Jón Friðrik Daðason & Kristín Bjarnadóttir
The Árni Magnússon Institute for Icelandic Studies
University of Iceland

Utilizing Constituent Structure for Compound Analysis

LREC 2014 Reykjavík
May 29
Session O20: Grammar, Lexicon and Morphology
Outline

• Compounding is extremely productive in Icelandic

• Unknown compounds are very common, representing a problem in NLP

• The tool described here splits Icelandic compounds into their binary constituent structure

Kristín: Icelandic compounds
Jón Friðrik: The process and the results
Icelandic compounds

- Extremely productive compounding
  - Tagged Icelandic Corpus (25M words)
  - 75.4% of known word forms are compounds

- Recursive compounding and derivation
  - [[\text{mal}]_{\text{noun}} + [\text{bik}]_{\text{noun}}]_{\text{noun}} ‘gravel + pitch’ i.e. ‘asphalt’
  - [[\text{malbik}]_{\text{noun}} + a_{\text{v-suff}}]_{\text{verb}} ‘to lay asphalt’
  - [[\text{malbik}(a)]_{\text{verb}} + \text{un}_{\text{n-suff}}]_{\text{noun}} ‘the laying of asphalt’
  - [[\text{malbikunar}]_{\text{noun}} + [\text{vél}]_{\text{noun}}]_{\text{noun}} ‘asphalt machine’

- Theoretically no limit to the number of constituents
- In practice, more than 6 constituents are rare (not counting affixes)
Variation in combining forms

- Form of nouns as modifiers:
  - Stem \( \text{bók}^{\text{stem}} + \text{sala} \) ‘book store’
  - Genitive singular \( \text{bókar}^{\text{gen.sg.}} + \text{kápa} \) ‘book cover’
  - Genitive plural \( \text{bóka}^{\text{gen.pl.}} + \text{búð} \) ‘book store’
  - Dative singular (rare) \( \text{ísi}^{\text{dat.sg.}} + \text{lagður} \) ‘ice covered’
  - Dative plural (rare) \( \text{fótum}^{\text{dat.pl.}} + \text{troðinn} \) ‘trodden by feet’
  - Link phoneme \( \text{labbi}^{\text{link}} + \text{túr} \) ‘walking trip’

- The form of modifier is arbitrary but not free, i.e., the form is lexicalized

- Semantic lexicalization has no effect on the form of the compounds
Constituent structure

fjármálaráðherra
<n.masc.>
'finance minister'

fjármála
<n.neut.gen.pl.>
'finance'

fjár
<n.neut.gen.sg.>
'money'

mála
<n.neut.gen.pl.>
'matters'

ráð
<n.neut.stem.>
'council'

ráðherra
<n.masc.>
'minister'

herra
<n.masc.>
'master'
Approaches

• Methods guided by a parallel corpus
  • Prefer splits if their translated parts match up with the translation of the compound itself
  • Example: liðsandi ‘team spirit’
    • lið<stem> ‘team’ + sandi ‘sand’
    • liðs<gen.sg.> ‘team’ + andi ‘spirit’

• Frequency-based methods
  • Estimate the probability of a split from the frequencies of its parts
  • Example: þingstörfunum
    • þing<stem> (23.846) + störfunum (297) ‘parliamentary work’
    • þings<gen.sg.> (4.688) + törfunum (18) ‘parliamentary bulls’

• Ensemble methods
Method

• Split the word into the shortest possible sequence of base words it could consist of
  • fjármálaráðherra ‘finance minister’ → fjár+mála+ráð+herra

• Join the two neighboring parts with the greatest probability of forming a compound together
  • Repeat until only one part remains

• If there are multiple potential splits into base words
  • Create a tree structure for each split
  • Choose the tree with the highest product of the probabilities of each joining operation performed
Compound corpus

• The Database of Modern Icelandic Inflection (DMII)
  • 270,000 Icelandic inflectional paradigms
  • 5.8 million inflectional forms

• Unpublished data on the binary split of 240,000 compounds

• The decompounder is trained using
  • The binary tree structures from the compound data
  • Word frequencies from a large web corpus
    • Íslenskur orðasjóður (Universität Leipzig), 500M running words
Estimating probabilities

• Known compounds
  • \( P(mod + head) = \frac{\text{count}(mod+head)}{N} \), where
  • \( \text{count}(mod + head) \) is the number of words in which this compound occurs, and
  • \( N \) is the total number of compounds

• Unseen compounds
  • \( P(mod + head) = \frac{\text{count}(mod+\ast)}{N} \ast \frac{\text{count}(\ast+head)}{N} \)
Compound templates

• Many compounds share a similar inner structure
  • \([\text{fjár+mála}]+[\text{ráð+herra}]\) ‘finance minister’
  • \([\text{dóms+mála}]+[\text{ráð+herra}]\) ‘justice minister’
  • \([\text{mennta+mála}]+[\text{ráð+herra}]\) ‘education minister’
  • \([\text{X+mála}]+[\text{ráð+herra}]\) ‘X minister’

• Estimate probability of \(\text{samgöngumála+ráðherra}\)
  • \text{samgöngu} ‘transport’ + \text{mála} ‘affairs’ is known
  • \([\text{X+mála}]+[\text{ráð+herra}]\) is known
  • Multiply the two probabilities
Example
Example
Example

```
0.00019

fjármálaraðherra

fjármála       ráðherra
  fjar         ráð
    fjar       herra
```


Example
Granularity

• NLP tasks can benefit from splits at varying levels of granularity

• Machine translation
  • mið+viku+dags+morgunn ‘mid week day morning’
  • miðvikudags+morgunn ‘Wednesday morning’

• Part-of-speech tagging
  • atvinnutækifærum ‘employment opportunities’
  • E.g., at+vinnu+tæki+færum
  • færum ‘went/move/bring/able/...’: 9 distinct tags
  • tækifærum ‘opportunities’: unambiguous
Results

- The decompounder was evaluated on a selection of manually annotated Icelandic Wikipedia articles
  - 6,098 word forms; 3,319 compounds
- Finding compounds and producing binary trees
  - Precision: 97.6%
  - Recall: 98.0%
  - Accuracy: 99.2%

| Parts | Count | Atoms  | Binary split | Binary tree |
|-------|-------|--------|--------------|-------------|
| 2     | 2,709 | 99.5%  | 99.6%        | 99.5%       |
| 3     | 513   | 93.2%  | 97.9%        | 91.8%       |
| 4+    | 97    | 91.8%  | 96.9%        | 88.7%       |
| Total | 3,319 | 98.3%  | 99.2%        | 98.0%       |
Thank you for your attention

Jón Friðrik Daðason & Kristín Bjarnadóttir
jfd1@hi.is, kristinb@hi.is

LREC 2014 Reykjavík
May 29