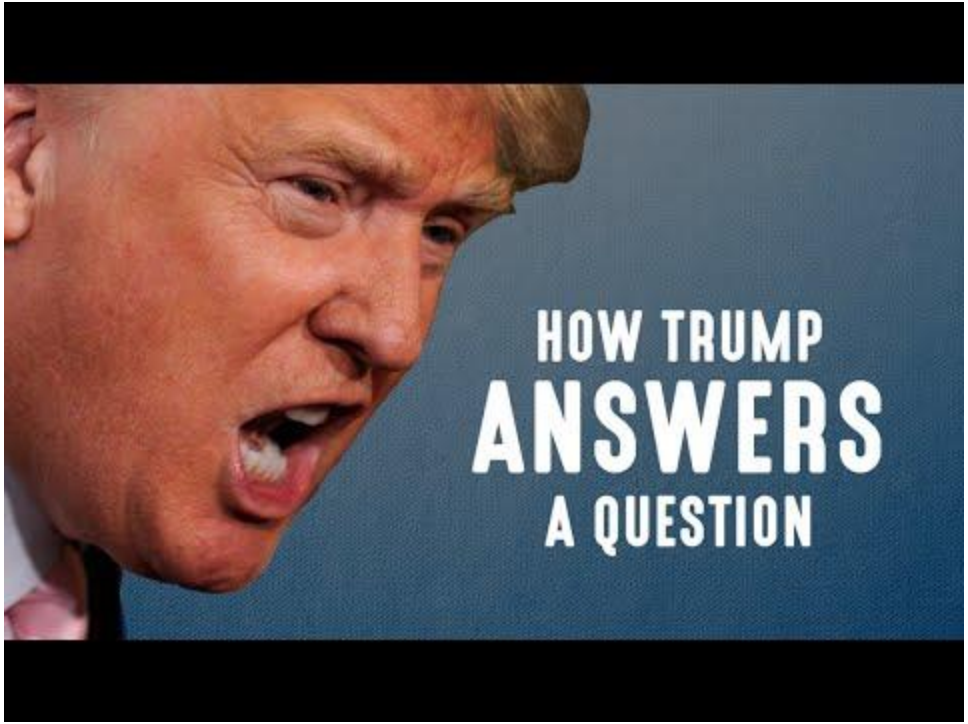
Disfluency Detection using Auto-Correlational Neural Networks

Lou et al, Macquarie University

EMNLP 2018

Presented by Andreea Deac


HOW TRUMP
ANSWERS
A QUESTION

Disfluency

Any interruption in the normal flow of speech:

- False starts: “You look at the two -- look at Paris”
- Corrections
- Repetitions: “with, with, you know..”
- Filled pauses

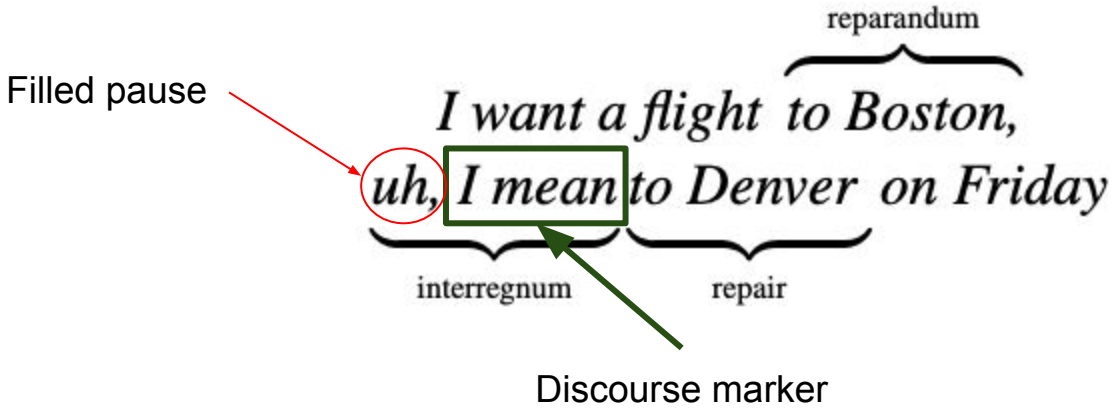
Disfluencies don’t exist in written language, need to identify them in speech to be able to interpret its meaning.

