Formal Models in NLP

Topics for Finite-State Methods

April 17, 2012
Deterministic part-of-speech tagging with finite-state transducers
Roche and Schabes, 1995

Brill’s rule-based part-of-speech tagger achieves best results by inferring rules from a training corpus. But the implementations of the tagger run more slowly than previous approaches. ($RKn$ elementary steps to tag an input of $n$ words with $R$ contextual rules requiring at most $K$ tokens of context).
Roche and Schabes present a finite-state tagger (based on the rule-based tagger) that operates in optimal time.

- Transform every extracted rule from the training corpus into a nondeterministic finite-state transducer.
- Turn the transducers into transducers that operate globally on the input.
- Combine all transducers into one single transducer.
- Determinize the single transducer.
A Short History of Two-Level Morphology
Karttunen and Beesley, 2001

Input: kaNpat
Rule1: \( N \rightarrow m / _p \)
Rule2: \( p \rightarrow m / m \_ \)
Output: kammat

- Rules are symbol-to-symbol constraints that are applied in parallel.
- Constraints can refer to the lexical context, to the surface context, or to both.
- Lexical lookup and morphological analysis are performed in tandem.
Problem: Representation of very large scale dictionaries.

Large dictionaries can be compiled into finite automata with distinct final states.

Mohri describes new methods for compiling dictionaries into $p$-subsequential transducers and compares the results of his experiments with those obtained with automata.

He shows two ways how to treat new states added to the transducer.
Generalized algorithms for constructing statistical LMs
Allauzen, Mohri and Roark, 2003

- **Counting**
  - LMs are based on the output of a SR system $\rightarrow$ word lattices.
  - LM is constructed by deriving the statistics for any given sequence from the lattices. BUT: even small automaton can have more than four billion paths.
  - Allauzen et al. compute the *expected count*.

- **Representation of language models by WFAs**
  - Classical $n$-gram LMs can easily be represented by a WFA.
  - But the size makes it impractical for offline optimizations.
  - The authors describe how to represent a $n$-gram LM by WFAs whose size is practical for offline use (vocabulary size about 500,000 words; $n = 6$).

- **Class-based models**
  - Classical class-based models are based on simple classes such as a list of words.
  - Allauzen et al. present a simple and more general approach to class-based LMs based on general weighted context-dependent rules (they can even deal with weighted regular languages).