Normalization of dialects and variants using FST technology

Overview

http://tinyurl.com/clcminde

Iñaki Alegria
(University of The Basque Country)

Outline of the tutorial

- Aims and tools
- *Foma: w*riting rules for morphological analysis and normalization using finite-state technology:
    - Syntax for writing rules
    - Compiling grammars (rewrite rules)
    - Examples

      OCR, normalization Galician-Portuguese, others
    - Exercises (afternoon):
        - Normalization of Spanish tweets
        - Wide coverage por2gal
        - American/UK English
        - Other proposals by students
- *Phonetisaurus: d*ata-driven approach
    - data: http://komunitatea.elhuyar.org/tweet-norm/

References

http://tinyurl.com/clcminde

Basic material:

- http://foma.sourceforge.net/lrec2010/index.html

Toolkits:

- Rule-based (foma): http://code.google.com/p/foma/
- Data-driven approach (Phonetisaurus) http://code.google.com/p/phonetisaurus/

Bibliography

- *Beesley, K. R., & Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. CSLI, Stanford.*
- *I. Etxeberria, I. Alegria, M. Hulden, L. Uria 2014. Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. SEPLN, 52, pp. 13-20.*
- *J. Porta, J.L. Sancho: Word Normalization in Twitter Using Finite-state Transducers. Tweet-Norm@SEPLN 2013: 49-53*

Aims and tools

- Normalization is a key tool for processing texts
  Specially:
  - Non-standardized languages
  - Dialects and diachronic variants (canonicalization)
  - Different alphabets (transliteration)
  - New variants (SMS, twitter...)
- Two approach:
  - Knowledge based: writing grammars (rules)
    - hard work, high precision
  - Data based: processing examples
    - quite good results when precision is not possible
    - not clear grammar, not experts...
  - Combination: (i.e.) rule based, but assigning weights to rules based on examples

Aims and tools (2)

- Two tools (FST technology in both)

  Foma: for writing, compiling and processing rules (grammars)

  - successful and easy to learn

  Phonetisaurus: for induction of weighted rules from examples

  - machine-learning: noisy-channel model (usual in speech)
  - (a bit) difficult to install, tune...

    dependencies with other softs
  - grapheme-to-grapheme (g2g)
- Our experience:
  - foma better for dialects
  - phonetisaurus more adequate for historical texts
  - both used in tweet-normalization

foma

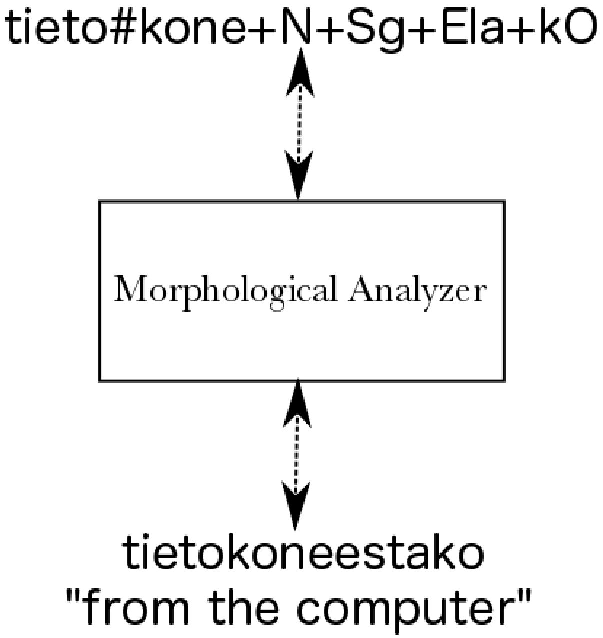
- Popular in computational morphology
- Open-source
- Similar to Xerox tools (lexc and xfst)
- Using *foma* for the morphology of several languages: Basque, Spanish, Quechua, Sami...
- And for normalization: Basque, Nahuatl, Quechua, tweets in Spanish...
- Two basic elements
  - Lexicon (and morphotactics/paradigms)
  - Phonological rules
- Compiled into FST (efficiency)
- Direct derivatives using the API:
  - spell checker/corrector, lemmatizer, verb conjugator and other ICALL and electronic dictionary tools

Morphological analysis/generation

Finnish example...

“tietokoneestako”

- compound noun tieto + kone
- singular
- elative case
- question particle



Normalization/ canonicalization

- Mainly phonological changes

  it will be our aim today
- For better results Language Model (LM) is necessarry:
  - Word-list or morphological description of the standard or pivot language
  - Easy way: word-list from the web (Wikipedia)
  - More sophisticated way: morphological description of the standard/pivot
    - Foma community and other open descriptions (*apertium*)
    - *Hunspell* and other spelling checkers

Installing foma

http://code.google.com/p/foma/

- Download (better on Linux, 32 or 64 bit)
  - From source:

```
make; make install;
```

  - Download the binary and set the PATH:

Save on Desktop/foma

```
PATH=$PATH:~/Desktop/foma/linux64
```

- Experimental support for FSM visualization
  - Linux: visualization requires “GraphViz” and “gqview”

```
sudo apt-get install graphviz
sudo apt-get install gqview    #or geeqie
```

  - Mac: Visualization requires GraphViz for OSX from http://www.pixelglow.net