Premodifying –ing participles in the parsed BNC

Turo Vartiainen¹ & Jefrey Lijffijt²

¹Department of English, VARIENG, University of Helsinki, Finland
²Department of Information and Computer Science, ALGODAN, Aalto University, Finland
Premodifying –ing participles

- Participles of the type:
  - An *amusing* story
  - *The running* men

- A theoretically debated class.
  - Verbs or adjectives? Both?
  - How to annotate the participles?
Other –ing forms

• An additional challenge: there are other kinds of (nominal) premodifying –ing forms, as in:
  – A parking attendant ‘traffic warden’
  – An eating contest ‘a contest in eating’

• Compare:
  – The parking man ‘a man who is parking’
  – The eating man ‘a man who is eating’
Premodifying –ing participles

• Furthermore, the premodifying –ing participle is a very infrequent item in English.
  – Large datasets need to be analysed.
  – The British National Corpus

• Dependency information required for accurate and efficient retrieval of the –ing participles.
  – The parsed BNC
The parsed BNC

• Parsed with RASP (Briscoe et al. 2006; Andersen et al. 2008)

• Based on BNC-XML
  – Does not modify corpus, just adds information
    • Word level: new POS tags, lemmatization
    • Phrase & sentence level: grammatical relations

• Grammatical relations
  – Relation (head word, dependent word)
  – *An amusing* story \(\rightarrow\) *ncmod* (story, amusing)
Mining premodifying –ing participles

• Constructed training set with ground truth
  – Three randomly selected texts
  – Approx. 3000 –ing forms
  – 351 premodifying –ing participles
  – 12 ambiguous cases discarded

• Q1: How have these been annotated?
  – Did the POS taggers produce the same annotation?

• Q2: Can we retrieve only the premodifying –ing forms?
  – Does the parser give us the necessary information to query the corpus?
CLAWS 5: Annotation of premodifying –ing forms

93 % of positive samples

1 %

6 %
RASP POS: Annotation of premodifying –ing forms

- Negative
- Positive

<table>
<thead>
<tr>
<th>Category</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>1%</td>
<td>17%</td>
</tr>
<tr>
<td>JB</td>
<td>18%</td>
<td>18%</td>
</tr>
<tr>
<td>JJ</td>
<td>17%</td>
<td>18%</td>
</tr>
<tr>
<td>NN1</td>
<td>18%</td>
<td>65%</td>
</tr>
<tr>
<td>NNS1</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>NNT1</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>NP1</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>PN1</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>RR</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>VBG</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>VDG</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>VHG</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>VV0</td>
<td>65%</td>
<td>65%</td>
</tr>
<tr>
<td>VVG</td>
<td>1%</td>
<td>1%</td>
</tr>
</tbody>
</table>
Querying the parsed BNC (1/2)

• Construct rule / query
  – Word x relevant iff
  – “C5 (x) = ADJ” or “RASP (x) = JJ” (POS rule), and
  – “ncmod (y, x)” where y > x (premodifier rule)

• Decision tree classifier seems suitable

• Gives too simple a model
  – Many negative, few positive examples
    • Favours negative
  – Many tags with only few examples
    • Favours not using the attribute at all
Querying the parsed BNC (2/2)

• Solution:
  – Cross-tabulate all possible rules
  – Incrementally select rules using precision

• Simple model works fine!
  – “C5 = ADJ” (BNC-XML)
    • 70 % precision
    • 93 % recall
  – Rule with BNC-XML and RASP
    • 71 % precision
    • 96 % recall
    • Or very high precision / recall
    • Still room for improvement
Trade-off curve for different features

- CLAWS-5 POS
- CLAWS-5 + RASP POS + GRAM
- 50% PRECISION
- RANDOM -ING

- 99% Recall, 51% Precision
- 93% Recall, 70% Precision
- 38% Recall, 99% Precision

Trade-off curve for different features:

- 99% Recall, 51% Precision
- 93% Recall, 70% Precision
- 38% Recall, 99% Precision
- 50% Precision
- RANDOM -ING
Pilot study

• Preliminary comparison of two genres:
  1. Academic and non-academic scientific texts (natural sciences; social sciences)
  2. Imaginary prose (novels)

• 50 files from the parsed BNC
• 2,304,371 words
• 5,106 premodifying –ing participles
Pilot study

• The average frequency of –ing participle tokens is high in the scientific domain
  – However, the number of participle types is consistently lower in scientific texts than in imaginary prose.
Type/token ratio per genre

Number of unique participles in text vs Number of participles in text

- Novel
- Scientific

Graph comparing the type/token ratio for novels and scientific texts, showing the number of unique participles in text versus the number of participles in text.
Explaining the differences

• Scientific texts:
  – Topical words (e.g. the leading stars)
  – Cohesive words (e.g. following, preceding, foregoing, succeeding...)

• Imaginary prose:
  – Cohesive participles rare
  – More variation in the use of –ing participles in general
Conclusion

• We can efficiently find premodifying –ing participles using information both from the BNC-XML and the parsed BNC.

• The pilot study will provide the basis for a detailed study of –ing participles in the BNC.
References
