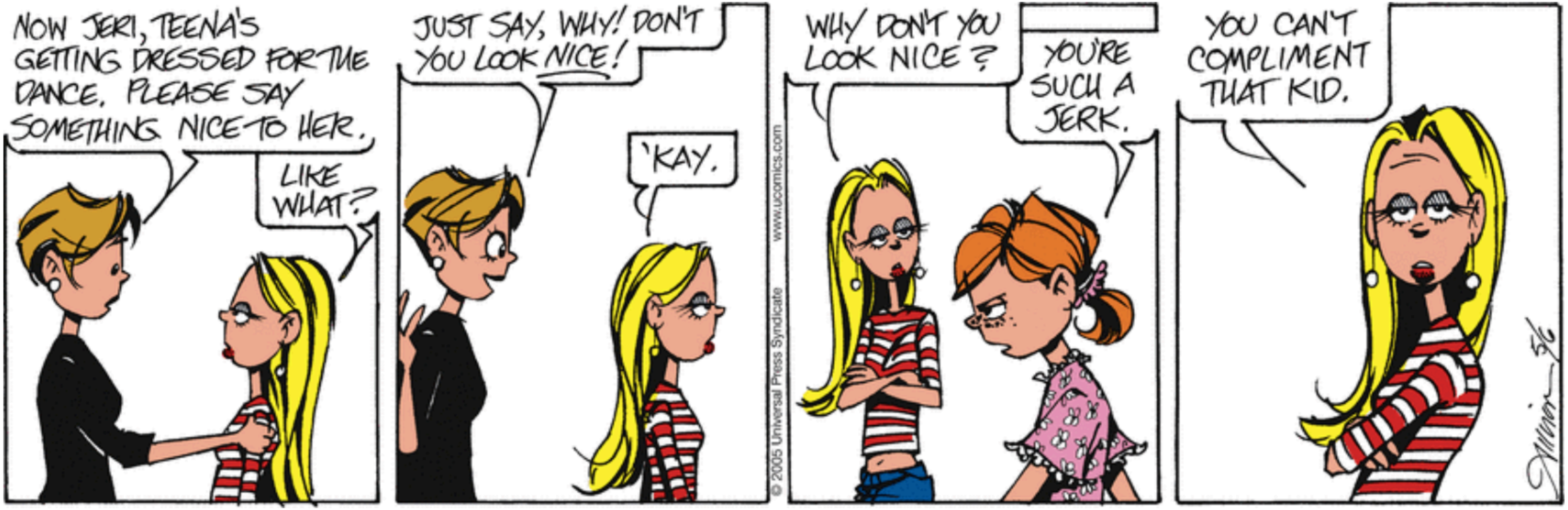


PreTeena by Allison Barrows -- May 6, 2005

On the Role of Style in Parsing Speech with Neural Models

Trang Tran[1] Jiahong Yuan[2] Yang Liu[2] Mari Ostendorf[1]

[1]Electrical & Computer Engineering, University of Washington

[2]LAIX Inc.

Overview: Parsing Speech

- Parsing:
  - Core technology for intermediate language understanding
  - Focus of parsing research and resources: written text
  - But many applications are speech based: dialog systems, translation, spoken document retrieval
- Speech transcripts ≠ written text
  - Not always grammatical; contains disfluencies; lacks punctuation
  - Has prosody: signals structure, intent, focus, …

