Stephen D. Richardson (Ed.)

LNAI 2499

Machine Translation: From Research to Real Users

5th Conference of the Association for Machine Translation in the Americas, AMTA 2002, Tiburon, CA, USA, October 2002 Proceedings

Springer

Lecture Notes in Artificial Intelligence 2499

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Stephen D. Richardson (Ed.)

Machine Translation: From Research to Real Users

5th Conference of the Association for Machine Translation in the Americas, AMTA 2002
Tiburon, CA, USA, October 8 – 12, 2002
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editor

Stephen D. Richardson
Microsoft Research
1 Microsoft Way, Redmond, WA 98052, USA
E-mail: steveri@microsoft.com

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Machine translation: from research to real users : Tiburon, CA, USA, October 8 - 12, 2002 ; proceedings / Stephen D. Richardson (ed.). - Berlin ; Heidelberg ; New York ; Hong Kong ; London ; Milan ; Paris ; Tokyo : Springer, 2002
(... Conference of Association for Machine Translation in the Americas, AMTA ... ; 5)
(Lecture notes in computer science ; Vol. 2499 : Lecture notes in artificial intelligence)
ISBN 3-540-44282-0

CR Subject Classification (1998): I.2.7, I.2, F.4.2-3, I.7.1-3

ISSN 0302-9743
ISBN 3-540-44282-0 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York,
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by Markus Richter, Heidelberg
Printed on acid-free paper SPIN: 10870740 06/3142 5 4 3 2 1 0

Preface

AMTA 2002: From Research to Real Users

Ever since the showdown between Empiricists and Rationalists a decade ago at TMI 92, MT researchers have hotly pursued promising paradigms for MT, including data-driven approaches (e.g., statistical, example-based) and hybrids that integrate these with more traditional rule-based components.

During the same period, commercial MT systems with standard transfer architectures have evolved along a parallel and almost unrelated track, increasing their coverage (primarily through manual update of their lexicons, we assume) and achieving much broader acceptance and usage, principally through the medium of the Internet. Webpage translators have become commonplace; a number of online translation services have appeared, including in their offerings both raw and postedited MT; and large corporations have been turning increasingly to MT to address the exigencies of global communication. Still, the output of the transfer-based systems employed in this expansion represents but a small drop in the ever-growing translation marketplace bucket.

Now, 10 years later, we wonder if this mounting variety of MT users is any better off, and if the promise of the research technologies is being realized to any measurable degree. In this regard, the papers in this volume target responses to the following questions:

- Why aren't any current commercially available MT systems primarily data-driven?
- Do any commercially available systems integrate (or plan to integrate) data-driven components?
- Do data-driven systems have significant performance or quality issues?
- Can such systems really provide better quality to users, or is their main advantage one of fast, facilitated customization?
- If any new MT technology could provide such benefits (somewhat higher quality, or facilitated customization), would that be the key to more widespread use of MT, or are there yet other more relevant unresolved issues, such as system integration?
- If better quality, customization, or system integration aren't the answer, then what is it that users really need from MT in order for it to be more useful to them?

The contributors to this volume have sought to shed light on these and related issues from a variety of viewpoints, including those of MT researchers, developers, end-users, professional translators, managers, and marketing experts. The jury appears still to be out, however, on whether data-driven MT, which seems to have meandered along a decade-long path of evolution (instead of revolution, as many thought it might be), will lead us to the holy grail of high-quality MT. And yet, there is a sense of progress and optimism among the practitioners of our field.

I extend my sincere thanks to the members of the AMTA 2002 program committee, who sacrificed time and effort to provide detailed analyses of the papers submit-

ted to the conference. Many of the authors expressed gratitude for the insightful and helpful comments they received from the reviewers as they prepared their papers for publication.

Many thanks also go to the organizers of AMTA 2002 who spent untold hours to ensure the success of the conference:

Elliott Macklovitch, General Chair
Violetta Cavalli-Sforza, Local Arrangements Chair
Robert Frederking, Workshops and Tutorials
Laurie Gerber, Exhibits Coordinator
Jin Yang, Webmaster
Debbie Becker, Registrar

In particular, I am grateful to Elliott and Laurie, who provided me with a constant and sustaining flow of guidance and wisdom throughout the conception and assemblage of the program.

