3000 and 8000. Machine translation efforts have focused on a small set of these, many of them closely related Indo-European languages. Commercial MT systems exist for only 30–40 languages as a source or target, with the vast majority focusing on fewer than 10 languages. Why so few?
This panel seeks to address:

1. Economics of MT development: What conditions are necessary to justify the investment in MT for minority languages?
2. Resource constraints: Can methods be devised to handle new languages cheaply and quickly, despite the lack of electronic (or even printed) resources?
3. Theoretical considerations: Will work on new language families expose gaps in MT theory?
4. The role of public policy in language engineering: If "HLT is the key that can open the door to a true multilingual society," what are the risks of marginalization and isolation for language groups that are not included?
5. Linguistic minorities within societies: Can MT help meet the legal and administrative needs of states with large immigrant communities?
6. The role of MT preservation of endangered languages: Can loss of languages be slowed with MT?

Panelists (at time of printing):
Laurie Gerber (SYSTRAN Software Inc.), chair
Scott Bennett (LOGOS Inc.)
Denis Gachot (SYSTRAN Software Inc.)
Sergei Nirenberg (Computing Research Laboratory, NMSU)
John O'Hara (Globalink Inc.)
Harold Somers (UMIST)
Peter Wilkniss (CATANAL Project)

Breaking the Quality Ceiling

David Farwell (moderator)

Recent advances in MT have perhaps enabled us to acheive large-scale, low-quality MT faster, but they have failed to improve the quality of translation, say, for the purpose of dissemination.

In the session, panelists representing different approaches to MT will discuss the sorts of barriers their approaches must overcome to achieve high (or at least much improved) quality, and identify potential breakthroughs and how far they will take us. Specifically, each panelist will outline the basic characteristics of a core approach to MT, describe the associated quality pitfalls and suggest ways to overcome those pitfalls. In passing, participants may directly address the question of what quality is and, possibly, discuss various various 'hybridizations' of approaches in relation to overcoming pitfalls.

Panelists (at time of printing):
David Farwell (Computing Research Laboratory, NMSU), chair
Christian Boitet (University of Grenoble), structural-transfer-based MT
Sergei Nirenburg (Computing Research Laboratory), knowledge-based MT
Harold Somers (UMIST), example-based MT
Dekai Wu (Hong Kong University of Science and Technology), statistical MT