Morfessor 2.0: Toolkit for Statistical Morphological Segmentation – Codebase

Previous: Morfessor 1.0

- Written in perl
- Limited utf-8 support
- Older codebase, unsuitable for extensions

Morfessor 2.0

- Complete rewrite
- Inclusion of many previously published features
- Extensible for new features and algorithms

Usage

- Library interface
  - Directly use Morfessor from python scripts
- Command line interface
  - Run training evaluation and segmentation from the command line
  - Almost complete coverage of Morfessor functionality

Features

- On-line training
- Training speed-up with random skips
- Frequency threshold and dampening for words in training data
- Possibility to weight training data likelihoods
- Optimization of data likelihood weight based on development set
- Optimization of data likelihood weight based on average morph type length
- Optimization of data likelihood weight based on average morphs / word

Items in grey will be released in Morfessor 2.1

Implementation

- Python
  - Runs on Python2, Python3 and Pypy interpreters
  - Best performance: Pypy
- Unit-agnostic code
  - Split words into morphs, or sentences into phrases

Demo

- Build on top of library interface
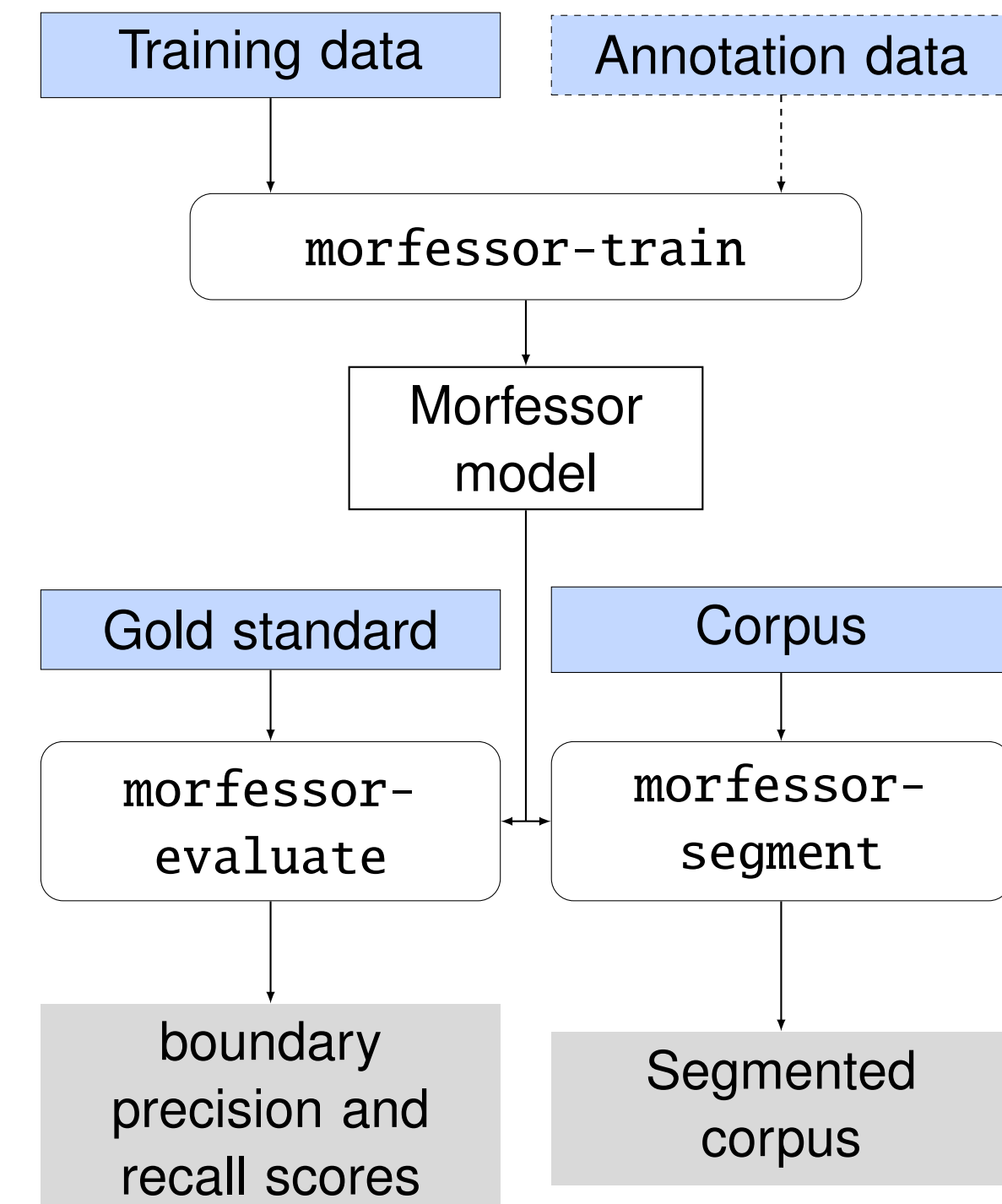- Will be available online in future

Distribution

- Source code availabe at GitHub
- Packages available from the PYthon Package Index (Pypi)
- BSD Open Source License

Upcoming developments

- Morfessor 2.0 based model with morphotactic constraints (categories)
- Data selection techniques

Acknowledgements

Links

- Homepage
  `http://www.cis.hut.fi/projects/morpho/`
- GitHub
  `https://github.com/aalto-speech/morfessor`
- Pypi
  `pip install morfessor`
  `https://pypi.python.org/pypi/Morfessor`
- Documentation
  `http://morfessor.readthedocs.org/en/latest/`

Aalto University

Morfessor 2.0: Toolkit for Statistical Morphological Segmentation – Codebase
Peter Smit, Sami Virpioja, Stig-Arne Grönroos and Mikko Kurimo
Aalto University