Final and special thanks go to Deborah Coughlin, who assisted me in managing the submissions to the conference and in overseeing all substantial aspects of the production of this volume, and to my other colleagues at Microsoft Research, who supported us during this process.

August 2002 Stephen D. Richardson

Program Committee

Arendse Bernth, IBM T.J. Watson Research Center
Christian Boitet, Pr. Université Joseph Fourier, GETA, CLIPS, IMAG
Ralf Brown, Carnegie Mellon University Language Technologies Institute
Robert Cain, MT Consultant
Michael Carl, Université de Montréal, RALI
Bill Dolan, Microsoft Research
Laurie Gerber, Language Technology Broker
Stephen Helmreich, New Mexico State University Computing Research Laboratory
Eduard Hovy, University of Southern California Information Science Institute
Pierre Isabelle, Xerox Research Centre Europe
Christine Kamprath, Caterpillar Corp.
Elliott Macklovitch, Université de Montréal, RALI
Bente Maegaard, Center for Sprogteknologi
Michael McCord, IBM T.J. Watson Research Center
Robert C. Moore, Microsoft Research
Hermann Ney, RWTH Aachen
Sergei Nirenburg, New Mexico State University Computing Research Laboratory
Franz Och, RWTH Aachen
Joseph Pentheroudakis, Microsoft Research
Jessie Pinkham, Microsoft Research
Fred Popowich, Gavagai Technology Inc.
Florence Reeder, MITRE Corp.
Harold Somers, UMIST
Keh-Yih Su, Behavior Design Corp.
Eiichiro Sumita, ATR
Hans Uszkoreit, Saarland University at Saarbrücken, DFKI
Lucy Vanderwende, Microsoft Research
Hideo Watanabe, IBM Tokyo Research Laboratory
Andy Way, Dublin City University
Eric Wehrli, University of Geneva
John White, Northrop Grumman Information Technology
Jin Yang, SYSTRAN
Ming Zhou, Microsoft Research

Tutorial Descriptions

Example-Based Machine Translation

Ralf Brown

Language Technologies Institute
Carnegie Mellon University
ralf@cs.cmu.edu

1 Description

This tutorial will introduce participants to the history and practice of example-based machine translation (EBMT). After a definition of EBMT and an overview of its origins (Sato and Nagao, among others), various types of approaches to example-based translation (such as deep versus shallow processing) will be presented. This discussion will lead into a overview of a number of recent example-based systems, both "pure" and hybrid systems combining rule-, statistics-, or knowledge-based approaches with EBMT. Candidates for discussion include EGDAR, Gaijin, ReVerb, and systems by Cranias, Guevenier/Cicekli, and Streiter. Finally, the tutorial will conclude with a more in-depth examination of the Generalized EBMT system developed at Carnegie Mellon University.

2 Outline

- Introduction: Example-Based Translation's definition and origins
- EBMT and its relation with other translation technologies
  - the Vaquois diagram and "depth" of processing
  - "shallow" EBMT and translation memories
  - "deep" EBMT and transfer-rule systems
  - relationship between EBMT and statistical MT
- Overview of EBMT Systems
  - EDGAR
  - Gaijin
  - ReVerb
  - etc.
- Hands-On Exercise in EBMT
- Carnegie Mellon University's Generalized EBMT system

- simple matching against an example base
- generalizing the examples into templates
- learning how to generalize
- inexact matching
- use as an engine in the Multi-Engine MT architecture

3 Biographical Information

Ralf Brown has been working on Example-Based Translation since 1995, using it in various applications such as the translation component of a speech-to-speech translation system and for document translation for event tracking in news streams. He received his Ph.D. in Computer Science from Carnegie Mellon University in 1993, where he is currently research faculty in the Language Technologies Institute.

