Statistical Machine Translation

Kevin Knight & Philipp Koehn

{knight@ISI.EDU, pkoehn@inf.ed.ac.uk}

Abstract

Automatic translation is a long-standing goal of computer science. 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. For general-purpose translation, the amount of required knowledge is staggering. Over the past few years, there has been a fair amount of research into extracting translation knowledge automatically from bilingual texts, using statistical modeling techniques. This tutorial covers the state-of-the-art in those techniques.

Data for Statistical MT

- bilingual and monolingual corpora
- data acquisition, cleaning, and preparation
- how much data is necessary?

MT Evaluation

- manual and automatic evaluation
- BLEU, WER, NIST, TER, Meteor, plus
- counting required postedits, HTER
- return on investment (ROI)
- what is the best evaluation?

Extracting Translation Knowledge from Bilingual Text

- IBM Models 1-5 and HMM models, training
- word alignment and its evaluation
- phrase-pair extraction and counting
- other phrase-based features
- maximum entropy models, training
- what are the strengths and weaknesses of existing models?

Monolingual Target Language Models

- n-gram models and smoothing, multiple language models
- efficient storage in memory
- how does language model size and quality affect translation accuracy?

Decoding & Translation Algorithms

- features and weights
- search space of possible translation outputs
- navigating the search space, search errors
- searching techniques and heuristics, minimum Bayes risk
- how fast is statistical MT?

Discriminative Training

- small-scale and large-scale weight tuning
- how sensitive is MT accuracy to feature weights?

Syntactic Models

- synchronous grammars, treelets, and tree transducers
- syntactic algorithms for alignment
- syntactic translation models and language models
- decoding/parsing algorithms
- do linguistic categories help statistical MT?

Specialized Translation Components

- segmentation and morphology
- named entity translation, numbers and dates

Future of Statistical MT

Available Resources & References

- software tools and data sets
- bibliography

About the Presenters

Philipp Koehn received his PhD from the University of Southern California, where he was a research assistant at the Information Sciences Institute (ISI) from 1997 to 2003. He was a postdoctoral research associate at the Massachusetts Institute of Technology (MIT) in 2004, and joined the University of Edinburgh as a lecturer in 2005. His research centers on statistical machine translation, but he has also worked on speech in 1999 at AT&T Research Labs and text classification in 2000 at Whizbang Labs. Besides his research, his major contribution to the machine translation community are the preparation and release of the DE-News and Europarl corpora, the Pharaoh decoder, and the Moses toolkit.

Kevin Knight is a Research Associate Professor in the Computer Science Department at the University of Southern California, a Senior Research Scientist and Fellow at the USC/Information Sciences Institute, and Chief Scientist at Language Weaver. Dr. Knight received a PhD in computer science from Carnegie Mellon University in 1992, and a bachelor's degree from Harvard University. His research interests include machine translation, linguistic structure, automata theory, statistical modeling, language generation, and decipherment. He teaches at USC and led a six-week intensive workshop on statistical MT in the summer of 1999. Dr. Knight is also co-author (with Elaine Rich) of the widely-adopted textbook Artificial Intelligence. With collaborators at USC/ISI and elsewhere, he has authored over 50 papers on MT and natural language processing. .