Building FST spell checkers with freely available toolkits and corpora

Tommi A Pirinen
tommi.pirinen@helsinki.fi

University of Helsinki
Department of Modern Languages
in LREC 2010—Saltmil workshop
Malta, Valletta

2010-05-20

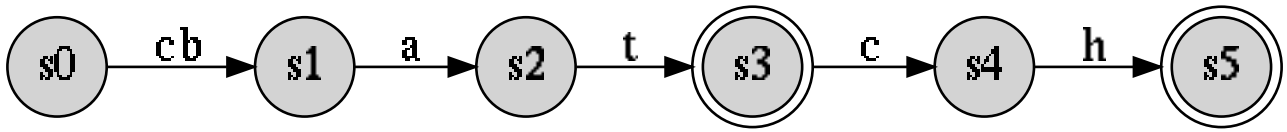
Outline

FSTs and Helsinki Finite State Technology

- simple free open source api for FSTs
- backed by Uni. Helsinki research projects and researchers
- lightweight bridging library for various free FST backends—no reinvented wheels or new FST toolkits
- implements everything needed for legacy interoperability:
  - Xerox tools (lexc, twolc; xfst under construction)
  - ispell, aspell, hunspell dictionaries (scripted, under construction)
  - AT&T/OpenFST tools (=command line interface to finite-state algebra)

FSTs for language models

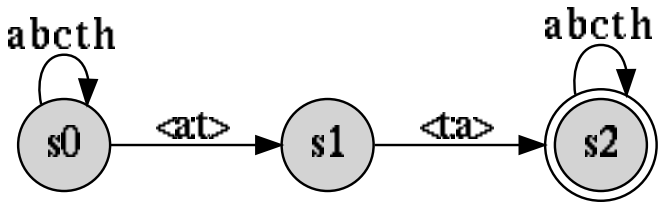
- common and tested strategy of implementing morphological analyzers in the past
- expressive enough to be able to encode most (all?) languages' morphological dictionaries
- theoretically efficient, among the fastest known methods for string matching
- weights can be used as probabilities of words, morphemes, etc.



A toy language model FSA for {cat,catch,bat,batch}

FSTs for error models

- defines translation from mispellings to correct forms
- can be used for other than spell checking
- models can be simply combined and extended with FST algebra (=not restricted by tool)
- weights can be used as probability of errors and their combinations



A toy error model FST for at->ta typo

Combining language models and error models

- error model is filter mapping wrong forms to correct ones
- the erroneous input is transformed to correct variants using composition over error model and language model
- if both are weighted, weight combining is done by fst algebra

c	t	a	0	input
c	t:a	a:t	10	error model
c	a	t	1	language model
c	a	t	11	result

correcting simple typo by composition and tropical (penalty) weighting

FST language model for spell checking

Any single-tape automaton containing correctly spelled words, e.g.:

- list of correctly written words
- corpus of word forms with frequencies
- *spell dictionaries
- FST morphologies with Xerox tools

Language models of different sources can be combined using FST union

Handmade models: Xerox tools, *spell, word-form lists

- large initial effort: requires lexicon, morphophonology
- usually maintainable
- easy to modify for specific purpose, e.g. take subset of correct language for spell checker
- may be weighted easily by hand, per word-form, per morpheme, etc.

Semi-automatic models: e.g. Wikipedia collecting

- `tokenize | sort | uniq -c` to get frequency lists; almost no initial effort
- gets some sort of popular subset of word forms with some estimate of correctness
- e.g. make likelihood of word from frequency f_w and corpus size CS by simply $\frac{f_w}{CS}$

Combination: Training hand-build model with Wikipedia

- take subset of correctly spelled word forms from Wikipedia and frequencies f_w
- assign weight to each word according to frequency and corpus size CS by $\frac{f_w}{CS}$
- assign small probability mass to word forms in language model that were not in wikipedia e.g. $\frac{1}{CS+1}$