Terminology

- The repair is frequently a “rough copy” of the reparandum

Terminology

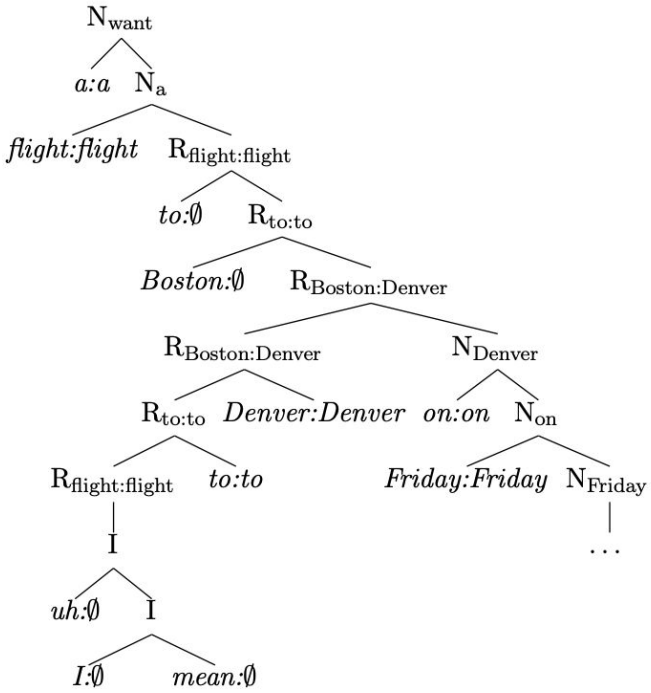
- Filled pauses and discourse markers belong to a closed set of words and phrases



Related work

Noisy channel models [1]

- Use complex tree adjoining grammar



Related work

Parsing-based approaches [2]

- Detect disfluencies while simultaneously identifying the syntactic structure
- Augment transition-based dependency parser with a new action to detect and remove disfluent parts and their dependencies from the stack
- Require large annotated trees!

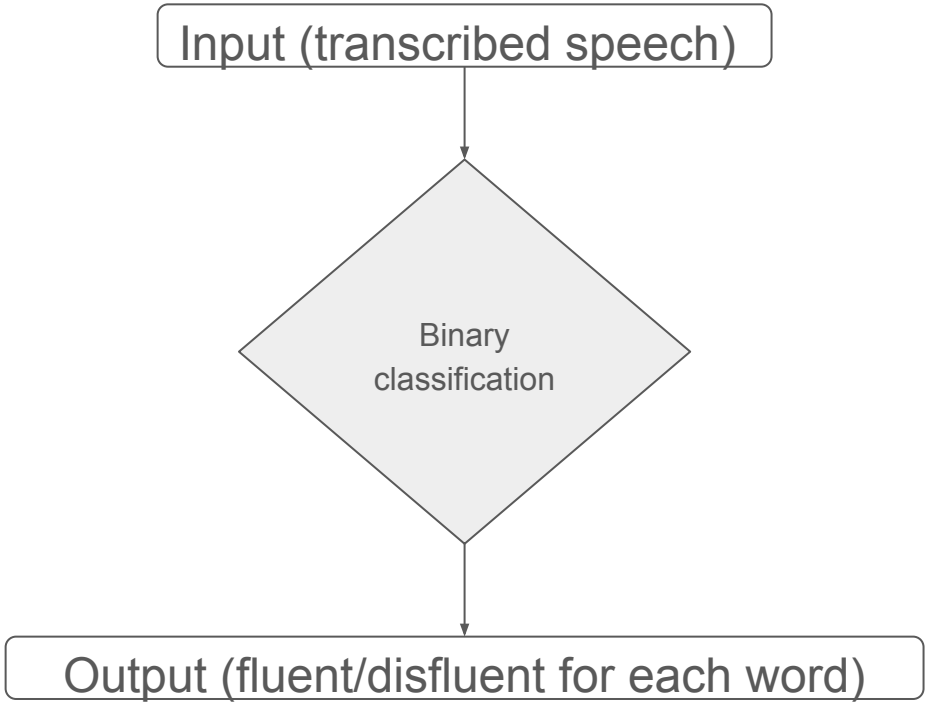
stack	word list	action	relation added
ROOT	she, likes, tea	SHIFT	
ROOT, she	likes tea	SHIFT	
ROOT, she, likes	tea	LeftArc	she ← likes
ROOT, likes	tea	SHIFT	
ROOT, likes, tea		RightArc	likes → tea
ROOT, likes		RightArc	ROOT → likes
ROOT		Done	

