Preface

Envisioning Machine Translation in the Information Future

When the organizing committee of AMTA-2000 began planning, it was in that brief moment in history when we were absorbed in contemplation of the passing of the century and the millennium. Nearly everyone was comparing lists of the most important accomplishments and people of the last 10, 100, or 1000 years, imagining the radical changes likely over just the next few years, and at least mildly anxious about the potential Y2K apocalypse.

The millennial theme for the conference, "Envisioning MT in the Information Future," arose from this period. The year 2000 has now come, and nothing terrible has happened (yet) to our electronic infrastructure. Our musings about great people and events probably did not ennoble us much, and whatever sense of jubilee we held has since dissipated. So it may seem a bit obsolete or anachronistic to cast this AMTA conference into visionary themes.

But the millennial concepts remain pertinent to MT because of what it is and what will be expected of it. Like the printing press, that archetypal breakthrough invention of the last millennium, MT will make information available to everyone, breaking open the language-bound cloisters of ideas. Like printing and publishing, MT has and will evolve from a tool that can only be calibrated and operated by skilled people to one which anyone can directly operate, even while the demand for the professionally developed product continues to grow. Like the printing press, in short, MT will create its own demand, and will go from a capability that we never thought we would need to one we cannot do without.

This process has begun. The papers in this volume capture the state of MT in the year 2000, and they will continue to be of value for researchers, developers, users, translators, and information consumers for many years to come. They cover breakthrough approaches to the science of knowledge representation, statistical modeling, interlinguas and transfer strategies, and deployment of systems. They express the ingenious application of MT systems and techniques to the demands of actual translation environments, and the collection and reuse of corpora. They delve into the visions of future needs, programs, and expectations, along with the means by which we will evaluate change.

The AMTA-2000 Program Committee deserves the credit for capturing the essence of the state of MT at the turn of the millennium. The members of the Program Committee are:

Jeff Allen, Softissimo
Robert Cain, Foreign Broadcast Information Service
Gary Coen, Boeing Phantom Works
Jennifer DeCamp, MITRE Corp.
Jennifer Doyon, Litton PRC
Ulrich Germann, University of Southern California Information Science Institute
Stephen Helmreich, New Mexico State University Computing Research Laboratory
Doug Jones, National Institute of Standards and Technology

Kevin Knight, University of Southern California Information Science Institute
Marjorie Leon, Pan American Health Organization
Dan Loehr, MITRE Corp.
Jackie Murgida, Lernout & Hauspie
Kathryn Taylor, Georgetown University

My thanks also go to the organizers of AMTA-2000: Ed Hovy, Muriel Vasconcellos, Laurie Gerber, and Dave Farwell, who picked up my sundry dropped balls, helped with reviews, and set the tone for a successful conference and proceedings volume. The venue of the conference, arranged masterfully by Muriel Vasconcellos, Marina Urquidi, and Nena Uranga, is the ancient and beautiful city of Cuernavaca, Mexico. The tutorials and workshops were organized by Laurie Gerber, who built balanced programs evocative of the issues on the edge of MT in this new century. Kimberly Kellogg Belvin continues in her polished, professional role as exhibits coordinator. My good friends Ed Hovy (AMTA President) and Dave Farwell (AMTA-2000 Conference Chair) have provided much needed support, vision, and hortatory expressions to stimulate the development of the program represented in this volume.

I wish to thank especially Florence Reeder of MITRE Corporation, whose command of the quaint art of LaTeX stitchery made the assembly of this volume possible for me, who had heretofore presumed that WYSIWYG word processing was, like indoor plumbing, an ordinary expectation.

Whatever the vision of the future holds, whether ubiquitous information access, information appliances that we wear, or whole new metaphors of what it means to communicate, two things should be clear: our predictions will be wrong (including this one), and variation in human language will remain. In this light, I hope that both the present and future reader of this volume will benefit from these papers, both in the context of today and across the changes that the future will have brought to the field of machine translation.

