Neural Constituency Parsing

Berkeley

Dan Klein
CS 288

Syntactic Parsing

She enjoys playing tennis.

Syntactic Parsing

Historical Trends

[Slide from Slav Petrov]
Output Correlations

\[
S \rightarrow NP\ VP
\]

Grammars

\[
S \rightarrow NP\ VP
\]

Input-Output Correlations

She enjoys playing tennis.

Span-Based Parsing

She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
... But Will We Get a Tree Out?

Reconciliation

Does It Work?

What’s Going on in There?

Grammar-Based
[Carreras et al, 08]

LSTM-Based
[Stern et al, 17]

92.6

91.0

F1 (English, dev)

Neural parsers no longer have much of the model structure provided to classical parsers.

How do they perform so well without it?
What's Going on in There?

**Why don't we need a grammar?**

Adjacent tree labels are redundant with LSTM features.

If we can predict surrounding tree labels from our LSTM representation of the input, then this information doesn't need to be provided explicitly by grammar production rules.

We find that for 92.3% of spans, the label of the span's parent can be predicted from the neural representation of the span.

---

**Do we need tree constraints?**

*Not for F1*

Many neural parsers no longer model output correlations with grammar rules, but still use output correlations from tree constraints.

Predicting span brackets independently gives nearly identical performance on PTB development set F1 and produces valid trees for 94.5% of sentences.

---

**Is distant context important?**

Almost a full point of F1 is lost by truncating context 5 words away from span endpoints and half a point with 10 words.

---

**Do LSTMs introduce useful inductive bias compared to feedforward networks?**

*Yes!*

We compare a truncated LSTM with feedforward architectures that are given the same inputs.

The LSTM outperformed the best feedforward by 6.5 F1.
She enjoys playing tennis.
Routing with Transformers

Query: She enjoys playing tennis

The verb is: enjoys

verb (VBZ)  | verb (VBG)  | noun | punctuation
---|---|---|---
✓  | ✗  | ✗  | ✗

F1 (English, dev)

<table>
<thead>
<tr>
<th></th>
<th>LSTM</th>
<th>Self-Attentive</th>
<th>+Factored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>61.3</td>
<td>63.3</td>
<td>65.6</td>
</tr>
<tr>
<td>French</td>
<td>92.7</td>
<td>92.7</td>
<td>93.6</td>
</tr>
<tr>
<td>German</td>
<td>77.7</td>
<td>80.3</td>
<td>83.8</td>
</tr>
<tr>
<td>Hebrew</td>
<td>91.0</td>
<td>91.5</td>
<td>92.6</td>
</tr>
<tr>
<td>Hungarian</td>
<td>77.7</td>
<td>80.3</td>
<td>83.8</td>
</tr>
<tr>
<td>Italian</td>
<td>82.5</td>
<td>84.0</td>
<td>84.5</td>
</tr>
<tr>
<td>Korean</td>
<td>89.8</td>
<td>89.9</td>
<td>90.4</td>
</tr>
<tr>
<td>Polish</td>
<td>71.3</td>
<td>71.3</td>
<td>71.8</td>
</tr>
<tr>
<td>Russian</td>
<td>89.8</td>
<td>90.4</td>
<td>90.9</td>
</tr>
<tr>
<td>Swedish</td>
<td>91.5</td>
<td>91.5</td>
<td>91.5</td>
</tr>
</tbody>
</table>

Results: Multilingual

- Björkland et al. (2014)
- Coavoux and Crabbé (2017)
- Cross and Huang (2016)
- Ours
Pre-Training

Problem: Input has more variation than output

Need to handle:
- Rare words not seen during training
- Word forms in morphologically rich languages
- Contextual paraphrase / lexical variation

Historical Trends

Knowledge Modularity

- Knowledge modularity: Learn domain-general knowledge from one data source and use it solve specific problems elsewhere

Context Embeddings and Pretraining

Key Idea: Embed contexts, not words. Use these embeddings for other tasks.

