Omorfi as a part of an MT pipeline in FinMT, Helsinki, 2016-09-12

https://github.com/flammie/omorfi/

Tommi A Pirinen ‹tommi.antero.pirinen@uni-hamburg.de ›

HZSK.de, de.CLARIN.eu, etc.

September 9, 2016

Introduction

- `http://flammie.github.io/omorfi`
- Omorfi is a lexical database that can be compiled into different sorts of rule-based parsers and similar stuff
- can be useful for machine translation with: morphological analysis/generation, {de,}segmentation, tokenisation, lemmatisation

Installation and use

- For serious use: modern Linux distro, install recent autotools, gcc or clang, python3 etc. basics, then HFST[1] with python bindings
- `git clone git@github.com:flammie/omorfi && cd omorfi && ./autogen.sh && make && make install`
- Scripts starting with `omorfi-` should now just work (tabtab).
- e.g. `$MOSES/scripts/tokenizer/tokenizer.perl < europarl-v8.fi.text | omorfi-factorise.py`

[1] http://hfst.github.io

Example sentence

newstestB2016-enfi.ref.fi.sgml docid="aamulehti.fi.22966" seg="1": "Poliisilla on ollut etsintätehtävä 9-vuotiaan pojan karkumatkan takia" police has had a search mission due to an escape journey of a 9-year-old boy

Omorfi can be used to add stuff to word-forms

Surface:	Poliisilla	on	ollut	etsintätehtävä
“Lemmas”:	poliisi	olla	olla	etsintätehtävä
UPOS:	NOUN	AUX	VERB	NOUN
CPOS:	SG.ADE	ACT.INDV PRES.SG3	ACT.PCP.SG	SG.NOM
morphs:	poliisi lla	o n	ol lut	etsintä tehtävä
(and more)				
Target:	Police	has	had	a rescue mission

more word-forms

Surface:	9-vuotiaan	pojan	karkumatkan	takia
“Lemmas”:	9-vuotias	poika	karkumatka	takia
UPOS:	ADJ	NOUN	NOUN	ADP
CPOS:	SG.GEN	SG.GEN	SG.GEN	POSTP
morphs:	9 -vuotiaan	poja n	karku matka n	takia
(and more)				
Target:	because of	9-year-old	boy’s	escape

In moses pipeline

- Put stuff in factors:
  `Poliisilla|poliisi|NOUN|SG.ADE|poliisi lla`
- Other pre-preprocessing like splitting compounds and morphs:
  `etsintä @@tehtävä` (a rescue mission)

In apertium pipeline

1. ambiguous parses

Poliisilla	on	ollut	etsint
poliisi.n.sg.ade	olla.vaux.pri.p3	olla.vaux.actv.past.conneg	etsint
	olla.vblex.pri.p3.sg	olla.vaux.actv.pp	etsint
		olla.vaux.pass.nut.pl.nom	
		...	

In apertium pipeline

2. disambiguate

Poliisilla	on	ollut	etsintätehtävä
poliisi.n.sg.ade	olla.vaux.pri.p3	olla.vblex.actv.pp	etsintätehtävä.n.s

In apertium pipeline

3. see the rest in the next presentation!