Weighted Finite-State Morphological Analysis of Finnish Compounding with HFST-LexC

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

University of Helsinki
Department of General Linguistics

2009-XX-XX

Finnish Compounding Analysis

- Finnish nominal compounding allows arbitrary noun chains with final noun inflected and other nouns in genitive or nominative form. E.g. *isä* (father), *isänisä* (grandfather), *isänisänisä* (great grandfather) and so forth.
- In lexicons, certain compounds can be lexicalized, i.e. treated as mainly non-compound unit. E.g. *talonmies* (janitor) $\sim$ *talon#mies* (man of the house)
- Productive compounding and lexicalization of compounds results in segmentational ambiguity, such as:
  - *paikassa* (in the place), *pai#kassa* (pie cash register)
  - *isän#isä* (grandfather), *isä#nisä* (father udder)
  - *avaruus#lentotukikohta* (space # flight base), *avaruuslento#tukikohta* (space flight # base)

Disambiguation criteria

- Most likely reading is the one with simplest structure
  - We use number of word boundaries as metric of structural complexity
  - e.g. prefer *paikassa* over *pai#kassa*
- Most likely reading has most common words
  - We use frequency of word forms in corpus as commonness of words in compound
  - e.g. usually prefer *isän#isä* over *isä#nisä*
  - e.g. preference between *avaruuslento#tukikohta* and *avaruus#lentotukikohta* may vary depending on corpora

Learning Corpus Frequencies from Word Forms

- Assume corpus size of *CS* entries, including:

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

- We use probabilities of word tokens in corpora to give weight to compound parts (converted with $-log()$ for tropical semiring):
  - talonmies = $-log(\frac{69}{CS})$
  - talon#mies = $-log(\frac{1674}{CS}) + -log(\frac{6251}{CS})$
  - talon#miehettä = $-log(\frac{1674}{CS}) + -log(\frac{1}{CS})$

Learning Corpus Frequencies from Word Forms

- Assume corpus size of CS entries, including:

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

- We use probabilities of word tokens in corpora to give weight to compound parts (converted with $-log()$ for tropical semiring):
  - talonmies = $-log(\frac{69}{CS}) + 0$
  - talon#mies = $-log(\frac{1674}{CS}) + -log(\frac{6251}{CS}) + -log(\frac{1}{CS+1})$
  - talon#miehettä = $-log(\frac{1674}{CS}) + -log(\frac{1}{CS}) + -log(\frac{1}{CS+1})$
- To account the structural penalty we add weight equal to (or greater than) $-log(\frac{1}{CS+1})$ per word boundary.

Weighting Finnish Compounding in Lexc Lexicon Formalism

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

LEXICON         Compound
#               CompoundNonFinal                ;
#               CompoundFinal                   ;

LEXICON         CompoundFinal
talo+sg+nom:talo    #       "house,"            ;
talo+sg+gen:talon   #       "house's,"          ;
talo+sg+ill:taloon  #       "in to the house,"  ;
```

- Any amount of compound initial forms, then word boundary, and final form with analyses

Weighting Finnish Compounding in Lexc Lexicon Formalism

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

LEXICON         Compound
#               CompoundNonFinal    "weight: -log(1/(CS+1))"    ;
#               CompoundFinal       "weight: -log(1/(CS+1))"    ;

LEXICON         CompoundFinal
talo+sg+nom:talo    #           "house, "               ;
talo+sg+gen:talon   #           "house's, "             ;
talo+sg+ill:taloon  #           "in to the house, "     ;
```

- Any amount of compound initial forms, then word boundary, and final form with analyses
- Assign maximum weight at word boundaries

Weighting Finnish Compounding in Lexc Lexicon Formalism

```
LEXICON          CompoundNonFinal
talo             Compound        "house, weight: -log(f('talo')/CS)"          ;
talon            Compound        "house's, weight: -log(f('talon')/CS)"       ;

LEXICON          Compound
#                CompoundNonFinal  "weight: -log(1/(CS+1))"                   ;
#                CompoundFinal     "weight: -log(1/(CS+1))"                   ;

LEXICON          CompoundFinal
talo+sg+nom:talo #                 "house, "                                  ;
talo+sg+gen:talon #                "house's, "                                ;
talo+sg+ill:taloon #               "in to the house, "                        ;
```

- Any amount of compound initial forms, then word boundary, and final form with analyses
- Assign maximum weight at word boundaries
- Calculate surface form frequency weights for initial elements

Weighting Finnish Compounding in Lexc Lexicon Formalism

```
LEXICON          CompoundNonFinal
talo             Compound          "house, weight: -log(f('talo')/CS)"          ;
talon            Compound          "house's, weight: -log(f('talon')/CS)"       ;

LEXICON          Compound
#                CompoundNonFinal  "weight: -log(1/(CS+1))"                     ;
#                CompoundFinal     "weight: -log(1/(CS+1))"                     ;

LEXICON          CompoundFinal
talo+sg+nom:talo #                 "house, weight: -log(f('talo+sg+nom')/CS)"   ;
talo+sg+gen:talon #                "house's, weight: -log(f('talo+sg+gen')/CS)" ;
talo+sg+ill:taloon #               "in to the house, weight: -log(f('talo+sg+ill')/CS)" ;
```

- Any amount of compound initial forms, then word boundary, and final form with analyses
- Assign maximum weight at word boundaries
- Calculate surface form frequency weights for initial elements
- Calculate analysis form weights for final elements

Results

- Comparing against results of another automatic disambiguating analyzer (i.e. not a gold standard)
- Using only structural penalty for boundaries gives 99.96 % precision and recall
- Adding corpus probability penalties for compound parts or only corpus weights we found virtually no disagreements with reference corpus, i.e. achieved 100 % precision and recall
- Discarding the structural penalty and retaining only the corpus penalty the result stays at 100 % precision and recall

Related research

- Similar research from other languages and methods:
  - Anne Schiller (2005) *German compound analysis with wfsc*
  - Fred Karlsson (1998) *Swetwol: A comprensive morphological analyzer for Swedish*
  - Jonas Sjörbergh and Viggo Kann (2004) *Finding the correct interperatioon of Swedish compounds a statistical approach*