Units of Meaning in Translation – Making Real Use of Corpus Evidence

Pernilla Danielsson (in co-operation with Prof Wolfgang Teubert)

Centre for Corpus Linguistics
Department of English
University of Birmingham
Birmingham B15 2TT
United Kingdom
Web-address: www.english.bham.ac.uk/ccl
pernilla@ccl.bham.ac.uk

1 Description

Birmingham's Centre for Corpus Linguistics offers to give a tutorial on how to use large corpora in language research, especially translation. This tutorial will focus on meaning. In modern corpus linguistics, we work from the hypothesis that meaning is in its use, as previously stated by researchers such as Terry Winograd or Wittgenstein in his 'Philosophical Investigations' stating that 'the meaning of a word is in its use'. This may be further interpreted into the statement that meaning is in text, which opposes the idea that meaning is constructed in the human brain. From a research point of view, this is a very important statement since it will remove the difficulties of trying to model human brains in order to interpret a text and instead gear us towards finding new methods of interpreting the complex systems governing the interpretation of texts. This tutorial will show how by carefully examining large corpora, meaning can emerge through patterning.

Units of meaning are often larger and more complex than the simple word. Most units of translation are compounds, collocations or even phrases. As for single words, most of them are

ambiguous. The participants will be shown methods to disambiguate words by investigating their contextual profiles. The tutorial will also focus on retrieving translation equivalents and learning how corpus data can help us produce translated texts that display the 'naturalness' of the target language.

2 Outline

The tutorial will cover the following topics:

- Working with Large Corpora
- Units of Meaning
- Translation Units in Parallel texts
- Contextual Profiles

The tutorial will be divided into two sessions in order to cover both monolingual and multilingual corpus methodologies. The first session is focused on the use of methods in monolingual corpus linguistics for extracting information about the languages in question. As a demonstration, we will make use of the Bank of English, a 450 million word corpus co-owned by the University of Birmingham and the publishing house HarperCollins. Once Units of Meaning have been established in the monolingual corpus we will move on to parallel texts. By using the newly discovered units of meaning, we will now discover that words are not as ambiguous as they first seem, at least not when treated within larger units.

3 Biographical Information

Dr. Pernilla Danielsson is the Deputy Director at the Centre for Corpus Linguistics, University of Birmingham. In 2000 she became the new Project Manager for the EU-funded project TELRI-II (Trans European Language Resources Infrastructure). With a background in Computational Linguistics, she is now focusing her research on the study of units of meaning in corpora.

Prof. Wolfgang Teubert (PhD Heidelberg 1979) was, until 2000, a senior research fellow at the Institut für Deutsche Sprache (IDS), Mannheim, Germany. In 2000, he was appointed to the Collins Chair of Corpus Linguistics, Department of English at the University of Birmingham. The focus of his research is the derivation of linguistic metadata from digital resources, particularly in multilingual environments with the emphasis on semantics. His other interest is the application of the methodology of corpus linguistics to critical discourse analysis. He is also the editor of the International Journal of Corpus Linguistics.

Supporting a Multilingual Online Audience

Laurie Gerber

On Demand Translation
61 Nicholas Road
Suite B3
Framingham, MA 01701
508-877-3430
lgerber@on-demand.biz

1 Description

The need to provide customer service and technical support to non-English-speaking customers is growing rapidly: Internet users are increasingly multilingual. IDC has reported that the number of Internet users in Western Europe surpassed users in the U.S. at the end of 2001. The trend is expected to continue, and English will eventually become a minority language on the Internet. The growth of the Internet overseas in turn tends to fuel sales of hardware and software applications. But companies that release localized products outside of the U.S. are often unprepared to fully support their increasingly diverse customer base. Simply translating web sites is rarely adequate. Customer communications are conducted through a variety of channels, and may contain very different types of text, speech and data. Translation products and services abound, but understanding which solutions can effectively address a specific need requires an understanding of the problem, as well as the solutions.

