OpenFst: An Open-Source, Weighted Finite-State Transducer Library and its Applications to Speech and Language

Introduction

Cyril Allauzen - allauzen@google.com
Martin Jansche - mjanchse@google.com
Michael Riley - riley@google.com

May 31, 2009

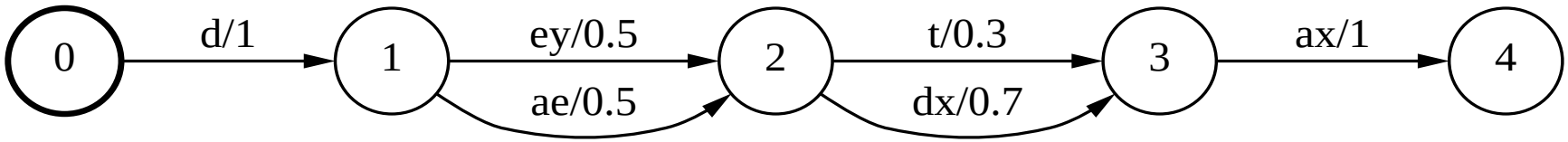
Thanks to Boulos Harb, Johan Schalkwyk, Masha Shugrina, Mehryar Mohri, Richard Sproat, and Wojtek Skut.

OpenFst Library

- C++ template library for constructing, combining, optimizing, and searching *weighted finite-states transducers (FSTs)*.
- Goals: Comprehensive, flexible, efficient and scale well to large problems.
- Origins: AT&T, merged efforts from Google and the NYU Courant Institute.
- Documentation and Download: http://www.openfst.org
- Released under the Apache license.

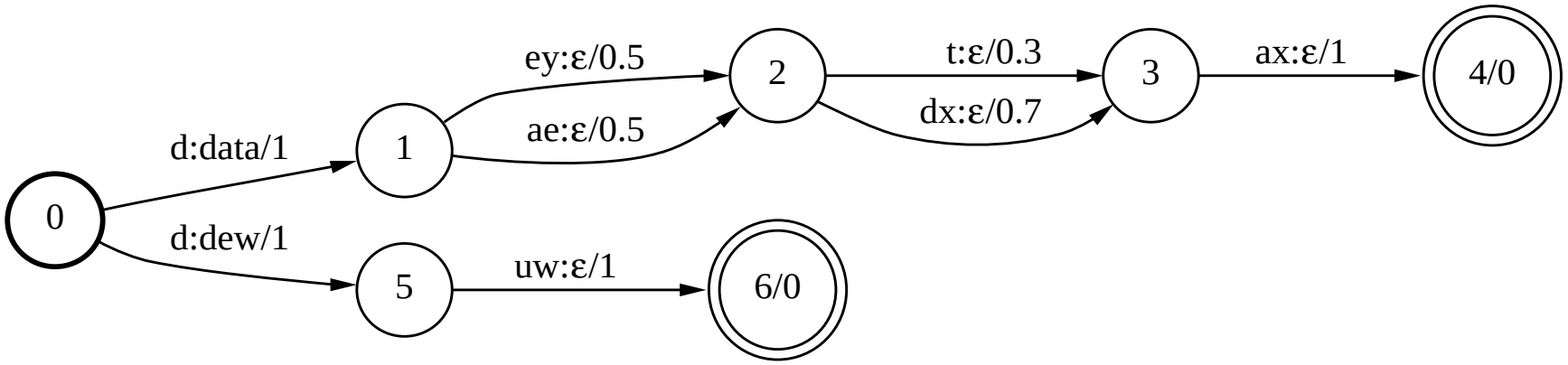
Weighted Acceptors

- Finite automata with labels and weights.
- Example: *Pronunciation model acceptor:*



Weighted Transducers

- Finite automata with input labels, output labels, and weights.
- Example: *Pronunciation lexicon transducer:*



Motivation

- Finite-State Acceptors: Compact representations of *regular (rational)* sets that are efficient to search. Examples: pattern matching (grep, PCRE), tokenization, compression.
- Finite-State Transducers: Compact representations of *rational* binary relations that are efficient to search and combine/cascade. Examples: dictionaries, context-dependent rules
- Weighted Automata: Weights typically encode uncertainty as e.g. probabilities. Examples: n-gram language models, language translation models.

References

- General Background:
  - John E. Hopcroft and Jeffrey D. Ullman. Introduction to Automata Theory, Languages, and Computation. Addison Wesley: Reading, MA, 1979.
  - Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest. Introduction to Algorithms. The MIT Press: Cambridge, MA, 1992.
- Theory of Transductions and Rational Power Series:
  - Jean Berstel. Transductions and Context-Free Languages. Teubner Studienbucher: Stuttgart, 1979.
  - Jean Berstel and Christophe Reutenauer. Rational Series and Their Languages. Springer-Verlag: Berlin-New York, 1988.
- Transducers Applied to Speech and NLP:
  - Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. The design principles of a weighted finite-state transducer library. Theoretical Computer Science, 231, 2000.
  - Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri, "OpenFst: A General and Efficient Weighted Finite-State Transducer Library", *Proc. of the 12th International Conference on Implementation and Application of Automata (CIAA 2007)*. Prague, CZ.
  - Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. Weighted Finite-State Transducers in Speech Recognition. Computer Speech and Language, 16(1):69-88, 2002.

Current OpenFst Applications

- Speech recognition (speech-to-text): lexicons, language models, phonetic context-dependency, recognizer hypothesis sets.
- Speech synthesis (text-to-speech): tokenization, text normalization, pronunciation models
- Optical character recognition: lexicons, language models
- Machine Translation: translation models, language model, translation hypothesis sets.
- Information extraction: pattern matching, text processing

Overview

1. Part I: Algorithms - Cyril Allauzen
2. Part II: Library Use and Design - Michael Riley
3. Part III: Applications - Martin Jansche