Deep Linguistic Processing

- Combinatory Categorial Grammar, Lexical-Functional Grammar, Head-Driven Phrase-Structure Grammar, etc.
- Government and Binding: D-structure VS S-structure

Question: Can we benefit from modeling deep elements?

- Perhaps. Deep grammar formalisms provide more transparent interface to semantics.
- Hard to prove. Grammar Formalism are heterogeneous and hard to be compared.
- Modeling empty category help dependency parsing.
  - Our CoNLL paper: Zhang, Sun and Wan (2017)
  - The dependency tree representation is augmented with empty nodes, which corresponds to unpronounced nominal words.
- Data-driven parsing based on global linear models

Question: How about neural models?

- Is it plausible to detect empty categories using RNNs rather than syntactic information?
- Can neural parsing benefit from modeling empty categories?

Pre-parsing neural empty category detection

- Context of empty categories: sequential context and hierarchical context
- A sequence-oriented model: we explore four sets of annotation specifications
- Tagging based on a BiLSTM-CRF model.

Joint ECD and dependency parsing

- Notation
  - a sentence s with n normal words
  - \( \mathcal{I}_0 = \{(i,j)|i,j \in \{1, \ldots, n\}\} \) : all possible overt dependency edges
  - \( \mathcal{I}_e = \{(i,\phi_i)|i,j \in \{1, \ldots, n\}\} \) : all possible covert dependency edges. \( \phi_i \) denotes an empty node that precede the jth word.
  - \( \mathcal{E} = \{x(i,j) : (i,j) \in \mathcal{I}_0 \cup \mathcal{I}_e\} \) : a dependency parse with empty nodes
- Parsing with ECD can be defined as a search for the highest-scored \( z^*(s) \) in all compatible analyses, just like parsing without empty elements:
  \[
  z^*(s) = \arg \max_{z \in \mathcal{Z}(s)} \text{SCORE}(s, z) = \arg \max_{z \in \mathcal{Z}(s)} \sum_{p \in \mathcal{P}(s)} \text{SCORE}_\text{PART}(s, p)
  \]

A second-order model

The score function over the whole syntactic analysis is defined as:

\[
\text{SCORE}(s, z) = \sum_{(i,j) \in \mathcal{I}(s)} \text{SCORE}_\text{DEP}(s, i, j) + \sum_{(i,\phi) \in \mathcal{I}(s)} \text{SCORE}_\text{EMPTY}(s, i, \phi) + \sum_{(i,j,k) \in \mathcal{I}_e} \text{SCORE}_\text{OVERT}(s, i, j, k) + \sum_{(i,j,k) \in \mathcal{I}_e} \text{SCORE}_\text{COVERT}(s, i, j, k) + \sum_{(i,j,k) \in \mathcal{I}_e} \text{SCORE}_\text{OVERT}(s, i, j, k) + \sum_{(i,j,k) \in \mathcal{I}_e} \text{SCORE}_\text{COVERT}(s, i, j, k)
\]

The score functions

- \( \text{SCORE}_\text{DEP}(s, i, j) \)
- \( \text{SCORE}_\text{EMPTY}(s, i, \phi) \)
- \( \text{SCORE}_\text{OVERT}(s, i, j, k) \)
- \( \text{SCORE}_\text{COVERT}(s, i, j, k) \)

Figure 1: An example of four kinds of annotations, "@@" means interspaces between words.

Overall results

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-parsing</td>
<td>67.3</td>
<td>54.7</td>
<td>60.4</td>
</tr>
<tr>
<td>In-parsing</td>
<td>72.6</td>
<td>55.5</td>
<td>62.9</td>
</tr>
<tr>
<td>In-parsing*</td>
<td>70.9</td>
<td>54.1</td>
<td>61.4</td>
</tr>
<tr>
<td>Xue and Yang (2013)*</td>
<td>65.3</td>
<td>51.2</td>
<td>57.4</td>
</tr>
<tr>
<td>Cai et al. (2011)</td>
<td>66.0</td>
<td>54.5</td>
<td>58.6</td>
</tr>
</tbody>
</table>

Table 1: The overall performance on test data. "+EC" indicates more stringent evaluation metrics.

Empty category helps neural parsing

<table>
<thead>
<tr>
<th></th>
<th>Unlabeled</th>
<th>Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>87.6</td>
<td>88.9</td>
</tr>
<tr>
<td></td>
<td>89.6</td>
<td>88.6</td>
</tr>
</tbody>
</table>

Table 2: Accuracies of both unlabeled and labeled parsing on development data. "EC" indicates parsing without empty categories. "+EC" indicates the second-order in-parsing models. "+EC" indicates parsing models both without and with ECs together.

LSTM is able to find some non-local dependencies

<table>
<thead>
<tr>
<th></th>
<th>Linear CRF</th>
<th>LSTM-CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without POS</td>
<td>With POS</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Interspace</td>
<td>74.6</td>
<td>20.6</td>
</tr>
<tr>
<td>Pre2</td>
<td>72.4</td>
<td>30.1</td>
</tr>
<tr>
<td>Pre3</td>
<td>71.3</td>
<td>30.2</td>
</tr>
<tr>
<td>Prepost</td>
<td>70.9</td>
<td>32.9</td>
</tr>
</tbody>
</table>

Table 3: The overall performance of the two sequential models on development data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61772036, 61331011) and the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

Contact Information

Email: {yuefei.chen, ws, wanxiaojun}@pku.edu.cn

http://www.icst.pku.edu.cn/lcwm