FST error models for suggestion generation

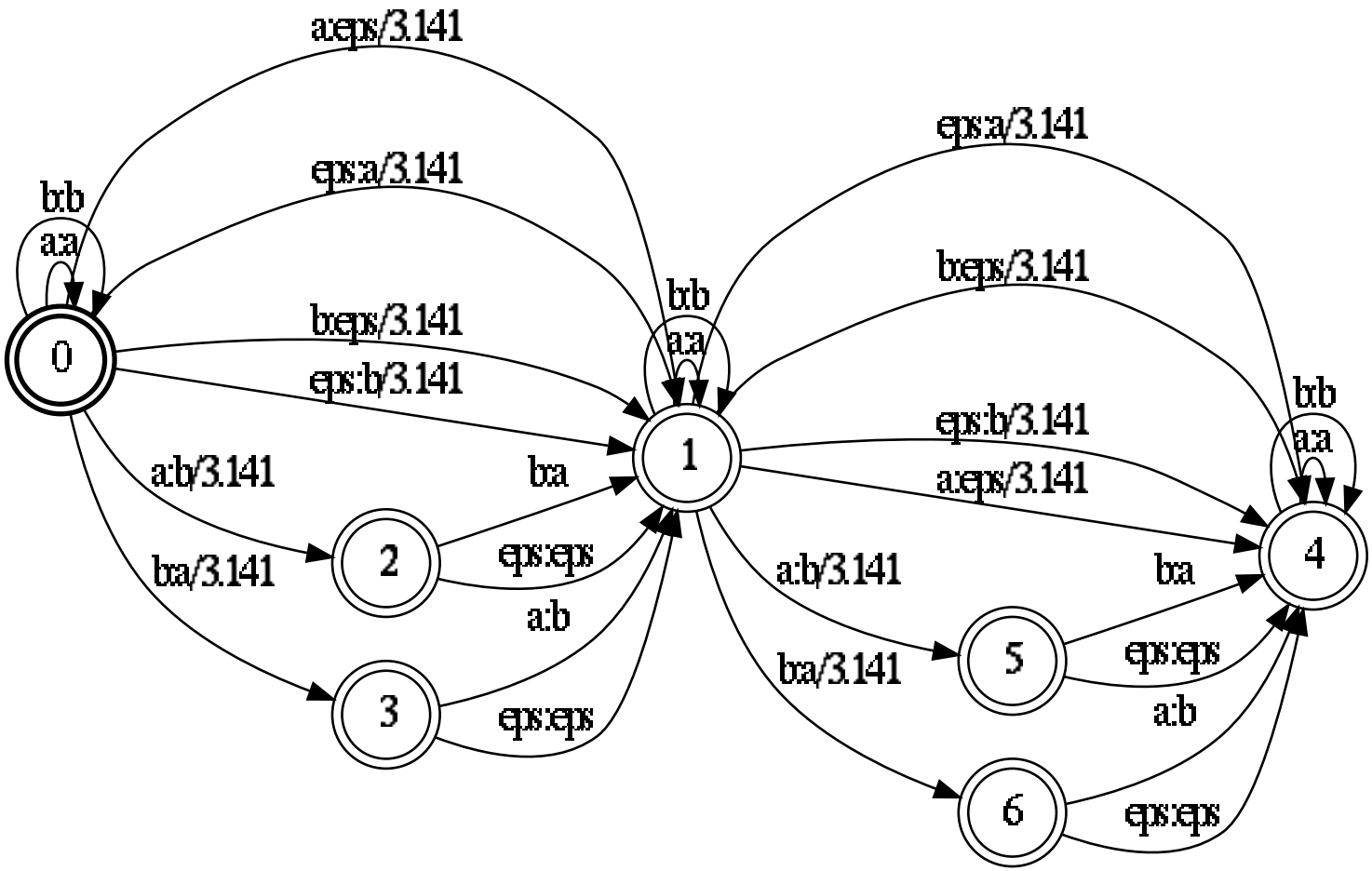
An error model is a two-tape FST mapping mispelt words into correct variants

