OpenFst: a General and Efficient Weighted Finite-State Transducer Library

Introduction

Cyril Allauzen - allauzen@google.com
Michael Riley - riley@google.com

December 12, 2010

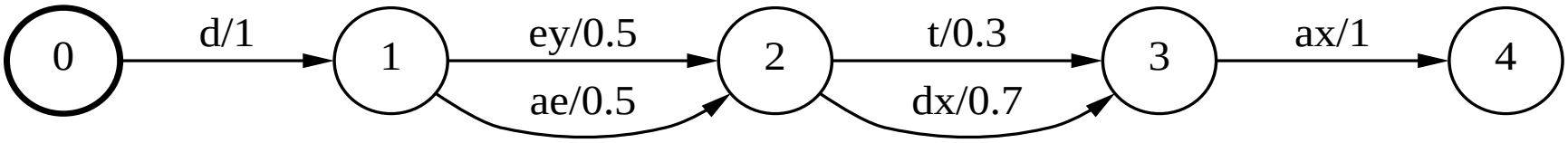
Special thanks to Johan Schalkwyk, Mehryar Mohri, and Wojtek Skut.

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.

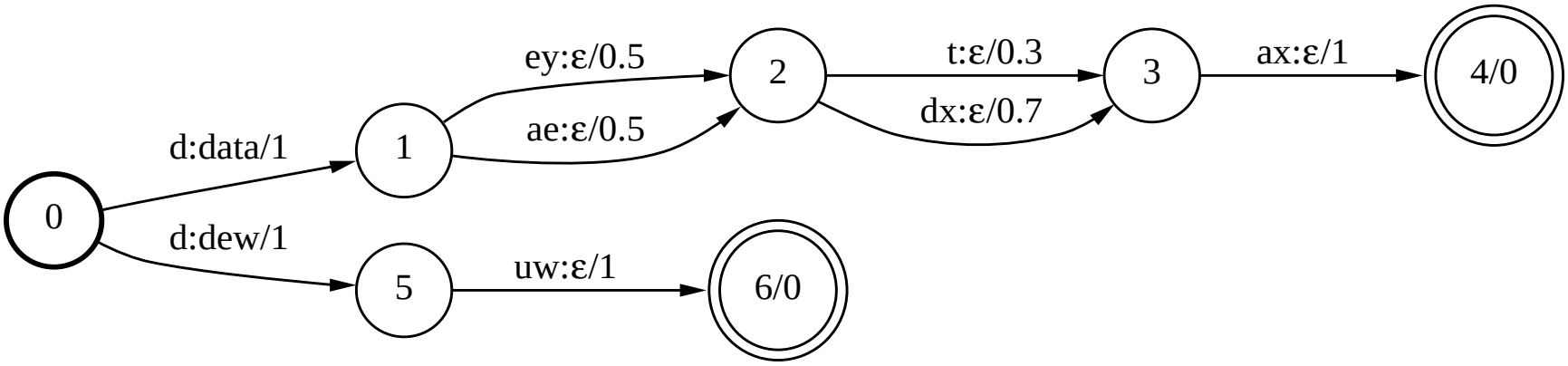
Weighted Automata

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



Weighted Transducers

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



Motivation

- Finite-State Automata: 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, 1979.
  - Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest.
    Introduction to Algorithms.
    The MIT Press, 1992.
- Theory of Transductions and Rational Power Series:
  - Jean Berstel.
    Transductions and Context-Free Languages.
    Teubner, 1979.
  - Jean Berstel and Christophe Reutenauer.
    Rational Series and Their Languages.
    Springer, 1988.
- Weighted Transducer Algorithms
  - Mehryar Mohri.
    Weighted automata algorithms.
    In *Handbook of Weighted Automata.* Springer, 2009.

- Transducers Applied to Speech and NLP:
  - Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley.
    Speech recognition with weighted finite-state transducers.
    In *Springer Handbook of Speech Processing*. Springer, 2008.
  - 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.
    In *Proc. CIAA 2007*. Springer, 2007.
  - Cyril Allauzen, Michael Riley, and Johan Schalkwyk.
    Filters for Efficient Composition of Weighted Finite-State Transducers.
    In *Proc. CIAA 2010*. Springer, 2010.

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: Library Use and Design
2. Part II: Applications