Parsing Speech: A Neural Approach to Integrating Lexical and Acoustic-Prosodic Information

Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, Mari Ostendorf
Why use prosody?

Pauses locations are correlated with syntax (Grosjean et al., 1979)

Listeners use prosody to resolve syntactic ambiguities (Price et al., 1991)

Prosody signals disfluencies by marking "interruption point" (Shriberg, 1994)
What is important about this work?

(Some) Prosodic features do not need to be manually engineered

Analysis of error types influenced by prosody

Parsing of "edit" nodes is built-in
The task

- Parsing Switchboard dataset
- Removed punctuation and casing (to mimic ASR setting)
- Using known sentence boundaries
- Outputs "linearized constituency trees with normalised preterminals"

Original parse tree

```
S — FRAG
   PP
      IN — about
      NP — PRP — yourself
```

Linearized parse tree

```
(S (FRAG (INTJ (UH uh)) (PP (IN about) (NP (PRP yourself))))
```

Final POS-normalized linearized parse tree

```
(S (FRAG (INTJ XX) (PP XX (NP XX))))
```
The system
Encoder/decoder

Input word representations (R to L)

Concatenation of word, pause, duration and learnt acoustic features

- **Word**
  - Learnable embeddings

- **Pauses**
  - Concatenation of pre and post pause vectors

- **Duration**
  - Duration of word / avg duration of word
  - Backs off to phonemes for rare/unseen
Encoder/decoder

Uses an attention layer

Two different systems are trialled:

- Content aware
- Location aware
Automatically learnt prosodic features

Uses time alignments in the corpus on a word level

Captures three fundamental frequency features and three energy features

Sampled at consistent intervals
Automatically learnt prosodic features

Each word has $N$ different filters applied of $m$ different sizes (creates $Nm$ feature matrix)

Why $m$ different sizes? To capture features on a different time scale

All filters are applied in strides of 1 and produce 1D convolutions

Convolutions are then max pooled
The system
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>flat-F1</th>
<th>fluent</th>
<th>disf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley</td>
<td>85.41</td>
<td>85.91</td>
<td>90.52</td>
<td>83.08</td>
</tr>
<tr>
<td>C-attn</td>
<td>83.33</td>
<td>83.20</td>
<td>90.86</td>
<td>79.94</td>
</tr>
<tr>
<td>CL-attn</td>
<td>87.85</td>
<td>87.68</td>
<td>92.07</td>
<td>85.95</td>
</tr>
</tbody>
</table>

Table 1: Scores of text-only models on the dev set: 2044 fluent and 3725 disfluent sentences. C-attn denotes content-only attention; CL-attn denotes content+location attention.
<table>
<thead>
<tr>
<th>Model</th>
<th>Parse</th>
<th>Disf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley (text only)</td>
<td>85.41</td>
<td>62.45</td>
</tr>
<tr>
<td>CL-attn (text only)</td>
<td>87.85</td>
<td>79.50</td>
</tr>
<tr>
<td>CL-attn text and</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ $p$</td>
<td>88.37</td>
<td>80.24</td>
</tr>
<tr>
<td>+ $\delta$</td>
<td>88.04</td>
<td>77.41</td>
</tr>
<tr>
<td>+ $p + \delta$</td>
<td>88.21</td>
<td>80.84</td>
</tr>
<tr>
<td>+ f0/E-CNN</td>
<td>88.52</td>
<td>80.81</td>
</tr>
<tr>
<td>+ $p + f0/E-CNN</td>
<td>88.45</td>
<td>81.19</td>
</tr>
<tr>
<td>+ $\delta + f0/E-CNN$</td>
<td>88.44</td>
<td>80.09</td>
</tr>
<tr>
<td>+ $p + \delta + f0/E-CNN$</td>
<td><strong>88.59</strong></td>
<td>80.84</td>
</tr>
</tbody>
</table>

Table 2: Parse and disfluency detection F1 scores on the dev set. Flat-F1 scores were consistently 0.1%-0.3% lower for our models, but 0.2% higher for the Berkeley parser (85.64).
<table>
<thead>
<tr>
<th>Model</th>
<th>Parse</th>
<th>Disfl</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL-attn</td>
<td>87.79 (0.11)</td>
<td>78.65 (0.46)</td>
</tr>
<tr>
<td>best model</td>
<td>88.15 (0.41)</td>
<td>80.48 (0.70)</td>
</tr>
</tbody>
</table>

