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?

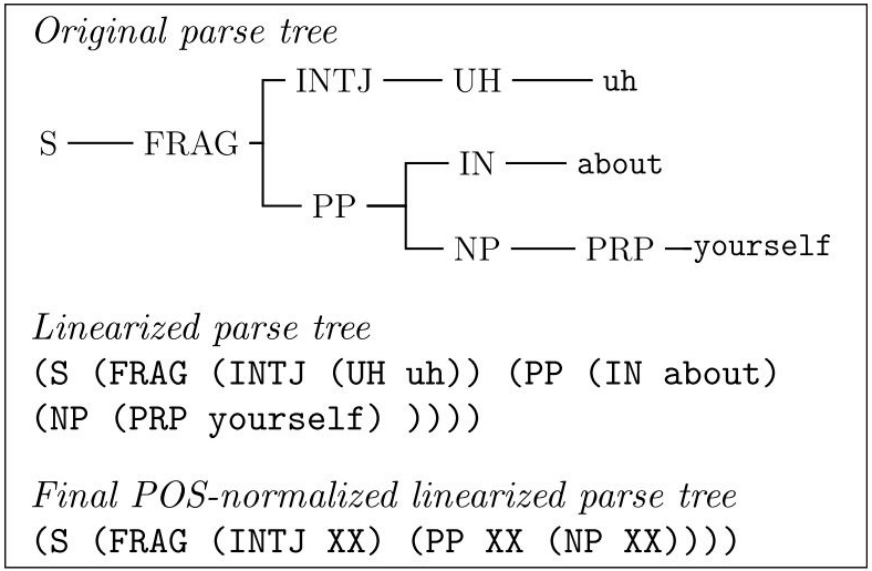
(Some) Prosodic features do not need to be manually engineered

