Neural Constituency Parsing

Dan Klein
CS 288
She enjoys playing tennis.
Syntactic Parsing

She enjoys playing tennis.
Historical Trends

[Slide from Slav Petrov]
Output Correlations

She enjoys playing tennis.
S → NP VP

NP^S → she

VP[enjoys] : S[playing]
She enjoys playing tennis.
Span-Based Parsing

She enjoys playing tennis.
Parsing as Span Classification

She enjoys playing tennis.
Routing with LSTMs

She enjoys playing tennis.
Routing with LSTMs

She enjoys playing tennis.
Routing with LSTMs

Verb at the start

She enjoys playing tennis.
She enjoys playing tennis.
Routing with LSTMs

She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
Non-Constituents

She enjoys playing tennis.

She enjoys playing tennis.

∅

VP
… But Will We Get a Tree Out?

She enjoys playing tennis.
She enjoys playing tennis.
Does It Work?

Grammar-Based
[Carreras et al, 08] 91.0

LSTM-Based
[Stern et al, 17] 92.6

F1 (English, dev)
What’s Going on in There?

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 predicted from the neural representation of the span.

![Diagram showing neural representation of a sentence]
What’s Going on in There?

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.
What’s Going on in There?

Is distant context important?

Yes!

Almost a full point of F1 is lost by truncating context 5 words away from span endpoints and half a point with 10 words.
What’s Going on in There?

**What word representations do we need?**

<table>
<thead>
<tr>
<th>Representation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Only</td>
<td>91.44</td>
</tr>
<tr>
<td>Word and Tag</td>
<td>92.09</td>
</tr>
<tr>
<td>Character LSTM Only</td>
<td>92.24</td>
</tr>
<tr>
<td>Character LSTM and Word</td>
<td>92.22</td>
</tr>
<tr>
<td>Character LSTM, Word, and Tag</td>
<td>92.24</td>
</tr>
</tbody>
</table>
What’s Going on in There?

What about lexicon features?

The character LSTM captures the same information

Heavily engineered lexicons used to be critical to good performance, but neural models typically don’t use them

Word features from the Berkeley Parser (Petrov and Klein 2007) can be predicted with over 99.7% accuracy from the character LSTM representation
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.
Routing with Transformers

Query: verb

She enjoys playing tennis.
Routing with Transformers

Query: verb

She enjoys playing tennis.
Routing with Transformers

Query: verb

She enjoys playing tennis.
Routing with Transformers

Query: verb

The verb is: enjoys

verb [VBZ]   verb [VBG]   noun   punctuation

She   enjoys   playing   tennis   .

The verb is: enjoys
Routing with Transformers

Query: She enjoys

The verb is: enjoys

word=She
verb=enjoys

verb [VBZ]
verb [VBG]
noun
punctuation

She enjoys playing tennis .
Query: verb

The verb is: enjoys

verb [VBZ]  
verb [VBG]  
noun  
punctuation

word=She  
verb=enjoys

She  
enjoys  
playing  
tennis  
.
What Helps?

F1 (English, dev)
Results: Multilingual

- Arabic:
  - Björkelund et al. (2014): 81.3
  - Coavoux and Crabbé (2017): 82.9
  - Cross and Huang (2016): 85.6
  - Ours: 85.6

- Basque:
  - Björkelund et al. (2014): 88.2
  - Coavoux and Crabbé (2017): 88.8
  - Cross and Huang (2016): 89.7
  - Ours: 90.5

- French:
  - Björkelund et al. (2014): 82.5
  - Coavoux and Crabbé (2017): 82.5
  - Cross and Huang (2016): 83.3
  - Ours: 85.5

- German:
  - Björkelund et al. (2014): 81.7
  - Coavoux and Crabbé (2017): 85.3
  - Cross and Huang (2016): 87.7
  - Ours: 90.5

- Hebrew:
  - Björkelund et al. (2014): 89.8
  - Coavoux and Crabbé (2017): 89.9
  - Cross and Huang (2016): 90.4
  - Ours: 90.5

- Hungarian:
  - Björkelund et al. (2014): 91.7
  - Coavoux and Crabbé (2017): 92.3
  - Cross and Huang (2016): 92.7
  - Ours: 85.5