Related work

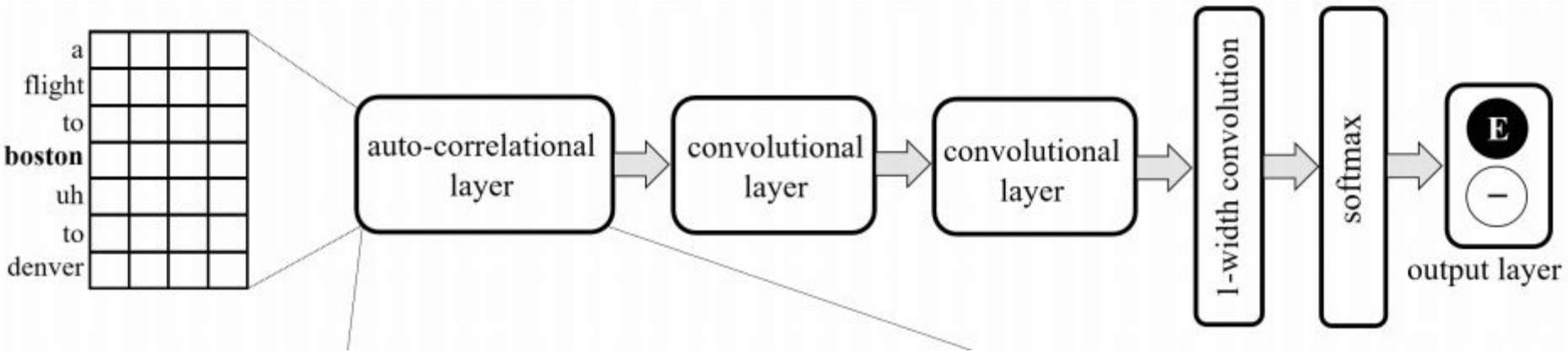
Sequence tagging methods

- Conditional random fields
- HMM
- Deep learning: BLSTM [3]

Improve performance by adding POS tags, hand-crafted pattern match...



ACNN - Architecture



ACNN - 1.Convolution Op recap

- Maps input matrix $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, to an output **y** of length **n**.

$$y_t = A \cdot X_{i:j} + \boldsymbol{b}$$

Where:

- $X_{i:j}$ is a representation of the input substring from word i to word j
- A is a learnable convolutional kernel with size j-i+1
- b is a learnable bias

ACNN - 2. Auto-correlation Op

- In order to capture rough copy dependencies, we want the layer to additionally process the similarity between parts of the input
- Generalisation of the convolution operator:

$$y_t = A \cdot X_{i:j} + B \cdot \hat{X}_{i:j,i:j} + \boldsymbol{b}$$

Where:

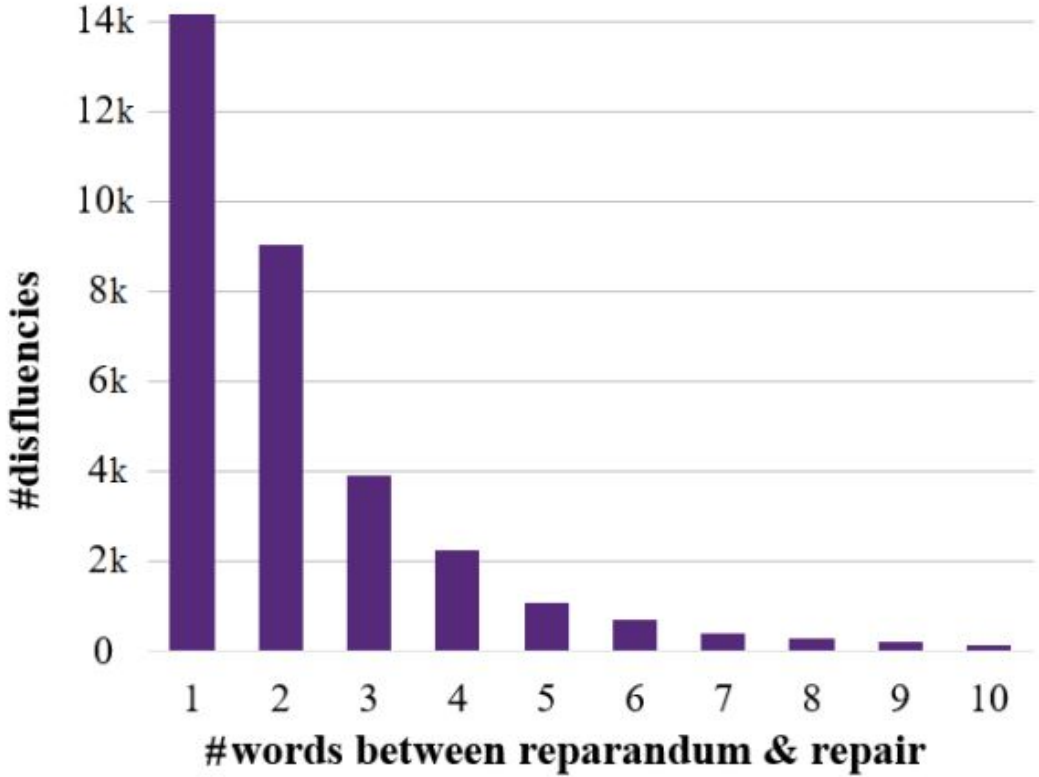
$\hat{X}_{i:j,i:j}$ is the relevant similarity submatrix between input words