Analysis of error types influenced by prosody

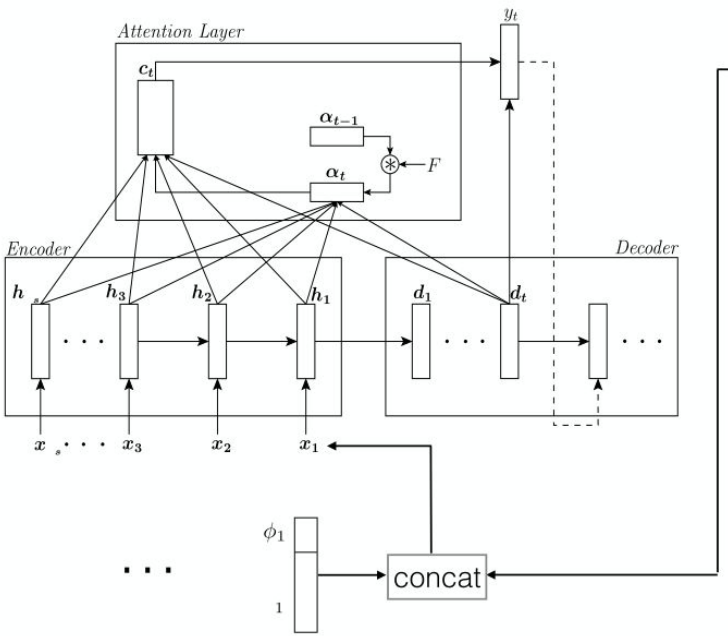
Parsing of "edit" nodes is built-in

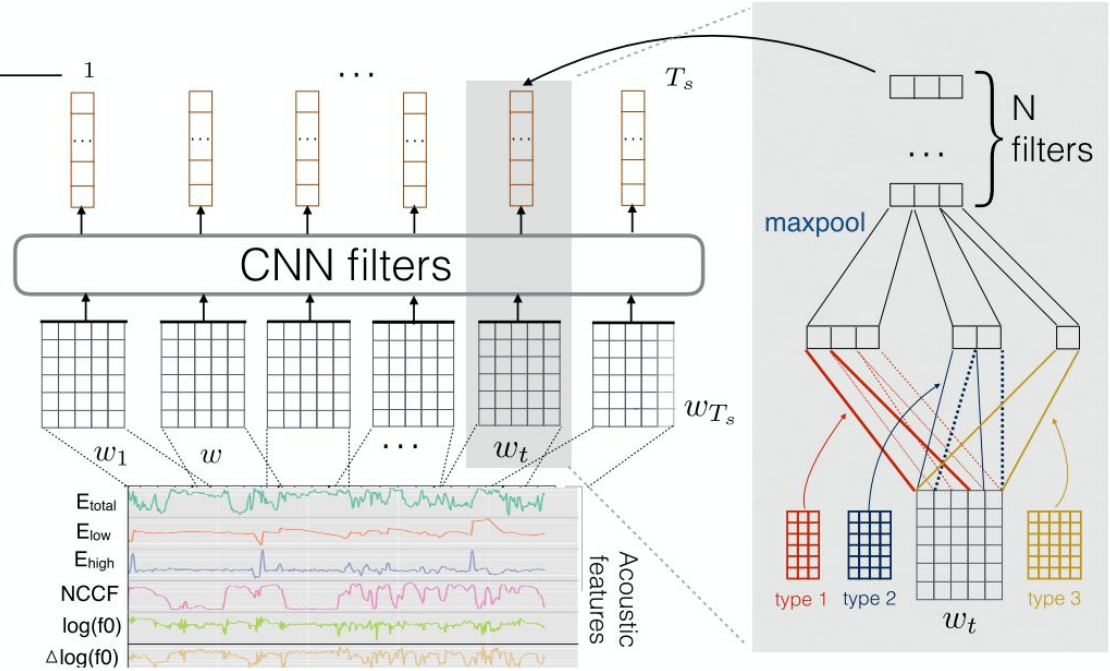
The task

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



The system


Attention Layer
Encoder
Decoder
concat

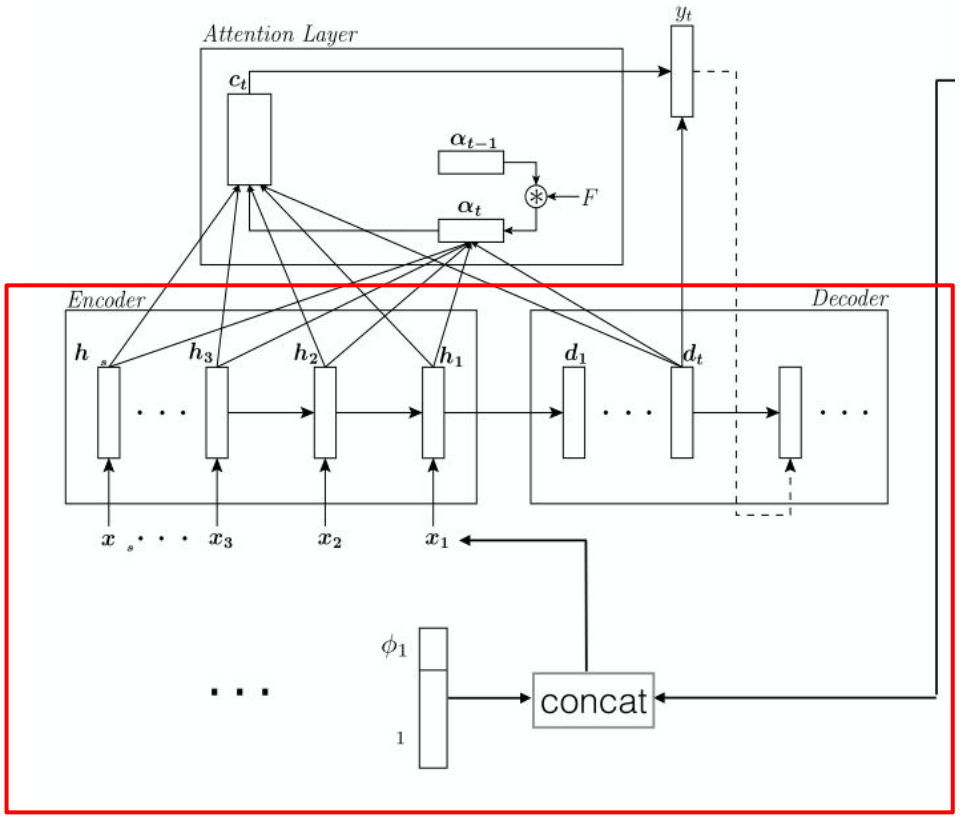

CNN filters
maxpool
N filters
type 1
type 2
type 3
Acoustic features
E_total
E_low
E_high
NCCF
log(f0)
Δlog(f0)