Background: Constituency Parsing

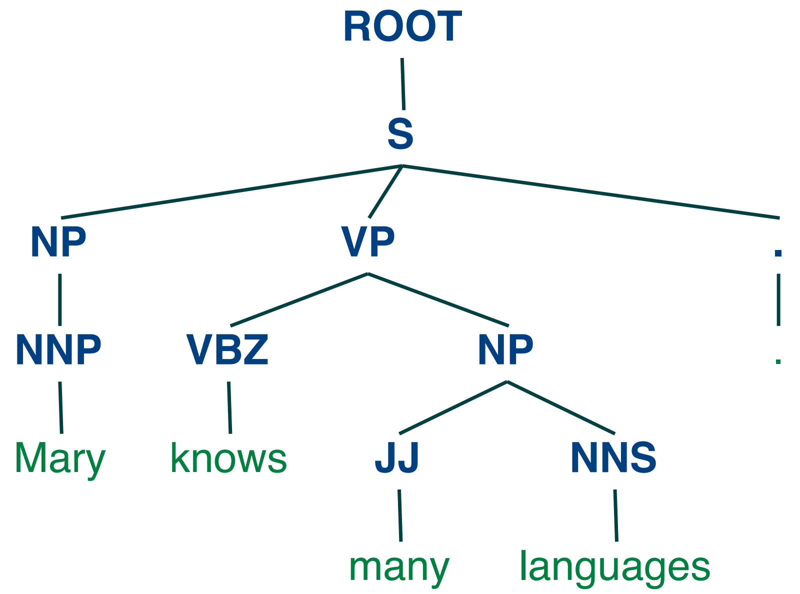
- Parsing: identifying syntactic structure

- SOTA parsers:
  - Neural multi-head self-attention (transformer)
  - Contextual word representations pretrained on large text corpora (ELMo, BERT)

Input: Mary knows many languages .

0 1 2 3 4 5

Output:



Set of spans:

[(ROOT, 0, 5), (S, 0, 5), (NP, 0, 1), (VP, 1, 4), …]

Background: Prosody

- Characteristics of speech beyond words; signals structure
- Acoustic correlates: energy, timing, intonation (f0)

```
      Mary knows many languages you know
```

[pause] [reduced]

“you know” = parenthetical

vs.

```
      Mary knows many languages you know
```

[prominent]

“you know” = subordinate clause

- Domain differences in both words and prosody

conversational/spontaneous speech ≠ read speech

Research Questions: Role of Style?

1. Do contextualized word representations learned for written text also benefit spontaneous speech parsers?

2. Does prosody improve further on top of the rich text information in neural parsers for spontaneous speech?

3. How is the use of prosody affected by mismatch between read and spontaneous speech styles?

Research Questions: Role of <u>Style</u>?

1. Do contextualized word representations learned for written text also benefit spontaneous speech parsers?
   [Spoiler: Yes!]

2. Does prosody improve further on top of the rich text information in neural parsers for spontaneous speech?

3. How is the use of prosody affected by mismatch between read and spontaneous speech styles?

Research Questions: Role of Style?

1. Do contextualized word representations learned for written text also benefit spontaneous speech parsers?
   [Spoiler: Yes!]

2. Does prosody improve further on top of the rich text information in neural parsers for spontaneous speech?
   [Spoiler: Yes!]

3. How is the use of prosody affected by mismatch between read and spontaneous speech styles?

Research Questions: Role of Style?

1. Do contextualized word representations learned for written text also benefit spontaneous speech parsers?
   [Spoiler: Yes!]

