Induction of Treebank-Aligned Lexical Resources

LREC 2008

Tejaswini Deoskar, Mats Rooth

Department of Linguistics
Cornell University
Overview

- **Goal:** Induction of probabilistic treebank-aligned lexical resources.

- *Treebank-Aligned Lexicon:* a systematic correspondence between features of a probabilistic lexicon and structural annotation in a treebank.

- **Features:**
  - complex subcategorization frames for verbs or nouns.
  - attachment preference of adverbs
Overview

- Treebank PCFG and lexicon.
  - Unlexicalised Treebank PCFG: Clear division between grammar and lexicon.
  - Good performance (Klein and Manning, 2003)
- Large-scale lexicon: Unsupervised acquisition from unlabeled data.
Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - Exports which played a key role in fueling growth over the last two years seem to have stalled.
Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - Exports which played a key role in fueling growth over the last two years seem to have stalled.

- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - Exports which played a key role in fueling growth over the last two years seem to have stalled.
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
- A sophisticated PCFG that captures the same phenomena as more expressive formalisms.
Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - Exports which played a key role in fueling growth over the last two years seem to have stalled.
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
- A sophisticated PCFG that captures the same phenomena as more expressive formalisms.
  - Linguistic theory neutral.
Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - Exports which played a key role in fueling growth over the last two years seem to have stalled.
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
- A sophisticated PCFG that captures the same phenomena as more expressive formalisms.
  - Linguistic theory neutral.
  - Focus on commonly observed phenomenon.
Treebank Transformation Framework

- Treebank Transformation: Johnson (1999), Klein and Manning (2003), etc.
- Training of PCFG on transformed treebank.
Treebank Transformation Framework

- Treebank Transformation: Johnson (1999), Klein and Manning (2003), etc.
- Training of PCFG on transformed treebank.

- Methodology for transformation based on addition of linguistically motivated features, and feature-constraint solving.
- Database of Penn Treebank trees annotated with linguistic features as a resource.
- Components usable for transforming existing PTB-style treebanks, and building accurate PCFGs from them.
Feature Constraint Framework

- Bare-bones CFG extracted from Penn Treebank.
- A feature-constraint grammar is built by adding constraints on CF rules (YAP, Schmid (2000)).
- Each treebank tree converted into a trivial context-free shared forest.
- Constraints in the shared forest solved by YAP constraint solver.
Adding Constraints

Features on auxiliary verbs:
Adding Constraints

Features on auxiliary verbs:

\[ VP \rightarrow VB \ ADVP \ VP \]
Adding Constraints

Features on auxiliary verbs:

\[
\begin{align*}
\text{VP} & \rightarrow \text{VB} \quad \text{ADVP} \quad \text{VP} \\
\text{VP} \{ \text{Vform} = \text{base}; \} & \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \}
\end{align*}
\]
Adding Constraints

Features on auxiliary verbs:

\[ \text{VP} \rightarrow \text{VB} \quad \text{ADVP} \quad \text{VP} \]

\[ \text{VP} \{ \text{Vform} = \text{base}; \} \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \} \]

\[ \text{VP} \{ \text{Vform} = \text{base}; \ \text{Slash} = s/; \} \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \ \text{Vsel} = \text{vf}; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \text{Slash} = s/; \ \text{Vform} = \text{vf} \} \]
Adding Constraints

Features on auxiliary verbs:

\[
\begin{align*}
\text{VP} & \rightarrow \text{VB} \quad \text{ADVP} \quad \text{VP} \\
\text{VP} \{ \text{Vform} = \text{base}; \} & \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \} \\
\text{VP} \{ \text{Vform} = \text{base}; \text{Slash} = \text{sl}; \} & \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \text{Vsel} = \text{vf}; \} \quad \text{ADVP} \{ \} \\
& \quad \text{VP} \{ \text{Slash} = \text{sl}; \text{Vform} = \text{vf} \} \\
\text{VP} \{ \text{Vform} = \text{base}; \text{Slash} = \text{sl}; \} & \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \text{Vsel} = \text{vf}; \} \quad \text{Prep} = -; \text{Prtcl} = -; \text{Sbj} = -; \} \quad \text{ADVP} \{ \} \\
& \quad \text{VP} \{ \text{Slash} = \text{sl}; \text{Vform} = \text{vf} \}
\end{align*}
\]
Adding Constraints