Table 3: Parse and disfluency detection F1 scores on the dev set: mean (and standard deviation) over 10 runs for the baseline text-only model (CL-attn) and the best model with prosody.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parse</th>
<th>Disfl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley</td>
<td>85.87</td>
<td>63.44</td>
</tr>
<tr>
<td>CL-attn</td>
<td>87.99</td>
<td>76.69</td>
</tr>
<tr>
<td><strong>best model</strong></td>
<td><strong>88.50</strong></td>
<td><strong>77.47</strong></td>
</tr>
</tbody>
</table>

Table 4: Parse and disfluency detection F1 scores on the test set. The best model has *statistically significant* gains over the text-only baseline with $p$-value $< 0.02$. 
Figure 2: F1 scores of the text-only model and our best model as a function of sentence length.
<table>
<thead>
<tr>
<th>Model</th>
<th>fluent</th>
<th>disfluent</th>
</tr>
</thead>
<tbody>
<tr>
<td>text-only</td>
<td>92.07</td>
<td>85.90</td>
</tr>
<tr>
<td>best model</td>
<td>92.03</td>
<td>87.02</td>
</tr>
</tbody>
</table>

Table 6: Dev set F1-score of text-only and best model on fluent (2029) vs. disfluent (3689) sentences.\(^\text{10}\)
Berkeley parser analyser

(Kummerfeld et al., 2012)

Image from: https://github.com/jkkummerfeld/berkeley-parser-analyser
<table>
<thead>
<tr>
<th>Error Type</th>
<th>Disfluent Sentences</th>
<th>best model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>text + $p$</td>
<td></td>
</tr>
<tr>
<td>Clause Att.</td>
<td>5.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>Diff. Label</td>
<td>7.6%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Modifier Att.</td>
<td>9.7%</td>
<td>19.1%</td>
</tr>
<tr>
<td>NP Att.</td>
<td>-2.7%</td>
<td>14.5%</td>
</tr>
<tr>
<td>NP Internal</td>
<td>7.8%</td>
<td>7.4%</td>
</tr>
<tr>
<td>PP Att.</td>
<td>10.1%</td>
<td>7.8%</td>
</tr>
<tr>
<td>1-Word Phrase</td>
<td>6.3%</td>
<td>6.8%</td>
</tr>
<tr>
<td>Unary</td>
<td>-1.1%</td>
<td>8.9%</td>
</tr>
<tr>
<td>VP Att.</td>
<td>0.0%</td>
<td>12.0%</td>
</tr>
</tbody>
</table>

Table 7: Relative error reduction over the text-only baseline in the disfluent subset (3689 sentences) of the development set. Shown here are the most frequent error types (with count $\geq 100$ for the text-only model).
How to interpret the results?

- Doesn't appear much better than existing work done over a decade ago
- Author justifications
  - Evaluation metrics: F1 vs flat F1
  - There are known errors in the parses
    - Messes up audio-word alignments
  - It is difficult to compare to existing work which use additional information
    - Includes punctuation (Charniak and Johnson, 2001)
    - Gold part of speech tags
    - Special segmentation (Kahn et al., 2005)
  - Work is on constituency parsing vs dependency
Incorrectly transcribed example

uh uh <i have had> my wife ’s picked up a couple of things saying uh boy if we could refinish that ’d be a beautiful piece of furniture

<missing> inserted
Qualitative examples
The county is mostly Hispanic.

The minorities are mostly Hispanic.

The minorities are mostly Hispanic.
Incorrect example

```
S
  S
  NP  ADVP  VP
  XX  XX  XX  NP  ADVP
  television  sure  makes  XX  XX  XX  PP
  child  rearing  easy  XX  NP
  on  XX
  you
```

```
S
  S
  NP  ADVP  VP
  XX  XX  XX  NP  ADJP
  television  sure  makes  XX  XX  XX  PP
  child  rearing  easy  XX  NP
  on  XX
  you
```
References