This tutorial will provide participants with:

- Practical skills for analyzing language support requirements
- Strategies for selecting and deploying appropriate language solutions
- An understanding of the range and capabilities of language technologies
- Knowledge of how to integrate language technologies into organizational workflows while avoiding common pitfalls
- Suggestions for measuring the effectiveness of your multilingual customer support

2 Outline

- Language products, services, solutions and their applications
- Assessing multilingual communication needs
    - Source materials factors
    - Target materials factors

  - Delivery factors
  - Organizational factors
  - Cost factors
  - Implementation factors
- Integrating multiple language technologies to work together
- Preparing your content
- Communicating with customers about language technology
- Measuring results

3 Biographical Information

With a background in Asian languages, Laurie Gerber was a central figure in SYSTRAN Software's Chinese-English and Japanese-English machine translation development efforts from 1986 to 1998 and served as Director of R&D from 1995 through 1998. Also in contact with users, Ms. Gerber developed a strong interest in usability issues for language technology. After earning a Master of Computational Linguistics from the University of Southern California in May 2001, she now works as an independent consultant on language technology implementation, and business development for commercializable prototype language technologies. She is currently Vice President (2000-2002) of the Association for Machine Translation in the Americas, and editor of Machine Translation News International, the newsletter of the International Association for Machine Translation.

The State of the Art in Language Modeling

Joshua Goodman

Machine Learning and Applied Statistics Group
Microsoft Research
http://www.research.microsoft.com/~joshuago
JoshuaGo@microsoft.com

1 Description

The most popular approach to statistical machine translation is the source-channel model; language models are the "source" in source-channel. Language models give the probability of word sequences. In a machine translation system they can assist with word choice or word order-- "The flesh is willing" instead of "The meat is willing" or “Flesh the willing is.” Most use of language models for statistical MT has been limited to models used by speech recognition systems (trigrams), despite the existence of many other techniques. In addition, many language modeling techniques could be adapted to improve channel models, or other parts of MT systems.

This tutorial will cover the state of the art in language modeling. The introduction will include what a language model is, a quick review of elementary probability, and applications of language modeling, with an emphasis on statistical MT. The bulk of the talk will describe current techniques in language modeling, including techniques like clustering and smoothing that are useful in many areas besides language modeling, and more language-model specific techniques such as high order n-grams and sentence mixture models. Finally, we will describe available toolkits and corpora.

A portion of the material in this talk was developed by Eugene Charniak.

2 Outline

- Introduction quickly reviewing key concepts in probability needed for language models, and then briefly the source-channel model for machine translation. Outline of remainder of talk describing specific language modeling techniques.

- Smoothing addresses the problem of data sparsity: there is rarely enough data to accurately estimate the parameters of a language model. Smoothing gives a way to combine less specific, more accurate information with more specific, but noisier data. I will describe two classic techniques -- deleted interpolation and Katz (or Good-Turing) smoothing -- and one recent technique, Modified Kneser-Ney smoothing, the best known.

- Caching is a widely used technique that uses the observation that recently observed words are likely to occur again. Models from recently observed data can be combined with more general models to improve performance.

- Skipping models use the observation that even words that are not directly adjacent to the target word contain useful information.

- Sentence-mixture models create separate models for different sentence types. Modeling each type separately improves performance.

- Clustering is one of the most useful language modeling techniques. Words can be grouped together into clusters through various automatic techniques; then the probability of a cluster can be predicted instead of the probability of the word.

- There are many recent, but more speculative language modeling techniques, including grammar-based language models, maximum entropy models, and whole sentence maximum entropy models.

- Finally, I will also talk about some practical aspects of language modeling. I will describe how freely available, off-the-shelf tools can be used to easily build language models, where to get data to train a language model, and how to use methods such as count cutoffs or relative-entropy techniques to prune language models.

Those who attend the tutorial should walk away with a broad understanding of current language modeling techniques, and the background needed to either build their own language models, or to apply some of these techniques to machine translation.

3 Biographical Information

Joshua Goodman is a Researcher at Microsoft Corporation. He previously worked on speech recognition at Dragon Systems. He received his Ph.D. in 1998 from Harvard University for work on statistical natural language processing. He then worked in the speech group at Microsoft Research, with a focus on language modeling. Recently, he moved to the Microsoft Research Machine Learning and Applied Statistics group, where he has worked on probabilistic models for natural language tasks, such as grammar checking.

