Detecting Errors in Corpus Annotation

Detmar Meurers
University of Tübingen

Introduction

Basic procedure

Example

A general linguistic insight: Endocentricity

Most natural language expressions analyzed as endocentric: a category projects to a phrase of the same category (e.g., X-bar Schema, Jackendoff 1977)

- Generally speaking, the category of the mother is constrained by the categories of the daughters.

Idea: Combine the two strands—variation detection and the insight behind endocentricity:

- To detect errors, search for variation in mother categories dominating the same daughters.

Our approach

The basic procedure

We implement this idea as follows:

1. Extract all local trees from treebank and index them by the daughters lists.
2. For each daughters list, determine the set of immediately dominating mothers in the corpus, the immediate dominance set (ID set).
3. If the ID set has more than one element, the daughters list shows ID variation, indicating a potential error.

A case study

This procedure returns 844 daughters lists with ID variation for the WSJ corpus as annotated in the Penn Treebank 3.

- Sampled and inspected 100 daughters lists/ID sets:
  - 74 pointed to at least one error
  - 24 correct ambiguities
  - 2 unclear

We count a daughters list as erroneous if for at least one of the mothers in the ID set, every occurrence of the daughters with that mother is incorrect.

For all 844 daughters lists, we can estimate that

- 625 point to at least one error (95% CI: 552–697)

Error examples

Three main kinds of errors in the 74 erroneous cases.

- Bracketing error (13): runs, up, and high commission costs should all be sisters
  (1) Frequent trading runs [\text{up} \text{ up} \text{ high commission costs}]
  - Mother label error (41): past it should be a PP
  (2) Turkey in any event is long \{\text{past/IN} \text{ it}\}.
  - Daughter label error (38): like should be VB
  (3) Mr. Friend's client [\ldots] didn't [\text{up} \text{ like/IN} \text{ the way 0 defense attorney Tom Alexander acted during the legal proceedings 'T'}].

From ID variation to error detection

- We have established that ID variation is useful for finding incorrectly annotated local trees.
- To make this practically useful, we want to define a heuristic for automatically detecting
  - which of the elements in the ID set of a given daughters list are errors and which aren't.
- What information will be useful/necessary for this?

Frequency-based error detection heuristics

Absolute frequency

Remove all rules in the ID sets which occur only once

- Based on idea that pruning low-frequency rules in parsing will not degrade performance (Gaizauskas 1995; Charniak 1996; Cardie & Pierce 1998)

Results:

<table>
<thead>
<tr>
<th>Types</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>74.75% (74/99)</td>
<td>58.27% (74/127)</td>
</tr>
</tbody>
</table>

- Fairly high precision
- Very low token recall

References

Detmar Meurers
University of Tübingen
Detecting Errors in Corpus Annotation
Detmar Meurers
University of Tübingen

Introduction
Basic procedure
A case study
Results
Impact of automatically removing errors
Impact of automatically removing errors

A combined heuristic: frequency + ambiguity

Results

Procedure
Relative frequency

Frequency-based error detection heuristic

Relative frequency

Removing rules which occur less than 10\% of the time within their ID sets.

† Results:

<table>
<thead>
<tr>
<th>Types</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>tokens</td>
<td>60.47% (78/129)</td>
<td>61.42% (78/127)</td>
</tr>
<tr>
<td></td>
<td>9.20% (499/5424)</td>
<td>58.91% (499/847)</td>
</tr>
</tbody>
</table>

† Fairly high recall
† Very low token precision

Adding an ambiguity measure

Idea:

† certain pairs of mother categories are likely to occur as alternatives, regardless of their frequency.

Example:

† NP vs. NX
  † NP labels noun phrases
  † NX is used for noun phrases which share a modifier with another noun phrase
  † 114 of the 844 ID sets include both NP and NX as mothers (the second-most common variation)
  † e.g.: NX → VBG NN occurs only three times, but in variation with NP as mother, and both are correct.

A combined heuristic: frequency + ambiguity

Results

<table>
<thead>
<tr>
<th>Types</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>tokens</td>
<td>73.03% (65/89)</td>
<td>51.18% (65/127)</td>
</tr>
<tr>
<td></td>
<td>65.59% (364/555)</td>
<td>42.98% (364/847)</td>
</tr>
</tbody>
</table>

† Much better token recall
† Precision still quite high

⇒ Results are encouraging enough to try to measure the impact of removing all rules detected by this method

Impact of automatically removing errors

Setup

We tested the impact of erroneous rules in training data on PCFG parsing, using LoPar (Schmid 2000) (unlexicalized, non-headed version).