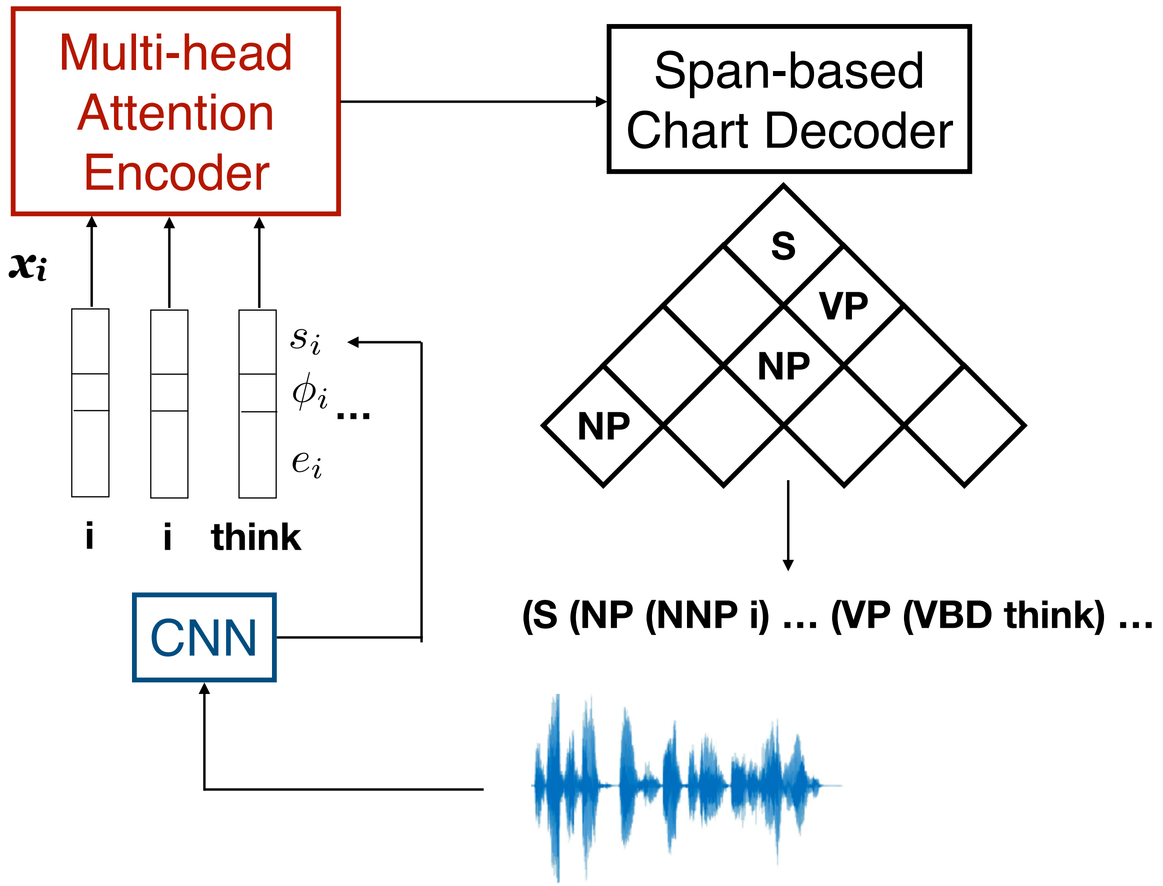
2. Does prosody improve further on top of the rich text information in neural parsers for spontaneous speech?
   [Spoiler: Yes!]

