Proceedings of HLT-NAACL
Tutorials , pg. 5
Edmonton, May-June 2003

What's New in Statistical Machine Translation

Kevin Knight and Philipp Koehn
USC/ISI

Tutorial Outline

1. Data for MT
   - bilingual corpora: what's out there?
   - acquisition and cleaning
   - what does three million words really mean?
2. MT Evaluation
   - manual and automatic
   - word error rate, BLEU, NIST measures
   - MT Evaluation versus MT
3. Core Models and Decoders
   - IBM Models 1-5 and HMM models, training, decoding
   - word alignment and its evaluation
   - alignment templates and phrase models
   - syntax-based translation and language models
   - weaknesses of existing models
   - maximum entropy models, training, decoding
4. Specialized Models
   - named entity MT
   - numbers and dates
   - morphology
   - noun phrase MT
5. Available Resources - tools and data

Abstract

Automatic translation from one human language to another using computers, better known as machine translation (MT), 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, and it is not clear how to prioritize knowledge acquisition efforts.

Recently, there has been a fair amount of research into extracting translation-relevant knowledge automatically from bilingual texts. In the early 1990s, IBM pioneered automatic bilingual-text analysis. A 1999 workshop at Johns Hopkins University saw a re-implementation of many of the core components of this work, aimed at attracting more researchers into the field. Over the past years, several statistical MT projects have appeared in North America, Europe, and Asia, and the literature is growing substantially. We will provide a technical overview of the state-of-the-art.