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

Shubham Toshniwal
TTI Chicago
May 10, 2018
Challenges in Parsing Speech

- Why not recognize speech (ASR) & then use a text parser?
  - ASR transcriptions lack punctuation and can have errors
  - Even assuming perfect transcriptions, need to deal with disfluencies
    - Interjections: hmm, uh, um
    - Speech repair: Why didn't he, why didn't she do it?
    - Parentheticals: I mean, I don't need a car
  - Why is conversational speech parsing important? Google Duplex!
Challenges in Parsing Speech

- Why not recognize speech (ASR) & then use a text parser?
- ASR transcriptions lack punctuation and can have errors
- Even assuming perfect transcriptions, need to deal with disfluencies
  - Interjections: hmm, uh, um
  - Speech repair: Why didn’t he, why didn’t she do it?
  - Parentheticals: I mean, I don’t need a car
- Why is conversational speech parsing important?
Challenges in Parsing Speech

- Why not recognize speech (ASR) & then use a text parser?
- ASR transcriptions lack punctuation and can have errors
- Even assuming perfect transcriptions, need to deal with disfluencies
  - Interjections: hmm, uh, um
  - Speech repair: Why didn’t he, why didn’t she do it?
  - Parentheticals: I mean, I don’t need a car
- Why is conversational speech parsing important? Google Duplex!
Prosodic boundaries found to co-occur with syntactic boundaries (Schepman, 2000)

Prosodic cues such as, pause length, pitch patterns, intensity etc can be useful
- Pauses can act like commas
- Rising pitch at the end of sentence can indicate question

*Chicago cops arrest man* *(pause)* *with knife*

*Chicago cops arrest man with knife*
Task

- Constituency parsing of conversational speech
- Assume transcription and word-level alignment of speech signal are given
- Follow the setup of (Vinyals, 2015) to linearize parse tree:

```
(S (S (PRN (S (NP (PRP I) ) (VP (VBP mean) ))) (INTJ (UH uh) ) (EDITED (NP (PRP you) )) (NP (PRP you) ) (VP (VBP try) (S (VP (TO to) ))))))
```

Linearized Parse Tree
(S (S (PRN (S (NP (PRP I) ) (VP (VBP mean) ))) (INTJ (UH uh) ) (EDITED (NP (PRP you) )) (NP (PRP you) ) (VP (VBP try) (S (VP (TO to) )))))))

Final POS-normalized linearized parse tree
(S (S (PRN (S (NP XX ) (VP XX ) ) ) (INTJ XX ) (EDITED (NP XX ) ) (NP XX ) (VP XX (S (VP XX ) ) ) ) ) )

```
 Linearized Parse Tree
(S (S (PRN (S (NP (PRP I) ) (VP (VBP mean) ))) (INTJ (UH uh) ) (EDITED (NP (PRP you) )) (NP (PRP you) ) (VP (VBP try) (S (VP (TO to) ))))))
```
Encoder-Decoder Models

- Use attention-based encoder-decoder model for outputting linearized parsed trees (Vinyals, 2015)
- Also experiment with location-aware attention models (Chorowski, 2015)
Acoustic-Prosodic Features

- Pause ($p$)
- Word duration ($d$)
- Fundamental frequency and Energy contours ($f_0/E$)
Proposed Model
Experimental Setup

- Switchboard-NXT corpus
- Roughly 100K sentences
- Operate at sentence level - remove punctuation and lowercase words (simulating speech recognition output)
- **Baselines:**
  - Text-only encoder-decoder model
  - Berkeley parser: Latent-variable probabilistic context-free grammar (PCFG) parser
- **Evaluation metric:** PARSEVAL F-score
Text-only Models

Dev set results for text-only model

- Refer to the best text-only model, location-aware attention model, referred to as “Best Text” model from hereon.
Text + Acoustic-Prosodic feature Models

Test set results

- Acoustic-Prosodic features improve parsing performance, in particular on disfluent sentences
Ablation on Acoustic-Prosodic Features

- A combination of all acoustic-prosodic features on top of text features gives the best result.
Effect of Sentence Length

Acoustic-Prosodic features help more on longer sentences
Cherrypicked Example
Performance Gain Categorization

Relative error reduction by adding acoustic-prosodic features

- Only analyze disfluent sentences for this analysis
- Analysis done using Berkeley Parser Analyzer (Kummerfeld, 2012)
Conclusion

• Acoustic-prosodic features are useful for constituency parsing
• Particularly useful for disfluent sentences and long sentences
• Future work:
  ◦ Removing the assumption of known sentence boundaries
  ◦ Cleaning up wrong transcriptions in Switchboard
  ◦ Extending this to dependency parsing