3. How is the use of prosody affected by mismatch between read and spontaneous speech styles?
   [No spoilers!]

Parser Model

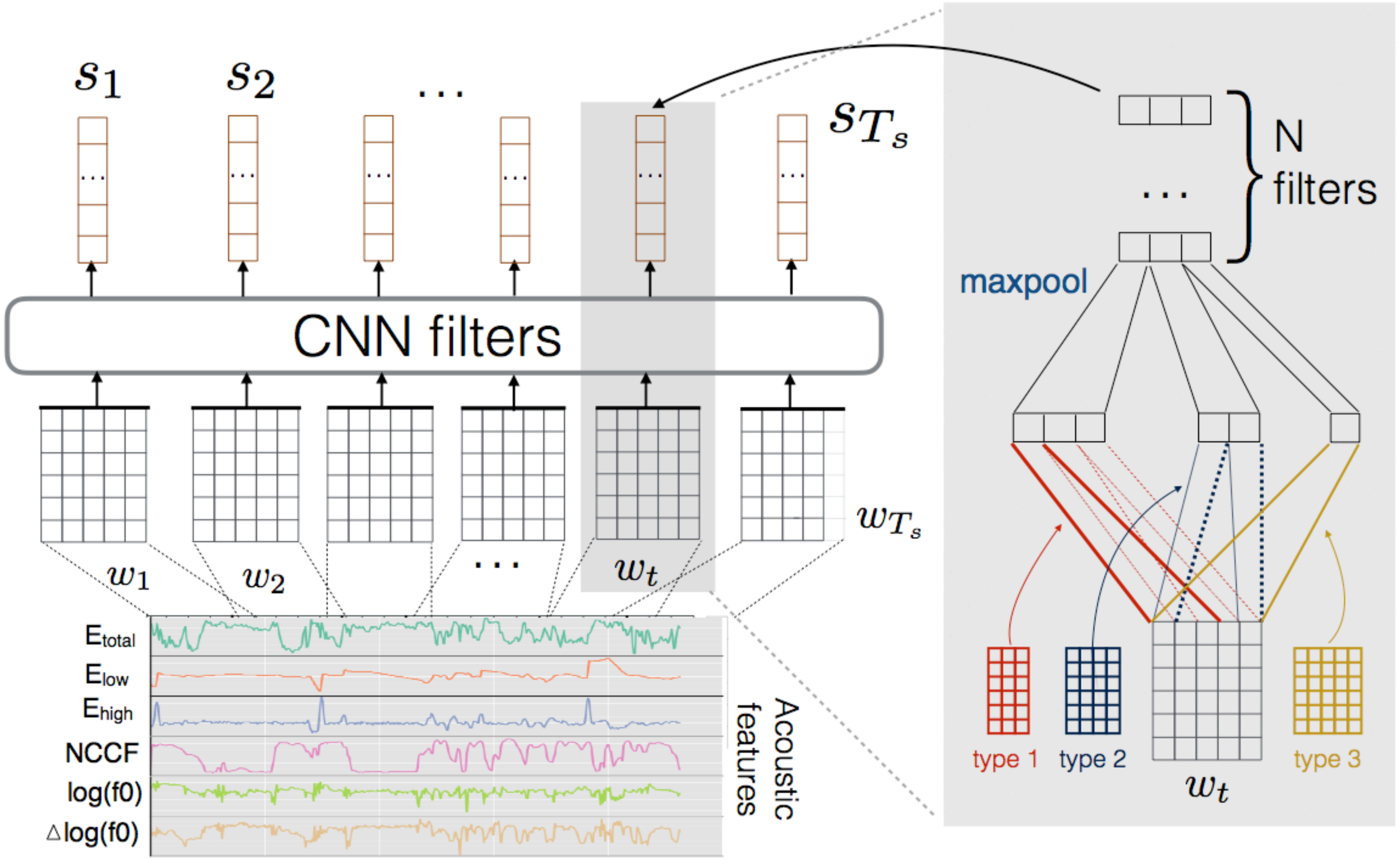
- Parser: transformer encoder + chart decoder (Kitaev & Klein, 2018)
- Word-level features $[x_1, x_2, \dots]$
  - $\boldsymbol{x}_i = [\boldsymbol{e}_i, \boldsymbol{s}_i, \boldsymbol{\phi}_i]$
  - $\boldsymbol{e}_i$: word embeddings
  - $\boldsymbol{s}_i$: f0, energy features
  - $\boldsymbol{\phi}_i$: pause, duration features
- This study: gold transcripts;

  word-level time alignments



Data-driven Prosody Features

- Represent variable-length sequence of features on the word-level
- CNN: summarize f0 & energy contour information (Tran et al., 2018)
- Jointly trained with parser



Results: Q1 – Contextual Embeddings Help

Train	Embedding	F1
WSJ (W)	BERT	77.5
SWBD (S)	Learned	91.0
	GloVe (Fisher)	91.0
	GloVe (Gword)	91.2
	ELMo	92.7
	BERT	93.2
S+W	BERT	93.4

- Training with text alone doesn't work, even with BERT embeddings
- Pretraining on large written text benefits parsing speech
- Training on both (SWBD+WSJ): marginal improvement