August, 2000 John S. White

Tutorial Descriptions

Ontological Semantics

Sergei Nirenburg

Computing Research Laboratory, New Mexico State University

In computational linguistics the term ontology has come to denote a world model used for specifying the meaning of lexical units in a language. Elements of the ontology, thus, can be viewed as the lexis of a metalanguage for describing the lexical semantics of a particular language. Once the ontological approach is chosen for describing lexical meaning, the lexicon and the ontology become coupled. Depending on the type of computational linguistic application that a lexicon is supposed to support, the ontology that underlies its semantic component will contain different (though possibly compatible) information. Among the possible applications are: knowledge-based machine translation (MT); lexical disambiguation as a module in transfer-based MT or in an information extraction (IE) system; text summarization; human-computer interaction; planning and plan recognition for a society of software and human agents; object and scene recognition; and others. To illustrate the application-oriented differences in ontology content, the work on agents requires detailed statements about "workflow scripts" that these agents follow as well as domain-related plans, both realizable as complex events, while the work on knowledge-based MT typically does not. Lexical disambiguation is often considered feasible without ontological underpinnings in the lexicon but based on a set of semantic features assigned to a lexical item (if not based entirely on corpus-based co-occurrence calculations).

The application on which we will concentrate will be knowledge-based MT. In the framework of knowledge-based MT, ontology supplies major chunks of the metalanguage not only for the semantic component of the lexicon but also for the language in which the meaning of texts is represented. The latter language (called TMR, for **T**ext **M**eaning **R**epresentation) is the interlingua in the KBMT system.

The tutorial will include the following topics:

Design of an MT system based on ontological semantics

The Static Knowledge Sources for KBMT: the TMR, the Ontology and the Lexicon

The TMR

a) The TMR content

b) The TMR format

The Ontology

c) The syntax of the ontology entry

d) The content of the ontology

e) A brief comparison with other ontologies used for language processing, notably, CYC, WordNet and Sensus.

Ontology Acquisition

a) The acquisition methodology

b) Examples of concept acquisition

The Lexicon
a) The analysis lexicon
b) The generation lexicon
c) The onomasticon

Lexicon Acquisition
a) The acquisition methodology
b) Examples of lexicon entry acquisition

Interaction among the TMR, the Ontology and the Lexicon in Mikrokosmos
Ontological support for **applications other than MT** (IE, summarization, agents).

The tutorial is intended for computational lexicographers, designers and implementers of NLP systems, including MT, IE, IR, and text summarizers.

A Gentle Introduction to MT: Theory and Current Practice

Eduard Hovy
Information Sciences Institute of the University of Southern California

This tutorial provides a non-technical introduction to machine translation. It reviews the whole scope of MT, outlining briefly its history and the major application areas today, and describing the various kinds of MT techniques that have been invented---from direct replacement through transfer to the holy grail of interlinguas. It briefly outlines the difficult questions of MT evaluation and provides an introduction to the newest statistics-based techniques (which are the topic of another tutorial).

Topics include:

- History and development of MT
- Theoretical foundations of MT
- Traditional and modern MT techniques
- Latest MT research
- Thorny questions of evaluating MT systems

Eduard Hovy is the director of the Natural Language Group at the Information Sciences Institute of the University of Southern California, and is a member of the Computer Science Departments of USC and of the University of Waterloo. His research focuses on machine translation, automated text summarization, automated question answering, multilingual information retrieval, and the semi-automated construction of large lexicons and terminology banks. He is the author or editor of four books and over 100 technical articles. Currently Dr. Hovy serves as the President of the Association of Machine Translation in the Americas (AMTA) and as Vice President of the ACL and as President-Elect of the International Association of Machine Translation (IAMT). Dr. Hovy regularly co-teaches a course in the new Master's Degree Program in Computational Linguistics at the University of Southern California, as well as occasional short courses on MT and other topics at universities and conferences.

Controlled Languages

Teruko Mitamura and Eric Nyberg
Carnegie Mellon University

The notion of Controlled Language (CL) is becoming increasingly important for both authors and translators working a large-scale document production environment. Good design, process and implementation of a Controlled Language can provide higher-quality documentation and more productive translation. Even so, there are some issues associated with introducing Controlled Language into document production environment which must be considered carefully. The goal of this tutorial is to introduce the concept of Controlled Language, to discuss design and deployment issues, and to summarize the state of the art in CL development.

Intended audience: MT users, Authors, Translators, anyone who would be interested in learning about CL.