- Korean:
  - Björkelund et al. (2014): 83.8
  - Coavoux and Crabbé (2017): 86.0
  - Cross and Huang (2016): 86.6
  - Ours: 86.6

- Polish:
  - Björkelund et al. (2014): 90.5
  - Coavoux and Crabbé (2017): 93.6
  - Cross and Huang (2016): 93.7
  - Ours: 85.5

- Swedish:
  - Björkelund et al. (2014): 84.0
  - Coavoux and Crabbé (2017): 84.5
  - Cross and Huang (2016): 85.5
  - Ours: 85.5

- Björkelund et al. (2014): 85.6
- Coavoux and Crabbé (2017): 85.4
- Cross and Huang (2016): 85.6
- Ours: 86.25
Data Hunger

Problem: Input has more variation than output

Need to handle:
• Rare words not seen during training
• Word forms in morphologically rich languages
Knowledge Modularity

- Knowledge modularity: Learn domain-general knowledge from one data source and use it solve specific problems elsewhere
**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
Recent Explosion of Pretraining Work

<table>
<thead>
<tr>
<th>Model</th>
<th>URL Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALBERT (Ensemble)</td>
<td>89.4</td>
</tr>
<tr>
<td>ALICE v2 large ensemble (Alibaba DAMO NLP)</td>
<td>89.0</td>
</tr>
<tr>
<td>FreeLB-RoBERTa (ensemble)</td>
<td>88.8</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>88.5</td>
</tr>
<tr>
<td>XLNet-Large (ensemble)</td>
<td>88.4</td>
</tr>
<tr>
<td>MT-DNN-ensemble</td>
<td>87.6</td>
</tr>
<tr>
<td>GLUE Human Baselines</td>
<td>87.1</td>
</tr>
<tr>
<td>Snorkel MeTaL</td>
<td>83.2</td>
</tr>
<tr>
<td>XLM (English only)</td>
<td>83.1</td>
</tr>
<tr>
<td>SemBERT</td>
<td>82.9</td>
</tr>
<tr>
<td>SpanBERT (single-task training)</td>
<td>82.8</td>
</tr>
<tr>
<td>BERT + BAM</td>
<td>82.3</td>
</tr>
<tr>
<td>Span-Extractive BERT on STILTS</td>
<td>82.3</td>
</tr>
<tr>
<td>BERT on STILTS</td>
<td>82.0</td>
</tr>
<tr>
<td>RGLM-Base (Huawei Noah's Ark Lab)</td>
<td>81.3</td>
</tr>
<tr>
<td>BERT: 24-layers, 16-heads, 1024-hidden</td>
<td>80.5</td>
</tr>
<tr>
<td>BERT + Single-task Adapters</td>
<td>80.2</td>
</tr>
<tr>
<td>Macaron Net-base</td>
<td>79.7</td>
</tr>
<tr>
<td>SesameBERT-Base</td>
<td>78.6</td>
</tr>
<tr>
<td>MobileBERT</td>
<td>78.5</td>
</tr>
<tr>
<td>StackingBERT-Base</td>
<td>78.4</td>
</tr>
<tr>
<td>TinyBERT</td>
<td>75.4</td>
</tr>
<tr>
<td>BILSTM+ELMo+Attn</td>
<td>70.0</td>
</tr>
</tbody>
</table>

GLUE Baseline (ICLR 2019)

GLUE SoTA (ICLR 2020)

Human

BERT

Insertion-based Generation

KERMIT

Transformer

Bi-directional LM

GPT

Larger model
More data

GPT-2

Defense

Grover

VideoBERT

CBT

VILBERT

VisualBERT

B2T2

Unicoder-VL

LXMERT

VL-BERT

UNITER

ERNIE (Tsinghua)

ERINE (Baidu)

ERNIE (Baidu) BERT-wwm

ERINE (Baidu) BERT-w7m

By Xiaochi Wang & Zhengyan Zhang @THUNLP

Cross-lingual

Multi-task

Generation

+Knowledge Graph

Cross-modal

Whole-Word Masking

Span prediction
Remove NSP
Longer time
Remove NSP
More data

Permutation LM
Transformer-XL
More data

ERNIE

KnowBert

Neural entity linker

B2T2

Unicoder-VL

LXMERT

VL-BERT

UNITER

ERNIE (Baidu)