Example: BERT (Devlin et al., 2019) -- bidirectional Transformer trained on masked language modeling and next-sentence prediction
Explosion of Pretraining Work

GLUE SoTA (ICLR 2020)

GLUE Baseline (ICLR 2019)

Insertion-based Generation
KERMIT

Parsing as Span Classification

Pretraining

Architecture

She enjoys playing tennis.
### Encoder Architectures

<table>
<thead>
<tr>
<th></th>
<th>LSTM</th>
<th>Self-Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pre-training</td>
<td>92.08 F1</td>
<td>93.55 F1</td>
</tr>
<tr>
<td></td>
<td>[Gaddy+ 2018]</td>
<td>[Kitaev &amp; Klein 2018]</td>
</tr>
<tr>
<td>Pre-training</td>
<td>95.13 F1</td>
<td>95.60 F1</td>
</tr>
<tr>
<td></td>
<td>(with ELMo)</td>
<td>(with BERT)</td>
</tr>
<tr>
<td></td>
<td>[Kitaev &amp; Klein 2018]</td>
<td>[Kitaev et al 2019]</td>
</tr>
</tbody>
</table>

### Results: Multilingual

<table>
<thead>
<tr>
<th>Language</th>
<th>Björkelund et al. (2014)</th>
<th>Coavoux and Crabbé (2017)</th>
<th>Cross and Huang (2016)</th>
<th>This work (one model per language)</th>
<th>This work (joint multilingual model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>50.1</td>
<td>61.4</td>
<td>83.2</td>
<td>87.5</td>
<td>89.3</td>
</tr>
<tr>
<td>German</td>
<td>76.7</td>
<td>85.5</td>
<td>84.6</td>
<td>86.4</td>
<td>86.8</td>
</tr>
<tr>
<td>Hebrew</td>
<td>56.9</td>
<td>61.4</td>
<td>84.8</td>
<td>84.6</td>
<td>84.6</td>
</tr>
<tr>
<td>Hungarian</td>
<td>74.7</td>
<td>82.7</td>
<td>84.4</td>
<td>88.3</td>
<td>89.4</td>
</tr>
<tr>
<td>Korean</td>
<td>82.6</td>
<td>83.3</td>
<td>84.6</td>
<td>85.8</td>
<td>86.4</td>
</tr>
<tr>
<td>Polish</td>
<td>90.8</td>
<td>90.8</td>
<td>90.8</td>
<td>90.8</td>
<td>90.8</td>
</tr>
<tr>
<td>Russian</td>
<td>89.4</td>
<td>89.4</td>
<td>89.4</td>
<td>89.4</td>
<td>89.4</td>
</tr>
<tr>
<td>Swedish</td>
<td>83.5</td>
<td>84.5</td>
<td>84.5</td>
<td>84.5</td>
<td>84.5</td>
</tr>
<tr>
<td>French</td>
<td>83.1</td>
<td>84.5</td>
<td>86.3</td>
<td>85.8</td>
<td>85.8</td>
</tr>
<tr>
<td>Hungarian</td>
<td>71.4</td>
<td>74.8</td>
<td>78.7</td>
<td>83.2</td>
<td>84.6</td>
</tr>
</tbody>
</table>

### Does Structure Help?

Figure 1: Labelled bracketing F1 versus minimum span length for the English corpora. F1 scores for the In-Order parser with BERT (orange) and the Chart parser with BERT (cyan) start to diverge for longer spans.
Out of Domain Parsing

<table>
<thead>
<tr>
<th></th>
<th>Berkeley F1</th>
<th>BLLIP F1</th>
<th>In-Order F1</th>
<th>Chart F1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Δ Err.</td>
<td>Δ Err.</td>
<td>Δ Err.</td>
<td>Δ Err.</td>
<td></td>
</tr>
<tr>
<td>WSJ Test</td>
<td>90.06</td>
<td>94.18</td>
<td>91.47</td>
<td>93.27</td>
<td>+0.0%</td>
</tr>
<tr>
<td>Brown All</td>
<td>84.64</td>
<td>85.89</td>
<td>85.60</td>
<td>88.04</td>
<td>+5.1%</td>
</tr>
<tr>
<td>Genia All</td>
<td>79.11</td>
<td>79.63</td>
<td>80.31</td>
<td>82.68</td>
<td>+10.2%</td>
</tr>
<tr>
<td>EWT All</td>
<td>77.38</td>
<td>79.91</td>
<td>79.07</td>
<td>82.22</td>
<td>+16.4%</td>
</tr>
</tbody>
</table>

Neural parsers improve out-of-domain numbers, but not more than in-domain numbers.

Open Source Release

Code and models are publicly available at: github.com/nikitakit/self-attentive-parser

Other Neural Constituency Parsers

- Back to at least Henderson 1998!
- Recent directions:
  - Shift-Reduce, eg Cross and Huang 2016
  - SR/Generative, eg Dyer et al 2016 (RNNG)
  - In-Order Generative, eg Liu and Zhang 2017