† Left-corner parser which allows for easy manipulation of the set of grammar rules.
† Used sections 2-21 of WSJ to train, section 23 to test.
† Training data, rules used:
  • All (15,246 rules): All grammar rules from the treebank without modification.
  • Reduced (14,798 rules): Grammar rules after removing rules flagged by combined frequency/ambiguity heuristic.

Impact of automatically removing errors

Results

<table>
<thead>
<tr>
<th>Types</th>
<th>Precision</th>
<th>Recall</th>
<th>F(_{\beta=1})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>70.39%</td>
<td>67.31%</td>
<td>68.82%</td>
</tr>
<tr>
<td>Reduced</td>
<td>71.48%</td>
<td>68.40%</td>
<td>69.91%</td>
</tr>
</tbody>
</table>

⇒ Changes significant at \(\alpha=0.001\) (using stratified shuffling)

Conclusion:

† Presence of erroneous rules in a grammar induced from a treebank is harmful for parsing precision and recall
† Targeting and eliminating erroneous rules can improve parser performance

Summary and Outlook

Summary:

† Introduced effective new way of detecting treebank errors
  † combines variation detection with endocentricity insight
† Demonstrated that removing erroneous training data detected by method improves PCFG performance

Outlook:

† Continue exploration of heuristics to improve precision/recall of errors
† Determine what exactly causes the improvement for PCFG parser
† Perform dependency-based evaluation measures
† Test methods on other treebanks with different annotation schemes

Endocentricity and Real Life Treebanks

Some treebank annotation guidelines violate endocentricity, e.g., WSJ guidelines for proper nouns (NNP, NNP):

† Rule for POS annotation (Santorini 1990, p. 32): capitalized words which appear in a title tagged NNP

4) A NNP Test/NNP eit/IN Two/NNP Cities/NNP

† Rule for syntactic annotation (Bles et al. 1995, p. 207): titles specified to be annotated like running text

5) [S-TO NP [NP-TO SV]] [NP Driving [NP Miss Daisy]]

⇒ WSJ includes VPs headed by NNP (VP → NNP PP):

6) [NP-TO PP [NP-TO PP]] [NP Saved/NNP [NP By/NNP [NP The/NNP Bell/NNP]]]

Frequency-based error detection heuristic

An example for relative frequency

Remove rules which occur less than 10\% of the time within their ID sets.

† e.g.: daughters list NNP CC NNP NNP
  † appears 86 times
  † with UCP as mother only twice (2.33\% of 86)
⇒ UCP can be removed
Frequency-based error detection heuristics

**Insufficiency of absolute frequency**

Two main reasons this predictor is insufficient:

- Frequently-occurring rules which are incorrect
  - e.g.: NP → VBG appears 177 times, despite being wrong
- Infrequently-occurring rules which are correct
  - Of the 99 rules in our set which occur once, a full 25 of them are correct
  - e.g.: S → NP S occurs once, but is correct

Frequency-based error detection heuristic

**Insufficiency of relative frequency**

Problems for relative frequency heuristic:

- Again, infrequently-occurring rules which are correct
  - e.g.: NX → NNP CC NNP NNP is correct, despite occurring only once out of 86 total token occurrences in ID set
- Very frequent rules are too dominant:
  - Despite appearing 102 times, NX → JJ NN is under 10% threshold (NP appears 5972 times as mother)

\[ \Rightarrow \text{Frequency-based heuristic is insufficient by itself} \]

### Exemplifying combined heuristic

Why combining frequency + ambiguity measure works

Sort out highly frequent rules based on something other than frequency

- With JJ NN, mother ADJP occurs 25 times as a mother but less than 10% of the time within the variation
- Incorrectly flagged as an error by relative frequency heuristic alone
- The pairing ADJP–NP is most frequent ambiguity, so rule is correctly not flagged as error by combined heuristic
- With IN NP, mother ADVP occurs 170 times
- Pairing ADVP–PP not one of the five most frequent, so correctly flagged as an error

### References


