6.7 Empirical aspects of finite-state language processing

Prins R.P., van Noord G.J.M.

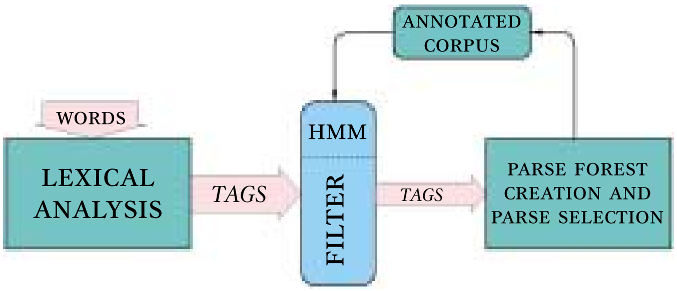
As computational grammars grow larger, increasing the number of syntactical constructions that can be recognized, they tend to become slower as well. At the same time ambiguity increases: the system will assign a larger number of analyses to one and the same utterance, making it harder to decide which analysis is the 'correct' one. We can reduce the inefficiency caused by grammar size and ambiguity by means of grammar approximation techniques. The research "Empirical Aspects of Finite-State Language Processing" is about using finite-state techniques to this end.

A finite-state automaton (FSA) is a conceptual machine that recognizes a language. At any moment the machine is in one of a finite number of states; reading in the next element from an input tape, it moves into another state if this transition is possible given a transition function. The input sequence is recognized as a string from the language defined by the FSA when all elements have been read and the FSA is in an accepting state. These automata can process input very quickly. Apart from the expected efficiency gain in practical applications, the FSA also shares characteristics with human language processing: a limited amount of available memory, problems with certain linguistic constructions such as center embedding, and linear processing time.

We approximated a computational grammar by means of a Hidden Markov Model (HMM), a simple kind of probabilistic FSA that can be used in computing the likeliness of a sequence of elements. The HMM was constructed on the basis of a large corpus, annotated by the computational grammar.

This HMM was implemented in a filter that speeds up the grammar by removing unlikely syntactic tags, assigned to words during lexical analysis (see figure 13). This way the system is able to reach at least the same level of accuracy as before, but the time needed is up to 20 times as short.

Figure 13 ▸


ANNOTATED CORPUS
WORDS
HMM
LEXICAL ANALYSIS
TAGS
FILTER
TAGS
PARSE FOREST CREATION AND PARSE SELECTION