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

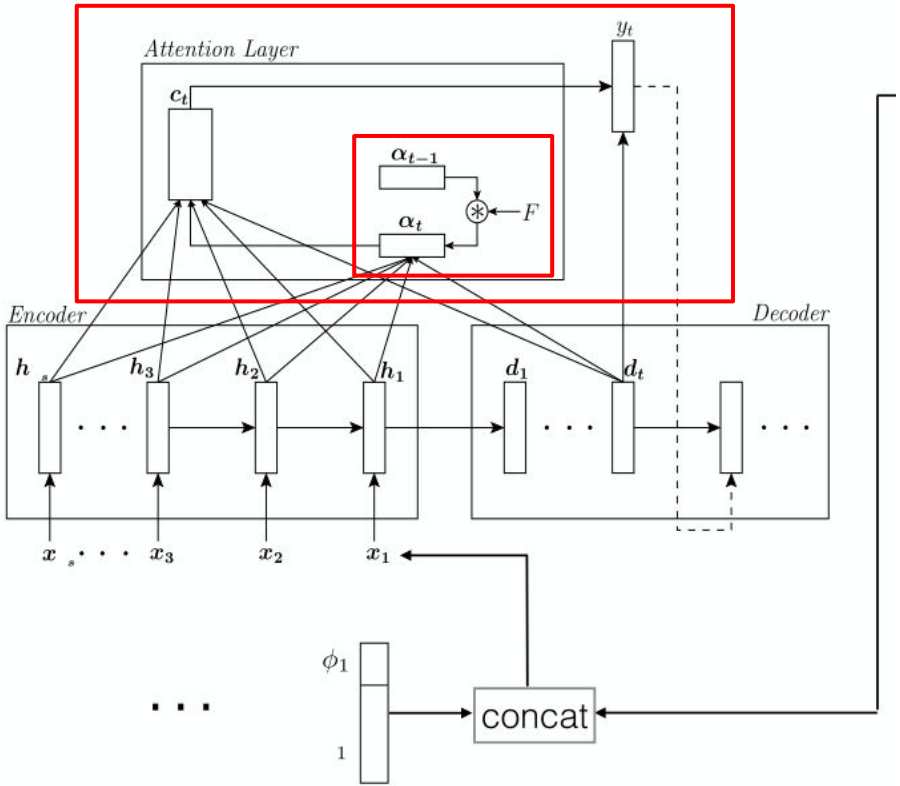


Encoder/decoder

Uses an attention layer

Two different systems are trialled:

