

2. What's New in Statistical Machine Translation

Kevin Knight and Philipp Koehn, Information Sciences Institute

<http://www.isi.edu/~knight>

<http://www.isi.edu/~koehn>

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. Recently, there has been a fair amount of research into extracting translation-relevant knowledge automatically from large collections of manually-translated texts, and over the past years, several statistical MT projects have appeared in North America, Europe, and Asia, and the literature is growing substantially. We will overview this progress.

2.1. Tutorial Outline

1. Data for MT.
 1. Bilingual corpora: what's out there?
 2. Acquisition and cleaning.
 3. What does three million words really mean?
2. MT Evaluation.
 1. Manual and automatic.
3. Core Models and Decoders
 1. IBM Models 1-5 and HMM models, training, decoding.
 2. Word alignment and its evaluation.
 3. Phrase models.
 4. Syntax-based translation and language models.
4. Specialized Models.
 1. Named entity MT, numbers and dates, morphology, noun phrase MT.
5. Available Resources.
 1. Tools and data.
6. Bibliography

2.2. Target Audience

The target audience for this tutorial is anyone interested in machine translation of human languages.

Kevin Knight is a Senior Research Scientist at the USC/Information Sciences Institute and a Research Associate Professor in the Computer Science Department at USC. He has written a number of articles on statistical MT, plus a widely-circulated MT workbook (<http://www.isi.edu/natural-language/mt/wkbk.rtf>). Dr. Knight has given several invited talks on machine translation at recent AMTA and EMNLP conferences.

Philipp Koehn completed his Ph.D. in Computer Science at the University of Southern California in Fall 2003. He has written a number of articles on topics in statistical machine translation, including bilingual lexicon induction from monolingual corpora, word-level translation models, and translation with scarce resources. He has also worked at AT&T Laboratories on text-to-speech systems, and at WhizBang! Labs on text categorization.