ERNIE (Baidu) BERT-wwm

By Xiaochi Wang & Zhengyan Zhang @THUNLP

Cross-lingual

Multi-task

Generation

+Knowledge Graph

Cross-modal

Whole-Word Masking

Span prediction
Remove NSP
Longer time
Remove NSP
More data

Permutation LM
Transformer-XL
More data

ERNIE

KnowBert

Neural entity linker

B2T2

Unicoder-VL

LXMERT

VL-BERT

UNITER

ERNIE (Baidu)

ERNIE (Baidu) BERT-wwm

By Xiaochi Wang & Zhengyan Zhang @THUNLP
She enjoys playing tennis.
She enjoys playing tennis.
She enjoys playing tennis.
Encoder Architectures

No pre-training

Pre-training (with ELMo)

Pre-training (with BERT)

LSTM

Self-Attention

92.08 F1

95.13 F1

95.60 F1

93.55 F1

93.55 F1

95.60 F1

[Kitaev & Klein 2018]

[Kitaev + 2018]

[Kitaev & Klein 2018]
Encoder Architectures

F1 Score (English)

- No pre-training: 93.6
- ELMo: 95.2
- BERT-base: 95.3
- BERT-large: 95.6
- XLNet-large: 96.0

Number of Parameters

- No pre-training: 26M
- ELMo: 107M
- BERT-base: 117M
- BERT-large: 343M
- XLNet-large: 361M
Results: Multilingual

- Arabic: Björkelund et al. (2014) 81.3, Kitaev and Klein (2018) 82.9, This work (one model per language) 88.0, This work (joint multilingual model) 87

- Basque: Coavoux and Crabbé (2017) 88.2, This work (one model per language) 89.7, This work (joint multilingual model) 92

- French: Cross and Huang (2016) 82.5, This work (one model per language) 84.1, This work (joint multilingual model) 87

- German: Björkelund et al. (2014) 81.7, This work (one model per language) 89.3, This work (joint multilingual model) 90

- Hebrew: Coavoux and Crabbé (2017) 89.8, This work (one model per language) 90.4, This work (joint multilingual model) 93

- Hungarian: Cross and Huang (2016) 91.7, This work (one model per language) 92.7, This work (joint multilingual model) 95

- Korean: Björkelund et al. (2014) 83.8, This work (one model per language) 86.0, This work (joint multilingual model) 89

- Polish: Coavoux and Crabbé (2017) 90.5, This work (one model per language) 93.6, This work (joint multilingual model) 96

- Swedish: Cross and Huang (2016) 84.0, This work (one model per language) 85.5, This work (joint multilingual model) 89
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

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

<table>
<thead>
<tr>
<th></th>
<th>Berkeley F1</th>
<th>∆ Err.</th>
<th>BLLIP F1</th>
<th>∆ Err.</th>
<th>In-Order F1</th>
<th>∆ Err.</th>
<th>Chart F1</th>
<th>∆ Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ Test</td>
<td>90.06</td>
<td>+0.0%</td>
<td>91.48</td>
<td>+0.0%</td>
<td>91.47</td>
<td>+0.0%</td>
<td>93.27</td>
<td>+0.0%</td>
</tr>
<tr>
<td>Brown All</td>
<td>84.64</td>
<td>+54.5%</td>
<td>85.89</td>
<td>+65.6%</td>
<td>85.60</td>
<td>+68.9%</td>
<td>88.04</td>
<td>+77.7%</td>
</tr>
<tr>
<td>Genia All</td>
<td>79.11</td>
<td>+110.2%</td>
<td>79.63</td>
<td>+139.1%</td>
<td>80.31</td>
<td>+130.9%</td>
<td>82.68</td>
<td>+157.4%</td>
</tr>
<tr>
<td>EWT All</td>
<td>77.38</td>
<td>+127.6%</td>
<td>79.91</td>
<td>+135.8%</td>
<td>79.07</td>
<td>+145.4%</td>
<td>82.22</td>
<td>+164.2%</td>
</tr>
</tbody>
</table>
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
Thank You!

Berkeley

nlp.cs.berkeley.edu