HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Weighted Finite-State Methods in Spell-Checking

thesis status report in research seminar

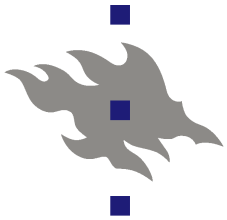
Tommi A Pirinen

‹tommi.pirinen@helsinki.fi›

February 12, 2013

University of Helsinki

Department of Modern Languages



Outline (of Thesis)

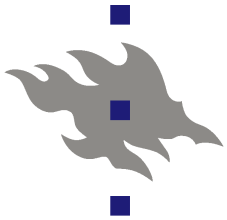
Introduction, backgrounds, motivations, history, articles

Language Models

Statistical Language Models

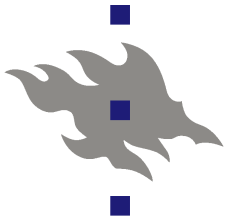
Error Models

Efficiency Evaluations



Finite-State + Spell-Checking

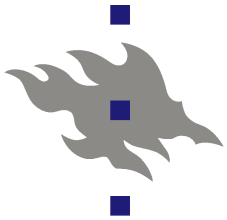
- A simple task: go through all words in text to see if they belong to language LL, if not, modify them with relation EE to fit into language LL
- In Finite-State world language model LL is any (weighted) single tape finite-state automaton recognising the words of the language
- The error model EE is any (weighted) two-tape automaton, that describes spelling errors, i.e. mapping from misspelt word into the correct one



Practical, Real-World Motivation(s)

- Fastest and most efficient way to deal with English is always finite data-structure of all word-forms (harvested from corpora) $\rightarrow$ “this problem is solved / trivial”†
- maybe finite-state approach, *infinite* lexicons, and such may be necessary for morphologically complexer languages?
  - e.g. cumulative amount of unique word-forms in texts of morphologically complex languages, the graphs for Finnish and English from Wiki are quite different
- esp. lesser resourced languages won’t get good spell-checker from corpora only

†latest measure from my FSA English speller is only beaten by aspell but not hunspell.



Theoretical Motivations?

- Regular grammars or FSAs are the weakest to describe fully morph. complex natural langs? (Not provable)
- Formal langs or methods for subset of regular languages do not generally improve efficiency? In terms of computational complexity; experimentally..?
- WFSAs provide a neat framework for bit of probabilistic and ruled combinatorics of preferences in spell-checking task

Thesis Articles I

Lind'en, K. and Pirinen, T. (2009a).

Weighted finite-state morphological analysis of finnish compounds.

In Jokinen, K. and Bick, E., editors, *Nodalida 2009*, volume 4 of *NEALT Proceedings*.

Lind'en, K. and Pirinen, T. (2009b).

Weighting finite-state morphological analyzers using hfst tools.

In Watson, B., Courie, D., Cleophas, L., and Rautenbach, P., editors, *FSMNLP 2009*.

Thesis Articles II

Pirinen, T., Lindén, K., et al. (2010).

Creating and weighting hunspell dictionaries as finite-state automata.

Investigationes Linguisticae, 21.

Pirinen, T., Silfverberg, M., and Lindén, K. (2012).

Improving finite-state spell-checker suggestions with part of speech n-grams.

Thesis Articles III

Pirinen, T. A. (2013).

Quality and speed trade-offs in weighted finite-state spell-checking.

forthcoming???

Pirinen, T. A. and Hardwick, S. (2012).

Effect of language and error models on efficiency of finite-state spell-checking and correction.

In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 1–9, Donostia–San Sebastián. Association for Computational Linguistics.

Thesis Articles IV

Pirinen, T. A. and Lind'en, K. (2010a).

Building and using existing hunspell dictionaries and TEX hyphenators as finite-state automata.

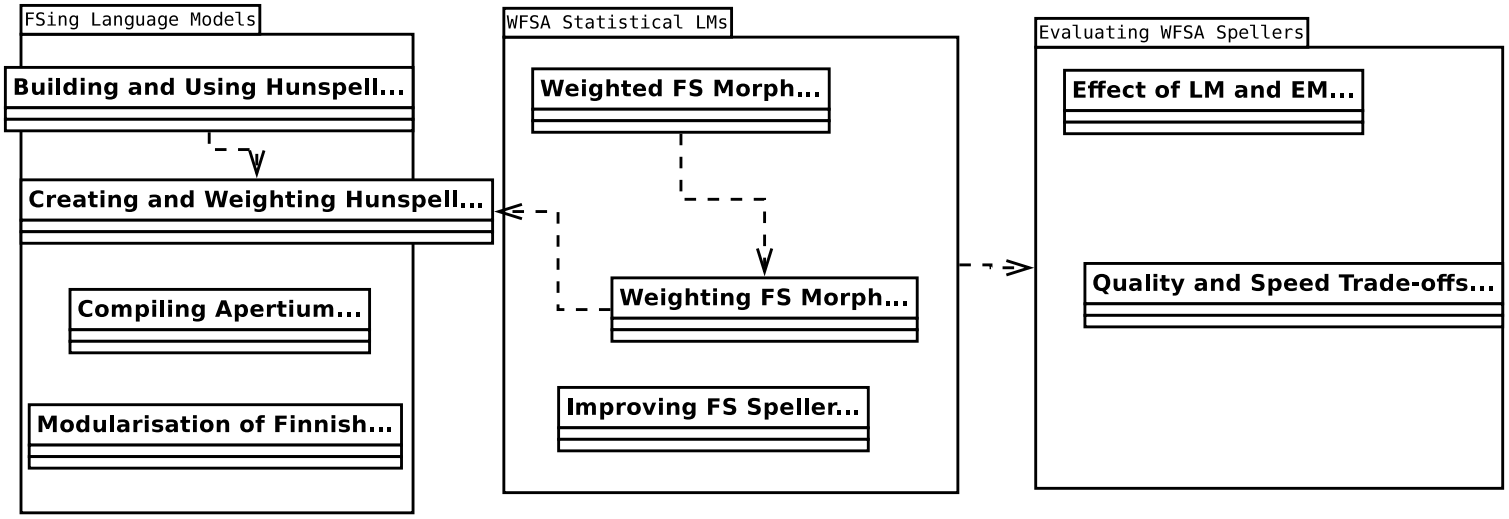
In *Proccedings of Computational Linguistics - Applications, 2010*, pages 25—32, Wisła, Poland.