- single typing errors, such as edit distance
- confusion sets for words or character sequences
- phonetic keying algorithms such as soundex
- e.g. from hunspell dictionaries: TRY/KEY/REP can be used

Edit distance models

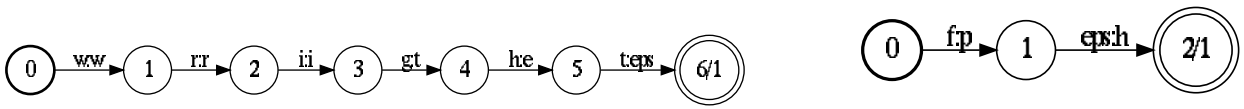
- relatively simple model for typoes: addition, deletion, substitution or swap of adjacent letters
- for each alphabet a draw arcs $a:0$, $0:a$ to end state
- for each alphabet pair a, b, draw arc $a:b$ to auxiliary ending state and afterwards $b:a$ to end state
- can be weighted using keyboard layouts, error corpora, rules, ...
- edit distance without swaps can be built with 1 state, with swaps Σ^2 states

Edit distance 2 for a and b



Confusion sets over words or character sequences

- simply modeled by FST paths attached aside other error model with lower or no weight
- word error like wright:write can be attached to star of the error model as separate path with low weight
- phonetic error f:ph can be attached by side of edit distance with lower weight



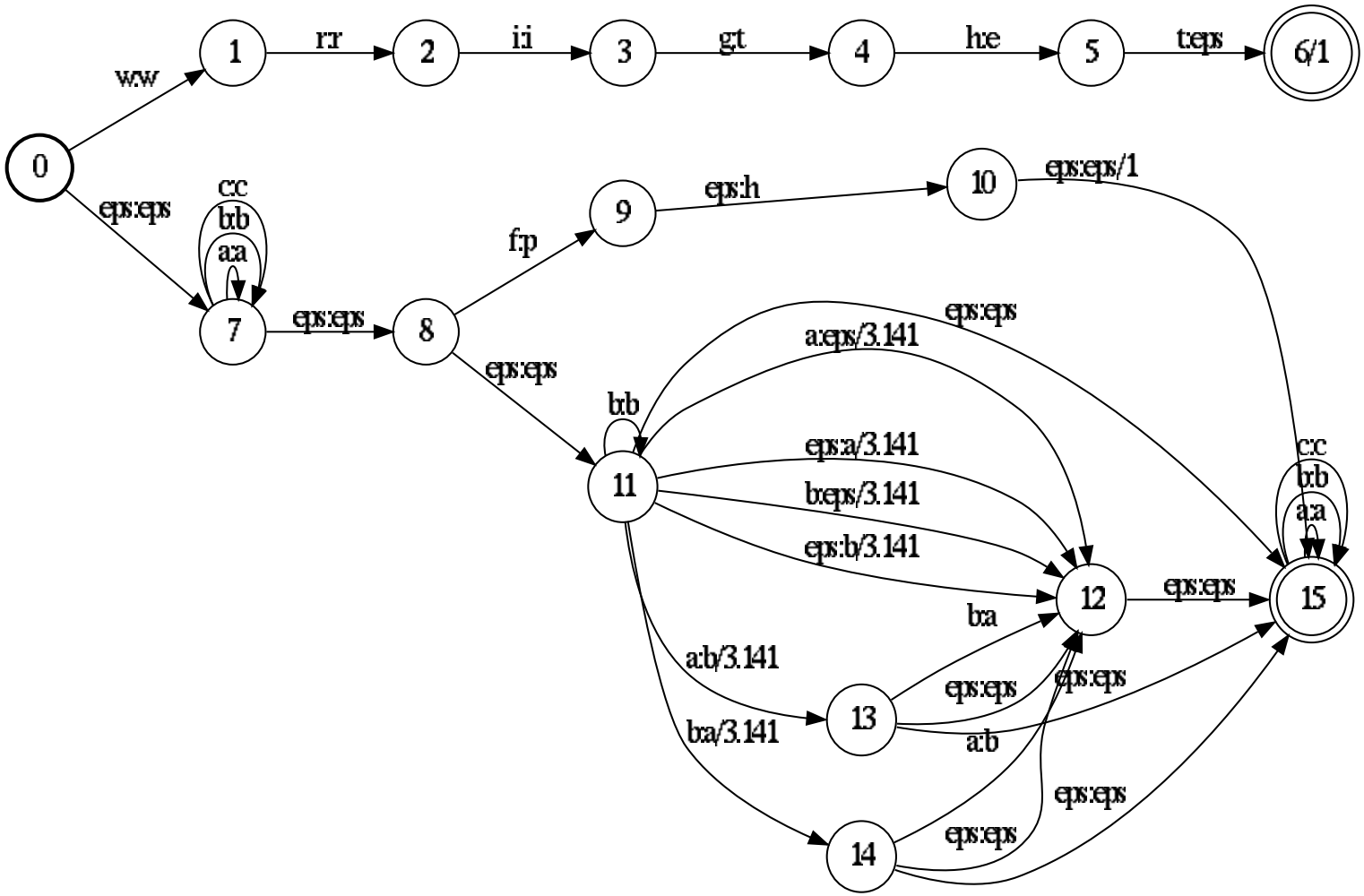
wright->write and f->ph typing errors as FSTs