- Content aware
- Location aware

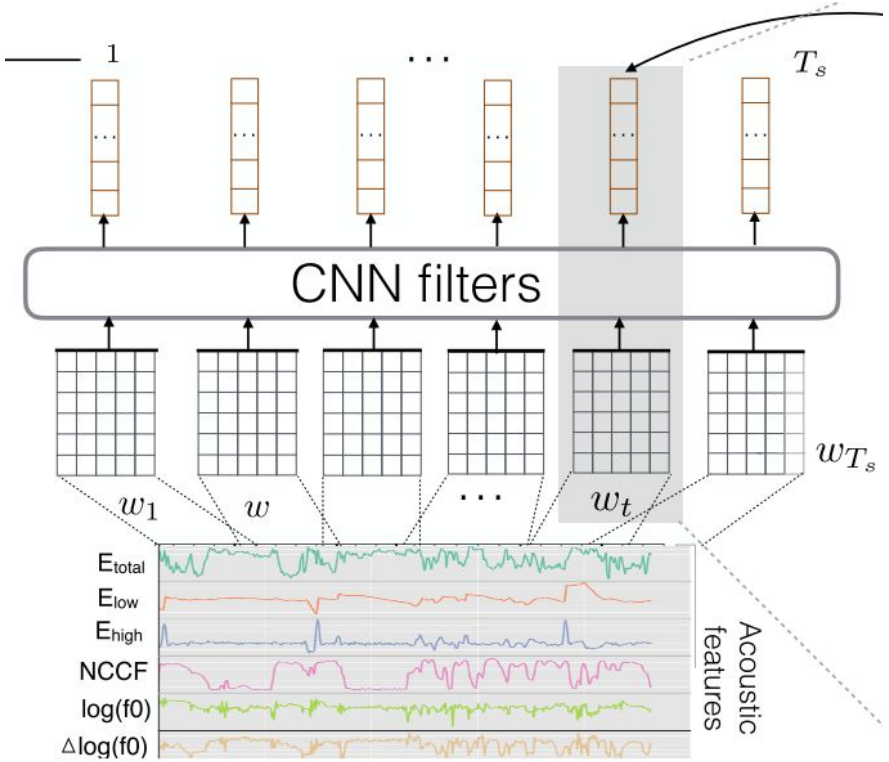


Automatically learnt prosodic features

Uses time alignments in the corpus on a word level

Captures three fundamental frequency features and three energy features

Sampled at consistent intervals



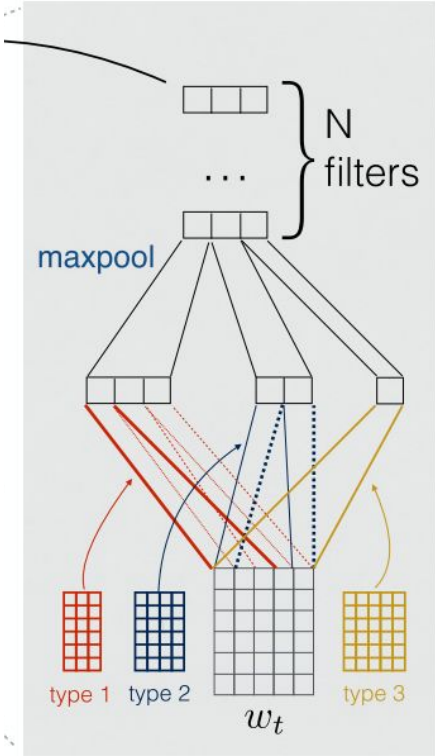
Automatically learnt prosodic features

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

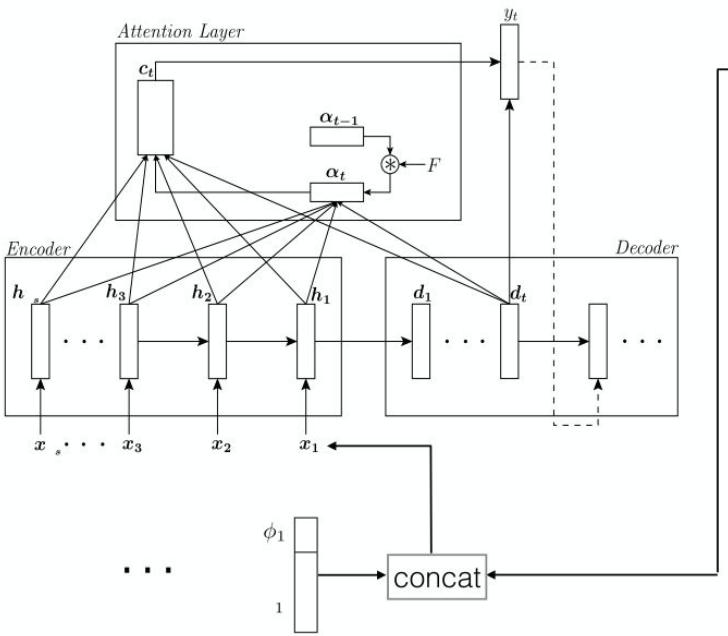
Why *m* different sizes? To capture features on a different time scale

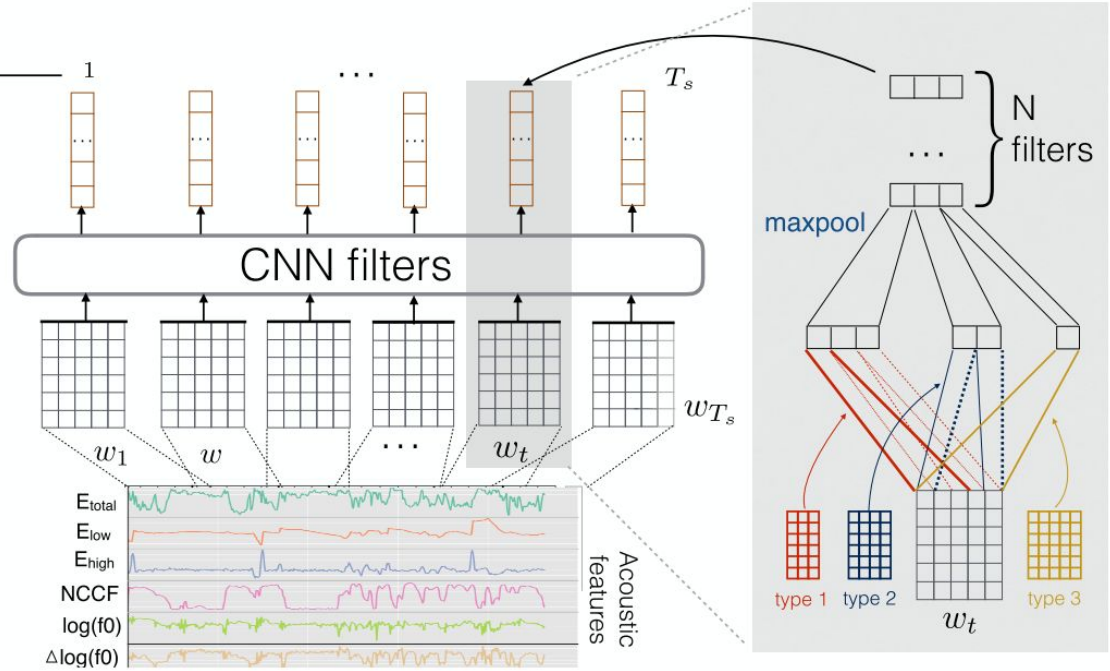
All filters are applied in strides of 1 and produce 1D convolutions

Convolutions are then max pooled



The system


Attention Layer
Encoder
Decoder
concat


CNN filters
maxpool
N filters
type 1
type 2
type 3
Acoustic features
E_total
E_low
E_high
NCCF
log(f0)
Δlog(f0)

Results

Model	F1	flat-F1	fluent	disf
Berkeley	85.41	85.91	90.52	83.08
C-attn	83.33	83.20	90.86	79.94
CL-attn	**87.85**	**87.68**	**92.07**	**85.95**

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.

Model	Parse	Disf
Berkeley (text only)	85.41	62.45
CL-attn (text only)	87.85	79.50
CL-attn text and		
+ p	88.37	80.24
+ δ	88.04	77.41
+ p + δ	88.21	80.84
+ f0/E-CNN	88.52	80.81
+ p + f0/E-CNN	88.45	**81.19**
+ δ + f0/E-CNN	88.44	80.09
+ p + δ + f0/E-CNN	**88.59**	80.84

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).

Model	Parse	Disf
CL-attn	87.79 (0.11)	78.65 (0.46)
best model	88.15 (0.41)	80.48 (0.70)

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.

Model	Parse	Disfl
Berkeley	85.87	63.44
CL-attn	87.99	76.69
best model	**88.50**	**77.47**

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.