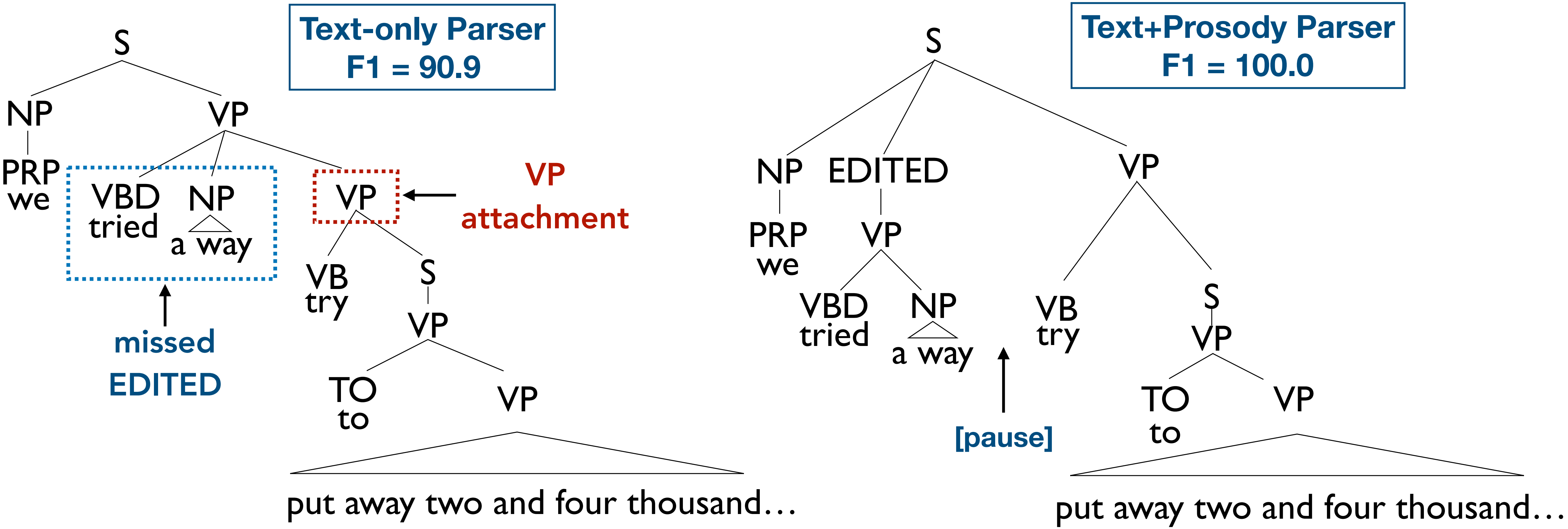
Results: Q2 – Adding Prosody Helps

Model		all	disfluent	fluent
text	ELMo	92.5	91.5	94.6
	BERT	92.9	91.9	94.9
+pros	ELMo	92.7*	91.7*	94.9*
	BERT	93.0*	92.1	95.2*

current best SWBD result

- Further improves over strong text-only parsers
- Helps in disfluent (and long) sentences
- Reduces attachment errors: 19% for VP

Results: Q2 – Adding Prosody Helps



we **[tried a way + try to]** put away two and four thousand…

Results: Q3 – Style Mismatch

Train/Tune	Model	SWBD (C)	GT-N (R)	GT-SW (RC)
SWBD (C)	text	92.9	92.4	98.0
CSR (R)	text	80.6	93.9	91.4
SWBD (C)	+prosody	93.0*	92.6*	98.0
CSR (R)	+prosody	80.4	94.2*	90.3

- Training on conversational (C) speech: minimal degradation on read (R) speech
- Training on (R) speech: significant degradation on (C) → (C) more useful for general training
- Use of prosody differs in (R) vs. (C): style mismatch is both in terms of words and acoustic cues

Results: Q3 – Style Mismatch

Train/Tune	Model	SWBD (C)	GT-N (R)	GT-SW (RC)
SWBD (C)	text	92.9	92.4	98.0
CSR (R)	text	80.6	93.9	91.4
SWBD (C)	+prosody	93.0*	92.6*	98.0
CSR (R)	+prosody	80.4	94.2*	90.3

- Training on conversational (C) speech: minimal degradation on read (R) speech
- Training on (R) speech: significant degradation on (C) → (C) more useful for general training
- Use of prosody differs in (R) vs. (C): style mismatch is both in terms of words and acoustic cues