Features on auxiliary verbs:

\[ \text{VP} \rightarrow \text{VB} \quad \text{ADVP} \quad \text{VP} \]
\[ \text{VP} \{ \text{Vform} = \text{base}; \} \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \} \]
\[ \text{VP} \{ \text{Vform} = \text{base}; \text{Slash} = sl; \} \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \text{Vsel} = vf; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \text{Slash} = sl; \text{Vform} = vf \} \]
\[ \text{VP} \{ \text{Vform} = \text{base}; \text{Slash} = sl; \} \rightarrow \text{VB} \{ \text{Val} = \text{aux}; \text{Vsel} = vf; \text{Prep} = -; \text{Prtcl} = -; \text{Sbj} = -; \} \quad \text{ADVP} \{ \} \quad \text{VP} \{ \text{Slash} = sl; \text{Vform} = vf \} \]
Relative Clause

..that has been seen.
Verbal Subcategorization Features

VP $\rightarrow$ VBD +EI-NP+ S
Verbal Subcategorization Features

\[ \text{VP} \rightarrow \text{VBD} +\text{EI-NP}+ \text{S} \]

\[ \text{VP}\{ \text{Vform} = \text{ns}; \} \rightarrow \text{VBD} \{ \text{Val} = \text{ns}; \} +\text{EI-NP}+ \text{S} \{ \} \]
Verbal Subcategorization Features

\[ VP \rightarrow VBD \ +EI-NP+ \ S \]
\[ VP\{ Vform = ns; \} \rightarrow VBD \{ Val = ns; \} \ +EI-NP+ \ S \{ \} \]
\[ VP\{ Vform = ns; \} \rightarrow VBD \{ Val=ns; Sbj = x; Vsel = vf; \} \]
\[ +EI-NP+ \]
\[ S \{ Sbj= x; Vform = vf; \} \]
Verbal Subcategorization Features

\[ VP \rightarrow VBD +E\text{I-NP}+ S \]
\[ VP\{ Vform = ns; \} \rightarrow VBD \{ Val = ns; \} +E\text{I-NP}+ S \{ \} \]
\[ VP\{ Vform = ns; \} \rightarrow VBD \{ Val = ns; Sbj = x; Vsel = vf; \} \]
\[ +E\text{I-NP}+ \]
\[ S \{ Sbj = x; Vform = vf; \} \]
\[ VP\{ Vform = ns; Slash = s/; \} \rightarrow VBD \{ Val = ns; Sbj = x; Vsel = vf; \}
Prep = -; Prtcl = -; \}
\[ +E\text{I-NP}+ \]
\[ S \{ Sbj = x; Vform = vf; Slash = s/; \} \]
Verbal Subcategorization

Structural information is projected onto lexical item: verbs, adverbs, nouns.
A feature-structure Treebank Tree

The product-design project he heads is scrapped
Treebank PCFG

- Frequencies collected from feature-annotated treebank database.
- Rule frequency table and frequency lexicon that can be used by a probabilistic parser.
Treebank grammar and lexicon

29092.0  ROOT → S.fin.-.-.root
14134.0  S.fin.-.-.- → NP-SBJ.nvd.base.-.-. VP.fin.-.-
13057.0  NP-SBJ.nvd.base.-.-.- → PRP
13050.0  PP.nvd.of.np → IN.of NP.nvd.base.-.-.-.-

tried  VBD.s.e.to.- 32.0 VBN.s.e.to.- 11.0 VBN.n.-.-.- 5.0
       VBD.z.-.-.- 1.0 VBD.n.-.-.- 1.0 VBD.s.e.g.- 1.0
       VBN.z.-.-.- 1.0
admiried  VBD.n.-.-.- 1.0
admit  VB.z.-.- 1.0 VB.n.-.- 1.0 VB.b.-.- 3.0
       VBP.z.-.- 1.0 VBP.p.-.- 1.0 VBP.b.-.- 2.0
admonishing  VBG.s.-.to 1.0
Treebank PCFG

- PCFG of variable granularity, based on attributes incorporated into the PCFG symbols.

