Finite-State Machines for Mining Patterns
in Very Large Text Repositories

Wojciech Skut

Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA, 94043, USA
wwskut@gmail.com

The emergence of WWW search engines since the 1990s has changed the scale of many natural language processing applications. Text mining, information extraction and related tasks can now be applied to tens of billions of documents, which sets new efficiency standards for NLP algorithms. Finite-state machines are an obvious choice of a formal framework for such applications. However, the scale of the problem (size of the searchable corpus, number of patterns to be matched) often poses a problem even to well-established finite-state string matching techniques. In my presentation, I will focus on the experience gained in the implementation a finite-state matching library optimized for searching large amounts of complex patterns in a WWW-scale repository of documents. Both algorithmic and implementation-related aspects of the task will be discussed. The library is based on OpenFST. An open-source release of the code is planned in the near future.