Weighting Finite-State Morphological Analyzers using HFST tools

Tommi Pirinen, Krister Lindén
tommi.pirinen@helsinki.fi

University of Helsinki
Department of General Linguistics

2009-18-07

Morphological Analyzer and Weights

Starting with

- An unweighted morphological FST mapping surface forms to analyses
- A corpus of words' and analyses' frequencies

Create

- A weighted FST to disambiguate analyses
- E.g. Finnish morphology with fully productive compounding contains following ambiguous analyses:
  - *isänisä* `isän#isä` 'grandfather' or `isä#nisä` 'father udder'
  - *talonmies* `talonmies` (janitor) or `talon#mies` (man of the house)
  - *kuin* `kuin` (conjunction) or `kuu` 'moon' plural instructive

Counting Weights from Corpus Tokens

- Assume corpus size of CS entries, including:

talo	1149	talon	1673	talotta	0
mies	6250	miehen	2998	miehettä	0
talonmies					68

- We approximate probabilities by token count $+1$ in corpora to get weight (converted with $-\log()$ for tropical semiring):
  - talon = $-\log(\frac{1673}{CS})$
  - mies = $-\log(\frac{6251}{CS})$
  - talonmies = $-\log(\frac{69}{CS})$

Counting Weights from Corpus Tokens

- Assume corpus size of CS entries, including:

talo	1149	talon	1673	talotta	0
mies	6250	miehen	2998	miehettä	0
talonmies					68

- We approximate probabilities by token count $+1$ in corpora to get weight (converted with $-\log()$ for tropical semiring):
  - talon = $-\log(\frac{1673}{CS})$
  - mies = $-\log(\frac{6251}{CS})$
  - talonmies = $-\log(\frac{69}{CS})$
  - talon#mies = $-\log(\frac{1674}{CS}) + -\log(\frac{6251}{CS})$

- For generated compounds we want the weight collected from compound parts as tokens

Counting Weights from Corpus Tokens

- Assume corpus size of CS entries, including:

talo	1149	talon	1673	talotta	0
mies	6250	miehen	2998	miehettä	0
talonmies					68

- We approximate probabilities by token count $+1$ in corpora to get weight (converted with $-\log()$ for tropical semiring):
  - talon = $-\log(\frac{1673}{CS})$
  - mies = $-\log(\frac{6251}{CS})$
  - talonmies = $-\log(\frac{69}{CS})$
  - talon#mies = $-\log(\frac{1674}{CS}) + -\log(\frac{6251}{CS})$
  - talon#miehettä = $-\log(\frac{1674}{CS}) + -\log(\frac{\frac{1}{CS+1}}{})$
- For generated compounds we want the weight collected from compound parts as tokens
- For unknown words and parts we want to assign *maximum weight* greater than corpus assigned weights e.g. $\geq -\log(\frac{1}{CS+1})$

Structure of Morphological Analyzer

- A Finnish morphology in (HFST) lexc & twolc style:

```
LEXICON                 Root
                        CompoundNonFinal                ;
                        CompoundFinal                   ;
LEXICON                 CompoundNonFinal
talo                    Compound        "house"         ;
talon                   Compound        "house's"       ;

LEXICON                 Compound
#                       CompoundNonFinal                ;
#                       CompoundFinal                   ;

LEXICON                 CompoundFinal
mies+sg+nom:mies        #               "man"           ;
mies+sg+gen:miehen      #               "man's"         ;
mies+sg+abe:miehettä    #               "into a man"    ;
```

- Arbitary nominal chains for compounds (i.e. Root CompoundNonFinal* CompoundFinal)

Structure of Morphological Analyzer

- A Finnish morphology in (HFST) lexc & twolc style:

```
LEXICON                 Root
                        CompoundNonFinal                ;
                                                        ;
LEXICON                 CompoundNonFinal
talo                    Compound        "house"         ;
talon                   Compound        "house's"       ;

LEXICON                 Compound
#                       CompoundNonFinal                ;
#                        #                              ;

LEXICON                 CompoundFinal
mies+sg+nom:mies        #               "man"           ;
mies+sg+gen:miehen      #               "man's"         ;
mies+sg+abe:miehettä    #               "into a man"    ;
```

- Arbitary nominal chains for compounds (i.e. Root CompoundNonFinal* CompoundFinal)
- Only compound initial forms (CompoundNonFinal*)

Structure of Morphological Analyzer

- A Finnish morphology in (HFST) lexc & twolc style:

```
LEXICON                 Root
                                                    ;
                        CompoundFinal               ;
LEXICON                 CompoundNonFinal
talo                    Compound        "house"     ;
talon                   Compound        "house's"   ;

LEXICON                 Compound
#                       CompoundNonFinal            ;
#                                                   ;

LEXICON                 CompoundFinal
mies+sg+nom:mies        #               "man"       ;
mies+sg+gen:miehen      #               "man's"     ;
mies+sg+abe:miehettä    #               "into a man" ;
```

- Arbitary nominal chains for compounds (i.e. Root CompoundNonFinal* CompoundFinal)
- Only compound initial forms (CompoundNonFinal*)
- Only compound final forms (CompoundFinal)

Weighting Decomposed Morphological FST by Composition

- Compose weighted tokens with word boundary marker over initial parts $A*$

Corpus	talo# <n.nn>	talon# <n.nn>	Σ∗ # <MAX>
Morpho-logy FST	talo# talo	talon# talon	XXX# XXX

- Compose weighted tokens with analyses over compound final parts and non-compounds B

Corpus	mies SG NOM <>	mies SG GEN <>	Σ∗ <MAX>
Morpho-logy FST	mies SG NOM mies	mies SG GEN miehen	MIES SG ABE miehettä

- Concatenate AB to get weighted compounds

Results of Weighted FST Unigram Tagger

- Both training and testing corpora are news papers analyzed with another, commercial, automatic disambiguating analyzer (i.e. not a gold standard)
- The weight-ranked analyses were classified to 4 types:
  - Correct reading at first position
  - Correct reading in position other than first
  - No correct reading among analyses
  - No analyses

Train Year	Test Year	1^{st} Correct (%)	n^{th} Correct (%)	No Correct (%)	No Analysis (%)
1995	1995	95.9	4.1	0.0	0.0
1995	1996	93.3	4.0	0.7	2.0
1995	1997	93.1	4.0	0.6	2.3
1996	1995	92.9	4.0	0.7	2.2
1996	1996	96.1	3.9	0.0	0.0
1996	1997	93.6	3.7	0.6	1.9
1997	1995	91.6	4.1	1.0	3.2
1997	1996	92.1	3.9	0.9	3.1
1997	1997	96.3	3.7	0.0	0.0

Error Analysis

- Definetely not solvable with unigram tagger:
  - ...