- Introduction
  - What is Controlled Language?
  - Goals of Controlled Language
- History of Controlled Language & Applications
  - Human Communications
  - Document Authoring
  - Document Translation
- Designing a Controlled Vocabulary
- Designing a Controlled Grammar
- How To Build and Deploy a Controlled Language
  - For authoring only
  - For authoring and MT
- Evaluating the Use of Controlled Language
  - Author's Perspective
  - Translator's Perspective
  - Developer's Perspective
- Current Status of Controlled Language
- The Future of Controlled Language

Statistical Machine Translation

Kevin Knight
Information Sciences Institute of the University of Southern California

The statistical approach to machine translation (MT) seeks to extract translation knowledge automatically from online bilingual texts (e.g., publications of the Canadian or Hong Kong governments). This idea can be traced back to suggestions

made by Warren Weaver in the 1940s. It was pioneered at IBM in the 1990s and continues to be inspired by relative successes in statistical speech recognition. We will present an accessible but technical tutorial that will cover the statistical MT literature to date. We will use graphical influence diagrams to explain statistical translation models used in different research projects around the world. We will also cover language models and "decoding" algorithms that perform online translations.

Outline:

- Introduction
  - History of statistical MT
  - Substitution ciphers, light probability, noisy channel framework
  - Substitution ciphers, light probability, noisy channel framework
  - Transliteration: a case study of MT as codebreaking
  - Sketch of a complete statistical MT system (training/translation modules)
- Building Blocks
  - Acquisition and cleaning of training data
    - monolingual and bilingual text corpora
    - sentence alignment
    - preprocessing
    - comparable text corpora
  - Language modeling and training
    - ngram models and smoothing
    - structured models
  - Translation modeling and training
    - word-internal translation models
    - word-for-word replacement and transposition models
    - phrase-for-phrase replacement and transposition models
    - tree-based models
  - Online translation ("decoding")
  - computational complexity and heuristics
  - word-for-word models, phrase-for-phrase models, tree-based models
- Assessment
  - Empirical results: does it work?
  - Strengths and weaknesses of statistical MT
  - Related applications
  - Immediate and long-term prospects
- Resources
  - Available software and text corpora
  - Full bibliography

The Diversity and Distribution of Languages

Laurie Gerber
Information Sciences Institute of the University of Southern California

Funding agencies and the market are placing greater emphasis on less common languages. Rapid response and short development times are crucial as economic or political events bring diverse regions and their languages to the front of the international stage. However, most MT development groups have worked on a relatively small set of languages - namely Indo-European. Even where other languages are addressed, the frameworks and architectures within which such development takes place were only designed to cover this relatively homogeneous group. Can extensions to existing paradigms cover the full diversity of the world's estimated 6,000 languages? Is it possible to build a single architecture that can handle the full range of diversity? How weird does it get? And are there any regularities that can be exploited in tackling the great diversity we face?

Outline:

- Classification methods:
  - What constitutes a language?
  - morphological, genetic, and word-order classification systems
    - *types of morphology
    - language families and areal contact
    - word order tendencies
  - variations in POS inventory... "Do all languages have nouns?"
- What's out there?
  - Where are 6,000 languages hiding?
  - Regional distribution and frequency of typological traits
    - or, How often will I have to worry about polysynthetic morphology and other exotic pheonomena?
- How can I use this information?
  - - Are there any useful universals?
  - What statistical tendencies and implicational universals can help in designing NLP systems?

MTranslatability

Arendse Bernth and Claudia Gdaniec
IBM T.J. Watson Research Center

Current MT systems are often unable to produce high-quality output on arbitrary, unseen input. The output frequently does not meet user needs and requirements. We

will address some of the reasons for the unsatisfactory quality of MT output, ways to improve translatability, and ways to measure the translatability of a document.

Intended audience: MT users and consultants, people in charge of information development.

Presenters: Arendse Bernth & Claudia Gdaniec, IBM T.J. Watson Research Center. The presenters have worked in the MT field for many years. Both have also worked on MT-related tools -- for pre-editing, and for automatically estimating the quality of MT output.

Outline

- Introduction
    - Why is MT output not better?
    - What aspects can the MT user control?
    - Is it possible to predict the output quality for given input automatically?
- Ways to Improve Translatability
    - Grammar Checkers
    - Controlled Language Checkers
- Other Helpful Tools
- Ways to Measure Translatability
    - Automatic readability scoring
    - Automatic detection of lexical inadequacies
    - Automatic MTranslatability scoring
- Conclusion
- Discussion of a Special Interest Group on Translatability