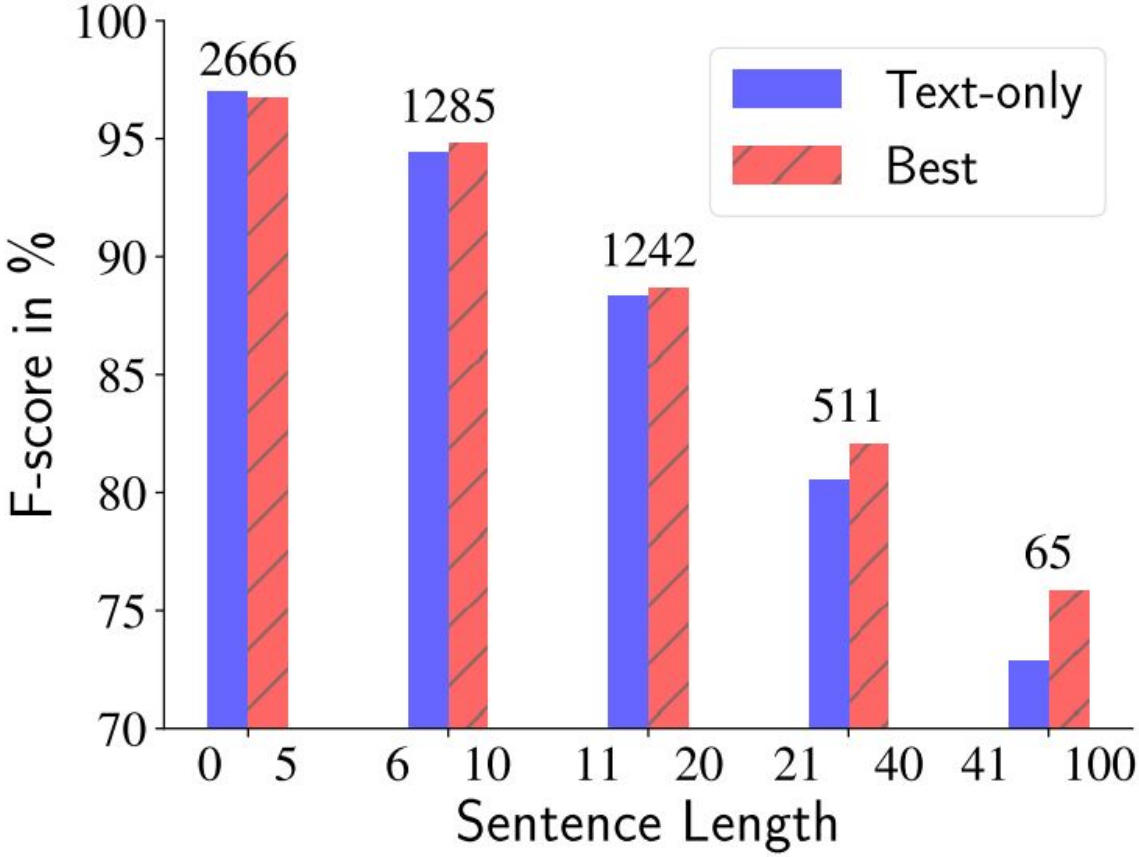


Figure 2: F1 scores of the text-only model and our best model as a function of sentence length.

Model	fluent	disfluent
text-only	92.07	85.90
best model	92.03	**87.02**

Table 6: Dev set F1-score of text-only and best model on fluent (2029) vs. disfluent (3689) sentences.[10]

Berkeley parser analyser

(Kummerfeld et al., 2012)

Image from: https://github.com/jkkummerfeld/berkeley-parser-analyser

Error Type	Disfluent Sentences	
	text + p	best model
Clause Att.	5.7%	1.3%
Diff. Label	7.6%	4.2%
Modifier Att.	9.7%	19.1%
NP Att.	-2.7%	14.5%
NP Internal	7.8%	7.4%
PP Att.	10.1%	7.8%
1-Word Phrase	6.3%	6.8%
Unary	-1.1%	8.9%
VP Att.	0.0%	12.0%

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 ≥ 100 for the text-only model).

How to interpret the results?

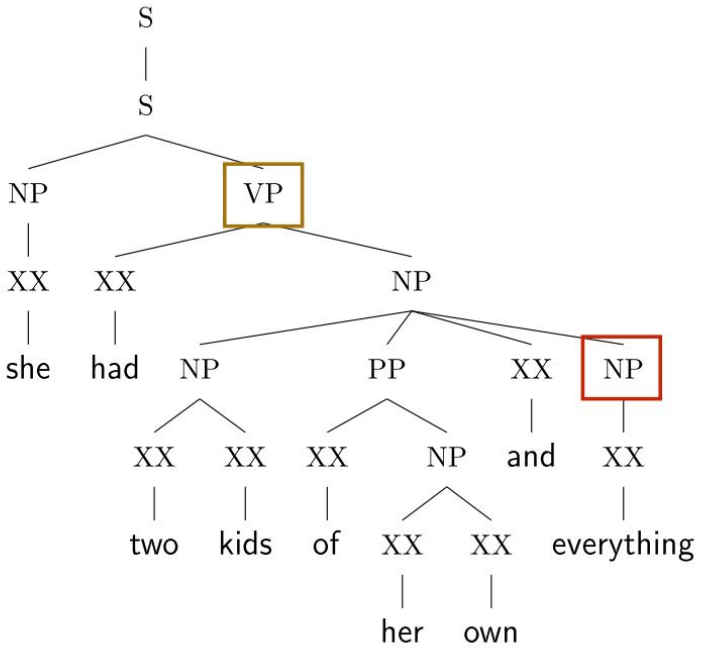
- Doesn't appear much better than existing work done over a decade ago
- Author justifications
  - Evaluation metrics: F1 vs flat F1
  - There are known errors in the parses
    - Messes up audio-word alignments
  - It is difficult to compare to existing work which use additional information
    - Includes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

