Automata for Transliteration and Machine Translation

Kevin Knight
Information Sciences Institute
University of Southern California
knight@isi.edu

Abstract

Automata theory, transliteration, and machine translation (MT) have an interesting and intertwined history.

Finite-state string automata theory became a powerful tool for speech and language after the introduction of the AT&T's FSM software. For example, string transducers can convert between word sequences and phoneme sequences, or between phoneme sequences and acoustic sequences; furthermore, these machines can be pipelined to attack complex problems like speech recognition. Likewise, n-gram models can be captured by finite-state acceptors, which can be reused across applications.

It is possible to mix, match, and compose transducers to flexibly solve all kinds of problems. One such problem is transliteration, which can be modeled as a pipeline of string transformations. MT has also been modeled with transducers, and descendants of the FSM toolkit are now used to implement phrase-based machine translation. Even speech recognizers and MT systems can themselves be composed to deliver speech-to-speech MT.

The main rub with finite-state string MT is word re-ordering. Tree transducers offer a natural mechanism to solve this problem, and they have recently been employed with some success.

In this talk, we will survey these ideas (and their origins), and we will finish with a discussion of how transliteration and MT can work together.

Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009, page 27,
Suntec, Singapore, 7 August 2009.