Beyond the Gist:
A Post-Editing Primer for Translation Professionals

Walter Hartmann

MTConsulting Co.
wh@mtconsult.com

1 Description

To keep up with ever-increasing demands on output and turn-around, the professional translator needs tools that help increase productivity. Among others, machine translation stands out as a timesaving tool that can greatly enhance translation output. It does take, however, preparation and practice to weave machine translation efficiently into the workflow.

This tutorial addresses the various topics to be considered when contemplating the integration of machine translation into the workflow for publication-ready translations.

2 Outline

The main topics to be covered are:

- Introduction: MT and the professional translator
- Evaluation techniques
  - Application variations
  - Analysis of MT output for a given task
  - Remedies: Pre-editing, dictionary updates, post-editing, when to pass on post-editing

- Editing techniques – hands-on practice using texts translated from various languages into English

3 Biographical Information

MT has been a valuable productivity tool for the presenter since 1985, when he began to post-edit publication-level texts for customers. When MT programs became available for PCs, he pioneered MT integration into the workflow of translation companies. In subsequent years, he integrated machine translation for high-volume translations for several other companies. Even today, as a part-time freelance translator, the presenter actively uses MT as a productivity tool.

MT Evaluation: The Common Thread

John S. White[1] and Florence Reeder[2]

[1]Northrop Grumman Information Technology
McLean, VA USA
White_john@prc.com

[2]The MITRE Corporation
McLean, VA USA
freeder@mitre.org

1 Description

MT Evaluation is central to the goals of both the research and product worlds, and so can form a common, uniting thread among the two. Useful, comparable measures are difficult to produce in both spheres, despite the lessons from the 1960's. This is due in part to the uniquely difficult aspects of evaluating translation in general. This tutorial sets out to explain some issues, structures, and approaches that will demystify evaluation, and make it possible to design and perform meaningful evaluations with a minimum of time and resource.

The tutorial will cover the topics of the difficulty of MT evaluation, and then presents different views (including the recent work from the ISLE project) on the stakeholders, uses, and types of MT, and the attributes, measurands, and metrics implied by these perspectives. We will present a number of historical methods which may have renewed usefulness in today's context, as well as some approaches over the last decade and new approaches modeled in the last year.

In particular we will address the potential for automatic evaluation of MT. The multiplicity of uses, users, and approaches to MT have traditionally made this a presumptively impossible goal for all of the evaluation metrics. However, new research in evaluation has shed light on the potential for the automatic prediction of certain key attributes of MT output which can be used to predict more general performance. Among the promising approaches for automation include capturing the perspectives of language learning; modeling "machine English" and human English; and predicting fidelity from intelligibility.

The tutorial will provide the participant with perspectives, tools, and data for determining the usefulness of particular approaches in particular MT contexts. The tutorial will be quite interactive, with exercises and MT data to help the participant understand the challenges and potential for MT evaluation.

2 Outline

- Evaluation overview
- What's hard about MT evaluation
- The stakeholders, purposes, and types of MT evaluation
    - Methods related to purposes
    - ISLE classification
- Some famous evaluation methods, old and new
- The search for automatic MT Evaluation
    - New methods
    - Rationales and Bases
    - Interactive Experiments
- Conclusions

3 Biographical Information

John White is Director of Independent Research and Development for Defense Enterprise Solution, Northrop Grumman Information Technology. In this capacity he is responsible for research and development initiatives in language systems evaluation, information assurance, software agent technology, modeling/simulation, collaborative computing, and imaging.

White holds a Ph.D. in Linguistic Anthropology from The University of Texas, and is widely published in machine translation, evaluation, artificial intelligence, and information assurance.

Florence Reeder is an Associate Technical Area Manager with The Mitre Corporation. She works with a variety of U.S. government agencies in developing machine translation and foreign language handling systems which multiply scarce expertise in critical languages. Florence is also a doctoral student at George Mason University, and is currently completing her dissertation work in the field of MT evaluation.

Table of Contents

Technical Papers

User Studies

System Descriptions