Statistical Machine Translation

Kevin Knight
USC/Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292, USA
knight@isi.edu

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 a technical, focused tutorial that will cover the statistical MT literature to date. This tutorial will not cover MT in the broad sense (transfer and interlingua approaches, evaluation, commercial products, etc.)—we will instead concentrate on statistical models proposed for the translation process, using accessible graphical influence diagrams to explain models used in different research projects around the world. We will also cover language models and "decoding" algorithms that perform online translations. The tutorial will be structured as follows:

- Introduction
    - History of statistical MT
    - 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
    - Language modeling and training
    - Translation modeling and training
    - Online translation ("decoding")
- Assessment
    - Empirical results: does it work?
    - Strengths and weaknesses of statistical MT
    - Related applications
    - Immediate and long-term prospects