Combining FST error models

Since error models were compiled to FSTs we can combine them using finite state algebra, e.g.:

- correct language model S is identity mapping of language alphabet without weight
- edit distance can be combined with other spelling errors and phonetical errors with union e.g. $ED \cup T_{ph:f}$
- edit distance of N is repetition of runs of correct spelling spliced with single edit distance errors: $ED_N = (SED_1 S)^N$
- full word mispellings combine with union to make final error model

Combined error model



Simple experiments

- existing free language models: Finnish, Northern Sámi (Xerox tools), English (word list with frequencies)
- wikipedia frequencies for language model training
- edit distance 2 with homogenous weights greater than Wikipedia frequency weight
- existing models were used as is for spell checking
- trained models were composed with error models for suggestion generation

Evaluation test setting

- gold standard of spelling errors hand collected from Wikipedia using original language model (Finnish)
- other hand made gold standards (English, Northern Sámi)
- automatically generated errors using simple algorithm generating edit disstance errors with probability of ~ 0.033 per character (all languages)

Evaluation results

Material	Rank 1	2	3	4	Lower	No rank	Total
Wikipedia word form frequencies and edit distance 2							
Finnish	451	105	50	22	62	84	761
	59 %	14 %	7 %	3 %	8 %	11 %	100 %
Northern	2421	745	427	266	2518	2732	9115
Sámi	27 %	8 %	5 %	3 %	28 %	30 %	100 %
English	9174	2946	1489	858	2902	17738	35106
	26 %	8 %	4 %	2 %	8 %	51 %	100 %

Table: gold standard

Evaluation contd.

Material	Rank 1	2	3	4	Lower	No rank	Total
Wikipedia word form frequencies and edit distance 2							
Finnish	4885	1128	488	305	1407	1635	10076
	49 %	11 %	5 %	3 %	14 %	16 %	100 %
Northern	1726	253	76	29	186	7730	10000
Sámi	17 %	3 %	1 %	1 %	2 %	77 %	100 %
English	5584	795	307	196	461	2657	10000
	56 %	8 %	3 %	2 %	5 %	27 %	100 %

Table: generated errors

Thank you.

slides and materials available through author's website

`http://www.helsinki.fi/%7Etapirine/`