<table>
<thead>
<tr>
<th>PTB Sec 23</th>
<th>No Prepositions</th>
<th>Prep. on verbs</th>
<th>Prep. on nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Recall</td>
<td>86.5</td>
<td>86.11</td>
<td>85.98</td>
</tr>
<tr>
<td>Labeled Precision</td>
<td>86.7</td>
<td>86.50</td>
<td>86.3</td>
</tr>
<tr>
<td>Labeled F-score</td>
<td>86.6</td>
<td>86.31</td>
<td>86.14</td>
</tr>
</tbody>
</table>

Number of features on all categories: 19
Some structural features, mostly linguistic features.
Scarcity of lexical data

In training sections of Penn Treebank, \( \sim45000 \) sentences

- Total verb types: \( \sim7450 \), tokens \( \sim125000 \).
- \( \sim2830 \) verb types with occurrence freq 1: 38\% of all types, 2.37\% of all tokens.

<table>
<thead>
<tr>
<th>verb</th>
<th>tag</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>admired</td>
<td>VBD.n.-.- 1.0</td>
<td></td>
</tr>
<tr>
<td>admit</td>
<td>VB.z.-.- 1.0 VB.n.-.- 1.0 VB.b.-.- 3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VBP.z.-.- 1.0 VBP.p.-.- 1.0 VBP.b.-.- 2.0</td>
<td></td>
</tr>
<tr>
<td>admonishing</td>
<td>VBG.s.-.to 1.0</td>
<td></td>
</tr>
<tr>
<td>adopted</td>
<td>VBN.aux.e.fin 2.0 VBD.n.-.- 15.0 VBD.np.-.- 1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VBN.n.-.- 16.0</td>
<td></td>
</tr>
</tbody>
</table>
Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
- Focus on learning lexical parameters.
Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
- Focus on learning lexical parameters.
  - ♦ Lexical parameters obtained from re-estimated model and treebank.
Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
- Focus on learning lexical parameters.
  - Lexical parameters obtained from re-estimated model and treebank.
  - Syntactic parameters obtained from treebank PCFG.
Inside-outside Re-estimation

- Raw data Corpus (Incomplete data)
- Initial PCFG
- Inside-Outside Algorithm
- New Model
Iterative Re-estimation

1. Raw data Corpus (Incomplete data)
2. Initial PCFG
3. Frequency Transformation (Lexical Parameters)
4. Inside-Outside Algorithm
5. New Model
Lexical Transformation

- Constraint on re-estimated lexicons.
- Ensures that re-estimated lexicons are similar to treebank lexicon.
- Linear interpolation of the treebank and the re-estimated lexicons.

\[
\begin{align*}
  d_i(w, \tau, \iota) &= (1 - \lambda)t(w, \tau, \iota) + \lambda \bar{c}_i(w, \tau, \iota) \\
\end{align*}
\]

where

- \( w, \tau, \iota \) word, POS tag, incorporation sequence

Scale Corpus frequencies:

\[
\bar{c}_i(w, \tau, \iota) = \frac{t(\tau, \iota)}{c_i(\tau, \iota)}c_i(w, \tau, \iota)
\]
Initial Model

- Non-novel words
Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
  - Average treebank distribution for that tag.
Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
  - Average treebank distribution for that tag.
  - The re-estimation procedure is expected to acquire a word specific distribution.
Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
  - Average treebank distribution for that tag.
  - The re-estimation procedure is expected to acquire a word specific distribution.
- Words in the corpus (both novel and non-novel) get all possible incorporation values for the POS tag.
The unlabelled corpus is tagged with POS tags in Penn Treebank style (Treetagger) and tokens of words and POS tags are tabulated to obtain a frequency table $g(w, \tau)$.

Each frequency $g(w, \tau)$ is split among possible incorporations $\iota$ in proportion to a ratio of marginal frequencies in $t_0$

$$g(w, \tau, \iota) = \frac{t_0(\tau, \iota)}{t_0(\tau)} g(w, \tau) \quad (2)$$

The tagged corpus is merged with the treebank corpus

$$t(w, \tau, \iota) = (1 - \lambda_{\tau, \iota}) t_0(w, \tau, \iota) + \lambda_{\tau, \iota} g(w, \tau, \iota) \quad (3)$$
Experimental Setup

- Re-estimation: 4 Million words of unannotated Wall Street Journal Corpus (year 1994), sentence-length < 25 words
- Each iteration results in a corresponding model.
Experimental Setup

- Re-estimation: 4 Million words of unannotated Wall Street Journal Corpus (year 1994), sentence-length < 25 words
- Each iteration results in a corresponding model.