Results: Q3 – Style Mismatch

Train/Tune	Model	SWBD (C)	GT-N (R)	GT-SW (RC)
SWBD (C)	text	92.9 →	92.4	98.0
CSR (R)	text	80.6 ←	93.9	91.4
SWBD (C)	+prosody	93.0* →	92.6*	98.0
CSR (R)	+prosody	80.4 ←	94.2*	90.3

- Training on conversational (C) speech: minimal degradation on read (R) speech
- Training on (R) speech: significant degradation on (C) → (C) more useful for general training
- Use of prosody differs in (R) vs. (C): style mismatch is both in terms of words and acoustic cues

Results: Q3 – Style Mismatch

Train/Tune	Model	SWBD (C)	GT-N (R)	GT-SW (RC)
SWBD (C)	text	92.9	92.4	98.0
CSR (R)	text	80.6	93.9	91.4
SWBD (C)	+prosody	93.0*	92.6*	98.0
CSR (R)	+prosody	80.4	94.2*	90.3

- Training on conversational (C) speech: minimal degradation on read (R) speech
- Training on (R) speech: significant degradation on (C) → (C) more useful for general training
- Use of prosody differs in (R) vs. (C): style mismatch is both in terms of words and acoustic cues

Conclusion

- Pretrained contextualized word embeddings on text helps constituency parsing of speech
- Using prosody gives further gains, especially in long and disfluent sentences; reducing attachment errors
- Conversational prosody ≠ read prosody
  Conversational prosody is more general, better for training

Thank you!

Backup Slides

Data

Data	Style	Available Material	# Sentences	Used in
WSJ	news text	(gold) parses	40k	Q1
SWBD	conv. speech (C)	audio, (gold) parses	96k	Q1, Q2, Q3
CSR	read news (R)	audio, (silver) parses	8k	Q2, Q3
GT-N	read news (R)	audio, (gold) parses	6k (3k unique)	Q3
GT-SW	read SWBD (RC)	audio, (gold) parses	31 (13 unique)	Q3

Background: Prosody

- Aspects of speech communicating information beyond written words
  - PERmit vs. perMIT; RECord vs. reCORD (meaning)
  - "Mary knows many languages, you know." vs.

    "Mary knows many languages *(that)* you know." (syntax)
  - "You want coffee?" vs. "You want coffee." (intent)
  - "Yeah, sure." vs. "YEAH! SURE!" (sentiment)
- Prosody in human communication: common & essential
- Prosody in AI systems: important but limited
  - Speech (input) understanding: recognition, parsing
  - Speech (output) generation: mostly neutral

Other results from paper: Q1

Train	ELMo	BERT
WSJ	76.0	77.5
SWBD	92.7	93.2
SWBD+WSJ	92.7	93.4

- Parsing result on the SWBD dev set, using only text information, comparing different types of training data.
- The differences between SWBD and SWBD+WSJ are not significant.

Other results from paper: Q2 (length)

Table 5: *Test set F1 scores for different sentence lengths. Prosody shows the most benefit in long sentences.*

Embedding	Model	Sentence lengths (# sents) [0, 5] (2885)	[6, 10] (1339)	[11, -] (1677)
ELMo	text	96.64	96.33	90.53
	+prosody	96.65	96.43	90.81
BERT	text	96.51	96.53	91.07
	+prosody	96.63	96.67	91.30

Other results from paper: Q2 (errors)

Table 6: *Percentage of error reduction counts from text to text+prosody models (first 2 columns) and from ELMo to BERT models (last 2 columns).*

Error Type	Δ(+pros, text)		Δ(BERT, ELMo)	
	ELMo	BERT	text	+pros
Co-ordination	-1.0	-5.1	18.2	14.9
PP Attach.	1.2	1.0	1.2	1.0
NP Attach.	-7.5	0.0	6.0	12.5
VP Attach.	19.2	19.6	-7.7	-7.1
Clause Attach.	8.3	-8.1	11.4	-4.4
Mod. Attach.	7.9	-1.4	11.8	3.0
NP Internal	2.7	7.0	6.5	10.6
1-Word Phrase	5.2	2.3	-3.5	-6.6
Different Label	1.0	7.3	-2.4	4.1