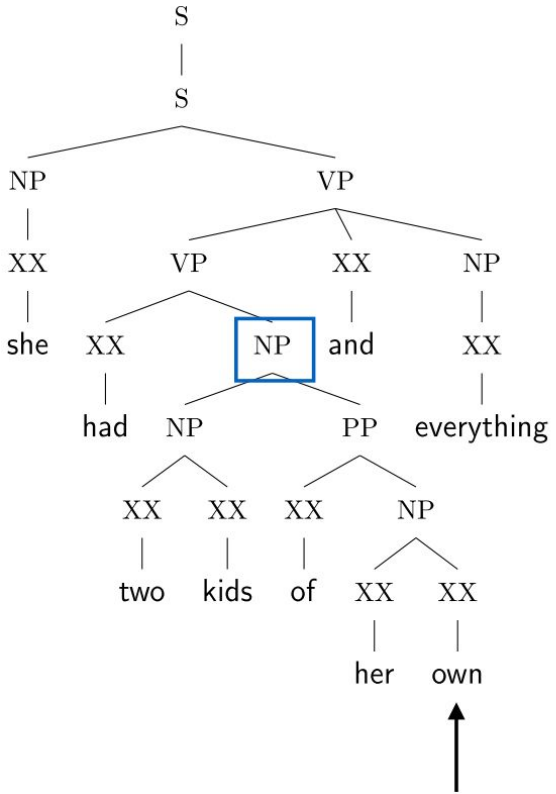
Incorrectly transcribed example

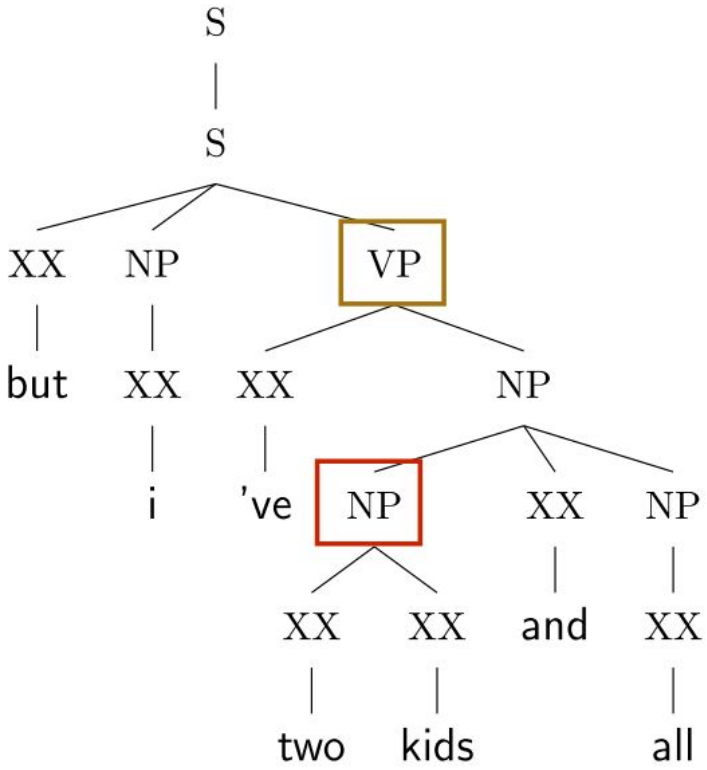
uh uh <i have had> my wife ’s picked up a couple of things saying uh boy if we could refinish that ’d be a beautiful piece of furniture