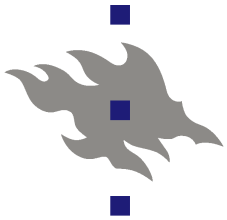
Pirinen, T. A. and Lind'en, K. (2010b).

Finite-state spell-checking with weighted language and error models.

In *Proceedings of the Seventh SaLTMiL workshop on creation and use of basic lexical resources for less-resourced languagages*, pages 13–18, Valletta, Malta.

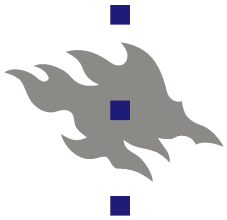
Thesis Structure in “UML”





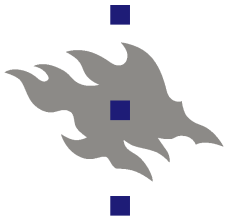
Short history of Speller LMs:

1. letter n-grams
2. word-form list $\rightarrow$ [Pirinen and Hardwick, 2012]
3. ispell, aspell, hunspell (stems, affix stripping) $\rightarrow$ [Pirinen and Lind'en, 2010a, Pirinen et al., 2010]
4. contexts (word-form trigrams, POSes) $\rightarrow$ [Pirinen et al., 2012]
5. finite-state automata $\rightarrow$ all cited
6. statistical LMs (Bayspell, winnow spell, etc.) $\rightarrow$ [Pirinen and Lind'en, 2010b]



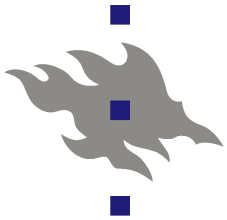
Short list of Weighted LMs:

- surface word-form unigram probabilities
- rules based on lemmas and tags
- surface word-form n-grams
- analysis probabilities almost require disambiguated gold corpora



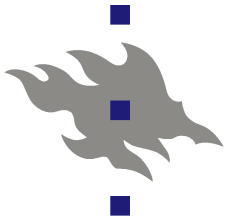
WF Probabilities(?) of morphologically complex Langs

- long compounds, derivation chains etc. that exist in these langs get rarer in corpora
- Using word-forms to train compounds in Finnish: weight of new compound *foo*+*bar* = weights of components foo, bar combined [Lind'en and Pirinen, 2009a, Lind'en and Pirinen, 2009b]
- Should be generalised: weights are counted per morph for for all languages $\rightarrow$ all languages become equally simple?



Short list of WFSA Error Models:

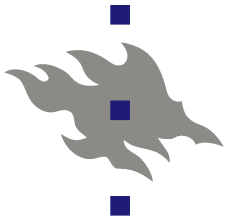
- Levenshtein-Damerau Edit distance (keyboard typing mistakes) [Pirinen and Lind'en, 2010b]
- its optimisations: no edit at first position, limiting distance, cutting parts of alphabet for mistakes [Pirinen and Hardwick, 2012]
- Confusion sets (competency errors); arbitrary string-to-string mappings
  [Pirinen and Lind'en, 2010a]
- Mistakes learnt from error corpora, may require manually verified good data
- Typical FSA ErrM would be a combination of all, done by simple disjuncting union join etc.



Speed measurements

System	WPS
English Hunspell	174
English aspell	20,000
English WFSA	999
North Saami Hunspell	3
North Saami WFSA	22
Finnish aspell	781
Finnish WFSA	1/3
Greenlandic WFSA	1/3

Table: The speed of finite-state spell-checking [Pirinen and Hardwick, 2012]



Quality measurements

System	Correct sug. 1st
English Hunspell	59.3 %
English aspell	55.7 %
English WFSA	73.7 %
Finnish aspell	21.1 %
Finnish WFSA	54.8 %
North Saami Hunspell	9.4 %
North Saami WFSA	3.5 %
Greenlandic WFSA	13.3 %

Table: The quality of spell-checkers in first suggestion correct statistics [Pirinen, 2013]