Evaluation: Acquiring subcategorization frames of novel verbs.

- 1360 tokens of 117 verb types: all occurrences heldout from treebank training data.
- Tokens of test verbs: preterminal (tag + incorporation sequence) extracted from Viterbi parse.
- Gold standard is the transformed treebank.
Subcat Frames

- Fine-grained subcategorization frames (81 subcategories)
- Intransitive, transitive, ditransitive, clausal, prepositional, etc.
- For clausal frames, the type and subject of clause.
## Subcat. error % for Novel verbs

<table>
<thead>
<tr>
<th>Iteration $i$</th>
<th>Interleaved Procedure</th>
<th>Standard Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>33.36</td>
<td>33.36</td>
</tr>
<tr>
<td>1</td>
<td>*24.40</td>
<td>28.69</td>
</tr>
<tr>
<td>2</td>
<td>*23.45</td>
<td>25.56</td>
</tr>
<tr>
<td>3</td>
<td>*23.05</td>
<td>27.86</td>
</tr>
<tr>
<td>4</td>
<td>*22.89</td>
<td>28.41</td>
</tr>
<tr>
<td>5</td>
<td>*22.81</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>22.83</td>
<td>-</td>
</tr>
</tbody>
</table>

10.55% absolute improvement and 31.6% error reduction
Evaluation

Overall Error reduction: 8.97% (16.8% overall error)
## Evaluation

Overall Error reduction: 8.97% (16.8% overall error)

Incorporating Prepositions into frame

<table>
<thead>
<tr>
<th>Iteration $i$</th>
<th>Subcat Error (No Prep.)</th>
<th>Subcat Error (Prep. on verbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>33.47</td>
<td>34.98</td>
</tr>
<tr>
<td>1</td>
<td>24.40</td>
<td>*25.52</td>
</tr>
<tr>
<td>2</td>
<td>23.45</td>
<td>*25.04</td>
</tr>
</tbody>
</table>
Conclusions

- Framework for adding features to Treebank PCFG: features of interest can be added.
- PCFG formalism simple, and estimation methods well defined.
- Using a Treebank-aligned grammar makes standard and reliable evaluations of re-estimated grammars possible.
- Lexical information induced for novel items; also useful for low frequency items.
Noun Valence

- Three valences s, sbar, p
- NN and NNS (common nouns)

<table>
<thead>
<tr>
<th>Iteration $i$</th>
<th>Noun valence Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>23.13</td>
</tr>
<tr>
<td>1</td>
<td>*20.35 (p &lt; 0.0001)</td>
</tr>
<tr>
<td>2</td>
<td>21.49</td>
</tr>
</tbody>
</table>

Table 1: Noun Valence errors, with 4M words of training data.
# Labeled Bracketing Evaluation

<table>
<thead>
<tr>
<th>Iteration $i$</th>
<th>Interleaved Procedure f-score</th>
<th>Standard Procedure f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>86.55</td>
<td>86.55</td>
</tr>
<tr>
<td>1</td>
<td>86.83</td>
<td>86.96</td>
</tr>
<tr>
<td>2</td>
<td>*86.93</td>
<td>85.93</td>
</tr>
<tr>
<td>3</td>
<td>*86.92</td>
<td>84.87</td>
</tr>
<tr>
<td>4</td>
<td>*86.92</td>
<td>83.77</td>
</tr>
<tr>
<td>5</td>
<td>86.92</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>86.86</td>
<td>-</td>
</tr>
</tbody>
</table>
## Larger Training Data

<table>
<thead>
<tr>
<th>Iteration $i$</th>
<th>Subcat Error 4M words</th>
<th>Subcat Error 8 M words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>33.47</td>
<td>33.47</td>
</tr>
<tr>
<td>1</td>
<td>24.40</td>
<td>24.64</td>
</tr>
<tr>
<td>2</td>
<td>23.45</td>
<td>*22.26 (&gt; 95% conf.)</td>
</tr>
<tr>
<td>3</td>
<td>23.05</td>
<td>22.34</td>
</tr>
<tr>
<td>4</td>
<td>22.89</td>
<td>23.05</td>
</tr>
<tr>
<td>5</td>
<td>22.81</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>22.83</td>
<td>-</td>
</tr>
</tbody>
</table>