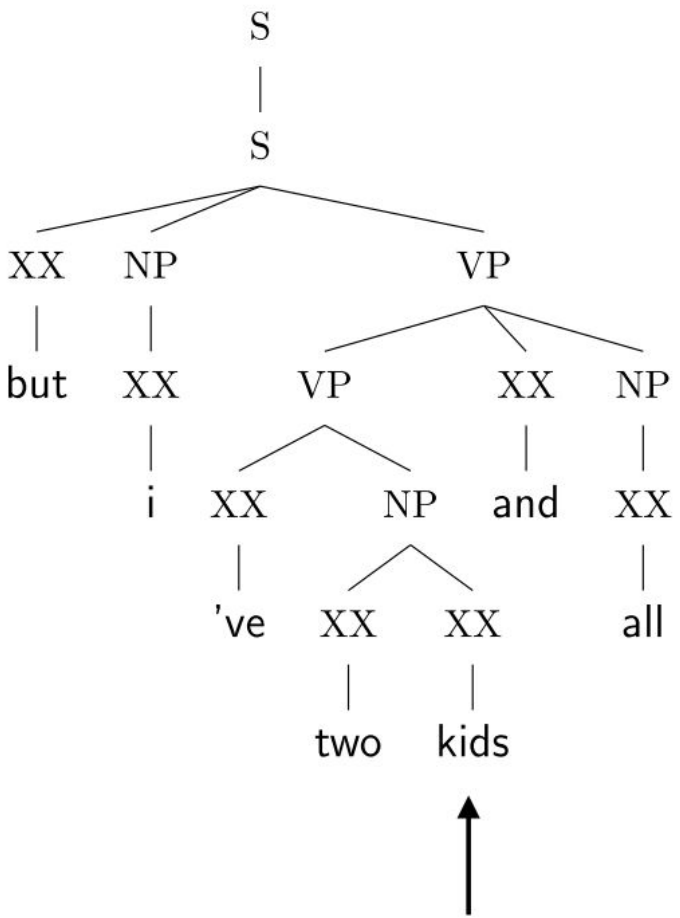
<missing> inserted

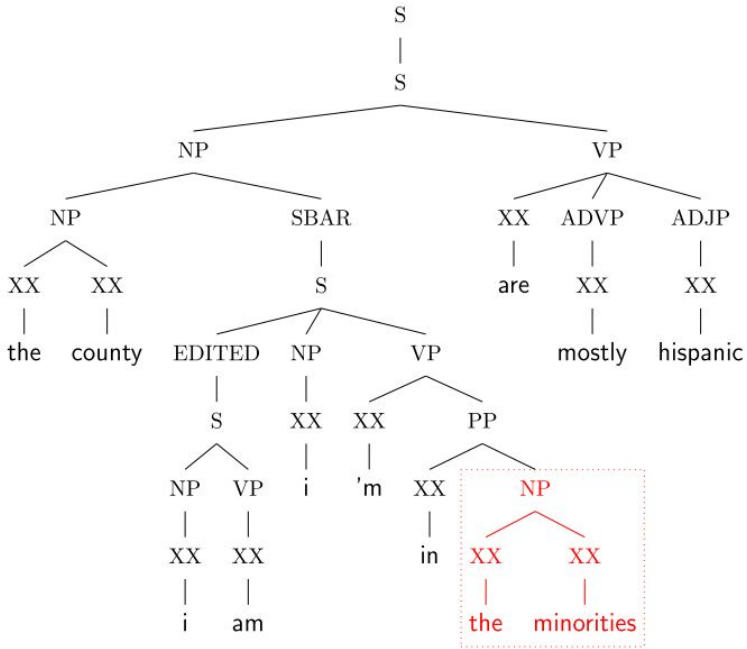
Qualitative examples

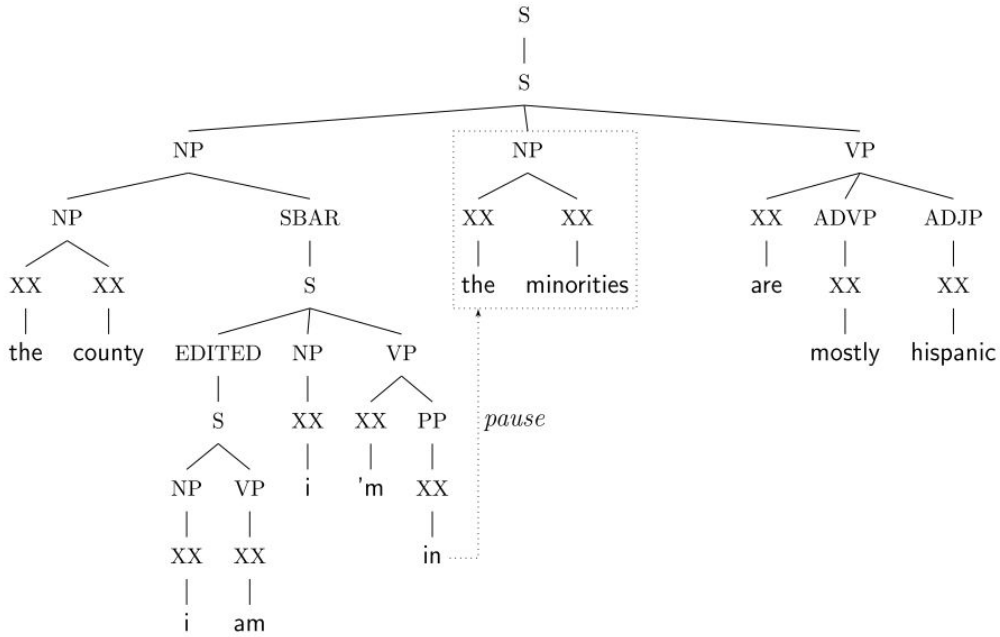




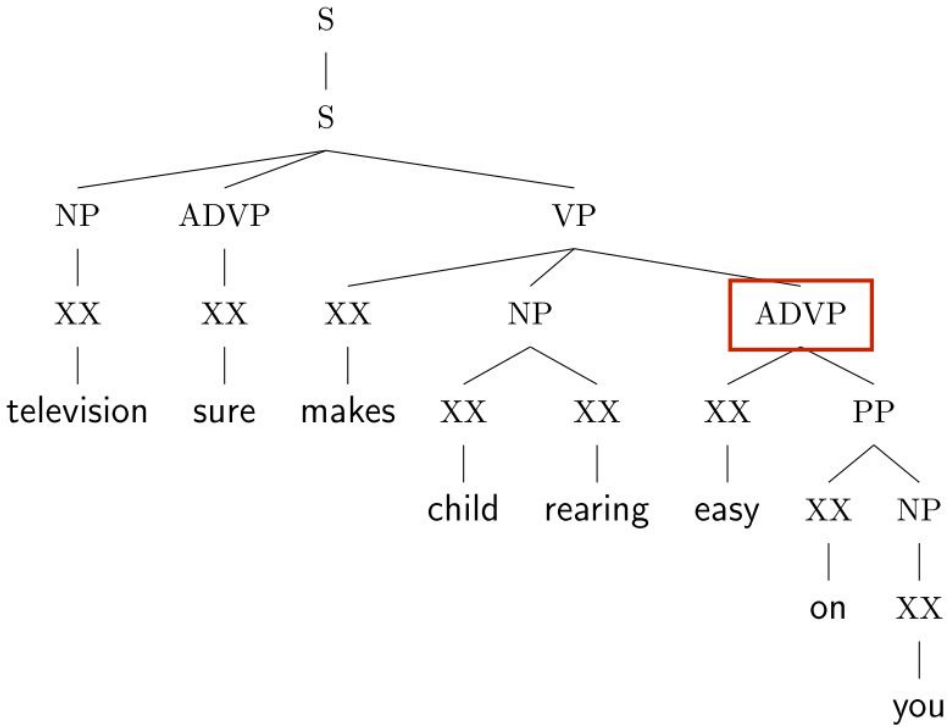

S
S
XX
NP
VP
but
XX
XX
NP
i
've
NP
XX
NP
XX
XX
and
XX
two
kids
all

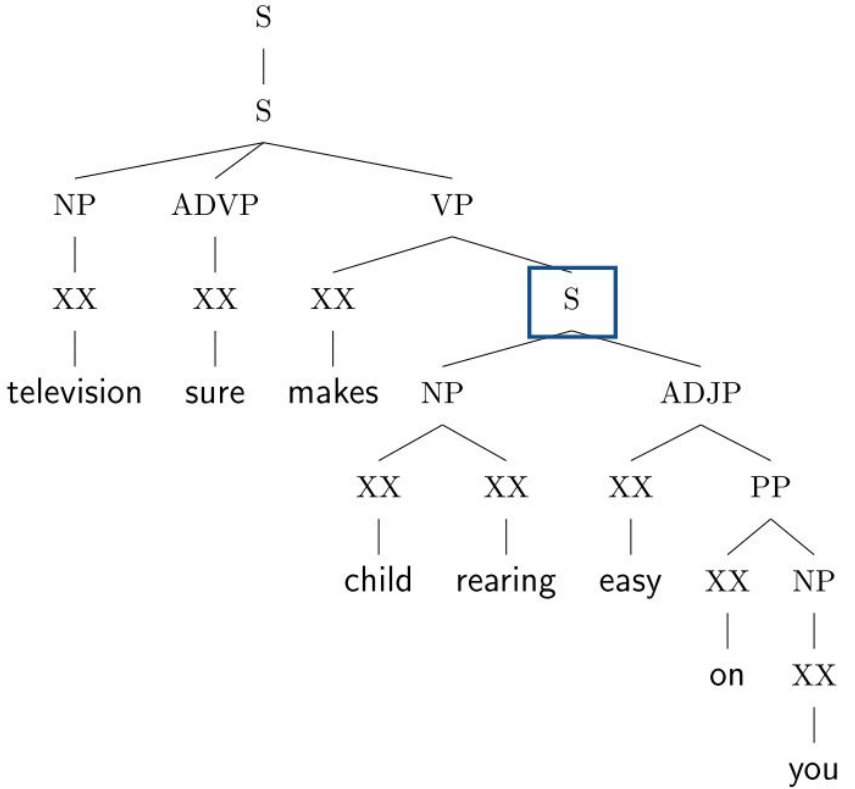

S
S
XX
NP
VP
but
XX
VP
XX
NP
i
XX
NP
and
XX
've
XX
XX
all
two
kids


S
S
NP
VP
NP
SBAR
XX
ADVP
ADJP
XX
XX
S
are
XX
XX
the
county
EDITED
NP
VP
mostly
hispanic
S
XX
XX
PP
NP
VP
i
'm
XX
NP
XX
XX
XX
XX
in
XX
XX
i
am
the
minorities


S
S
NP
NP
VP
NP
SBAR
XX
XX
XX
ADVP
ADJP
XX
XX
S
the
minorities
are
XX
XX
the
county
EDITED
NP
VP
mostly
hispanic
S
XX
XX
PP
pause
NP
VP
i
'm
XX
XX
XX
in
i
am

Incorrect example





References

Charniak, E., & Johnson, M. (2001, June). Edit detection and parsing for transcribed speech. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-9). Association for Computational Linguistics.

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive psychology*, *11*(1), 58-81.

Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. (2005, October). Effective use of prosody in parsing conversational speech. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 233-240). Association for Computational Linguistics.

Kummerfeld, J. K., Hall, D., Curran, J. R., & Klein, D. (2012, July). Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1048-1059). Association for Computational Linguistics.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, *90*(6), 2956-2970.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies* (Doctoral dissertation, University of California, Berkeley).