Treebank Usage

Martin Volk
Stockholm University

Usage - Overview

1. Training a chunker / parser on a treebank
   = Learning a probabilistic context-free grammar from a treebank
2. Evaluating a parser against a treebank
3. Using a treebank in education
   - for language learning
   - for linguistics education

Training a chunker / parser

Good introduction
  - Chap 11: Probabilistic Context-Free Grammars
  - Chap 12: Probabilistic Parsing

Parser usages

Three ways to use a probabilistic parser:
1. Probabilities for determining the best sentence
   - When the actual input is uncertain (e.g., word lattice in speech recognition), to determine the most probable sentence.
2. Probabilities for faster parsing
   - To find the best parse more quickly.
3. Probabilities for choosing between parses
   - To choose the most likely parse tree among the many parse trees for the input string.

Grammar Learning

- Automatic learning of grammars based solely on text input is impossible / hard.
  - unless negative evidence is included!
- But automatic learning of grammars based on treebanks is easy ...
- and provides probabilities on grammar rules.

Grammar learning from a treebank

- Count all derivations.
- Compute the probabilities based on the frequencies.
- The probabilities of all derivations with the same mother node must sum to 1.
- Penn Treebank
  > 10,000 rules
  ~4,000 appear more than once
  Which rule is the most frequent?

<table>
<thead>
<tr>
<th>Rule</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP → Det</td>
<td>2533</td>
</tr>
<tr>
<td>NP → Det AP</td>
<td>1265</td>
</tr>
<tr>
<td>NP → NN</td>
<td>501</td>
</tr>
<tr>
<td>NP → Det</td>
<td>388</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Problems with rule probabilities

Lexicalization needs to be taken into account.

In a pure PCFG the probability of a rule like

\[ VP \rightarrow V \text{ NP NP} \]

is independent of the verb. This is clearly wrong from a linguistic point of view.

Rule probabilities depend on grammatical functions.

- Compare subject and object positions in English:
  - An NP is much more likely to be realized as pronoun in subject position, (NP \rightarrow \text{Pron}) and
  - to be realized as NP with a prepositional attribute in object position (NP \rightarrow \text{NP PP}).

One solution: The grandparent node

- consider the tree on the right with one NP in subject position and one NP in object position

Distinguishing local trees based on the grandparent node \rightarrow via node relabeling leads to improved parsing results.

This is a way to take the derivation history into account!

"Transform" treebank trees, and proceed with PCFG extraction (Johnson, 1997)
- \~80\% labeled precision and recall

Relatives of probabilistic cf parsing

- DOP: Data oriented parsing (Rens Bod) is parsing via the recombination of parse trees of arbitrary depth.
DOP: Data-oriented parsing

• Example:
  
  \[
  \begin{array}{c}
  S \\
  \hspace{1cm} \text{NP} \\
  \hspace{2cm} \text{VP} \\
  \hspace{3cm} \text{NP} \\
  \hspace{4cm} \text{VP} \\
  \hspace{5cm} \text{NP} \\
  \hspace{6cm} \text{VP} \\
  \hspace{7cm} \text{NP} \\
  \end{array}
  \]

  \[
  \begin{array}{c}
  Mary \\
  \hspace{1cm} \text{heard} \\
  \hspace{2cm} \text{Sue} \\
  \end{array}
  \]

DOP: Data-oriented parsing

• Problems
  - How to store all possible trees.
  - Slow parsing since the highest probability tree cannot be found efficiently. Viterbi algorithm cannot be used.
• is similar to parsing with Probabilistic Tree Adjoining Grammars

Parser Evaluation: PARSEVAL

\[ P = \frac{\# \text{Correct Constituents}}{\# \text{Constituents in parser output}} \]

\[ R = \frac{\# \text{Correct Constituents}}{\# \text{Constituents in gold standard}} \]

Labeled Precision and Recall of constituents

In our example:

\[ \text{Precision} = \frac{6}{6} = 1.0 \quad \text{Recall} = \frac{6}{7} = 0.86 \]
Problems of PARSEVAL

PARSEVAL measures are not very discriminating.
- Charniak’s (’96) vanilla PCFG which ignores all lexical content worked well.
  - PARSEVAL measure is quite easy at reproducing the tree structures given by the Penn Treebank.
- PARSEVAL measures the success at the level of individual decisions.
- In NLP consecutive decisions are more important and harder.

Evaluation

- Penn Treebank’s problem
  - Too flat.
  - Non-standard adjunct structure given to post noun-head modifiers
- PARSEVAL measure seems too harsh on some specific problems.

Language / Linguistics Learning

Learning tasks over treebanks
- Viewing / searching trees
- Labeling trees
- Combining subtrees
- Comparing trees
- Evaluating trees
- Drawing trees

Some problems
- How to find rare constructions?
- How to avoid confusing the student with ungrammatical examples?

Interactive syntactic trees
(from Eckhard Bick)
Treebanks in Linguistics Courses

- H.v.Halteren "Syntactic Databases in the Classroom"
- Experiments in English syntax courses at Nijmegen University based on the TOSCA Treebank
- CLUES: Computer Library of Utterances for Exercises in Syntax

CLUES Exercise Types

1. Mark empty node, ask for label
   - What is the label for node X?
2. Give label, ask for node (unlabeled tree)
   - Which node is a prepositional complement?
3. Show partial tree, ask for reconstruction
4. Show incorrect tree, ask for correction

Studien-CD Linguistik

- an introduction to (German) linguistics
- developed at the University of Zurich (2001-2004)
- published with an introductory linguistics book
- contains 100 German syntax trees across 10 different text genres (novel, medical abstract, weather report, interview, newspaper report, fairy tale)
- in two views (complex vs. "easy")
- to be used in self-learning as
  - examples for word classes
  - examples for syntax structures

Summary

- There is a straightforward way to derive a probabilistic context-free grammar from a treebank. But this PCF grammar will need optimization (e.g. lexicalisation, context) for high accuracy parsing.
- It is difficult to establish a good measure for parser evaluation (i.e. tree comparison). PARSEVAL is the measure with wide-spread use.
- Treebanks can be used in syntax education.