$\hat{X}_{i,j,:}$ is given by $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where f is a similarity function, e.g. $f(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u} \circ \boldsymbol{v}$

B is a learned convolutional kernel

Data - Switchboard

- 260 hours of speech
- 2,400 two-sided telephone conversations among 543 speakers
- 70 topics
- Size of gap between reparandum and repair follows a power law



Results - Quantitative analysis

model	P	R	F
BLSTM (words)*	87.8	71.1	78.6
LSTM (words)*	87.6	71.4	78.7
CNN	89.4	74.6	81.3
ACNN	**90.0**	**82.8**	**86.2**

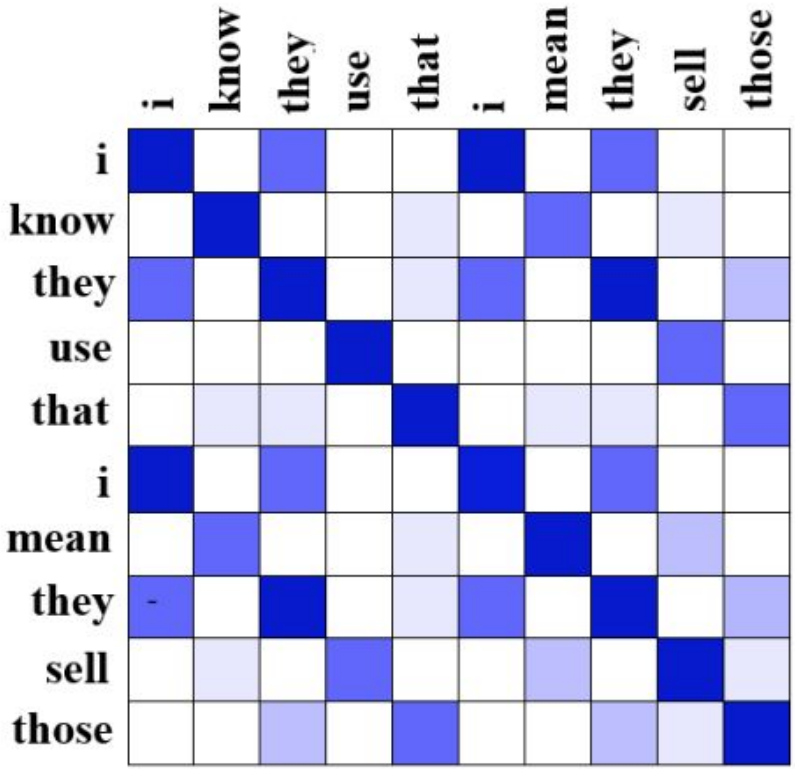
Results

model	Rep.	Cor.	Res.	All
CNN	93.3	66.0	57.1	80.4
ACNN	97.5	80.0	57.1	88.9

Results

model	P	R	F
Yoshikawa et al.(2016) ⋄	67.9	57.9	62.5
Georgila et al. (2010) †	77.4	64.6	70.4
Tran et al. (2018) ⊗ ⋆	-	-	77.5
Kahn et al. (2005) ⋆	-	-	78.2
Johnson et al. (2004) ≀	82.0	77.8	79.7
Georgila (2009) †	-	-	80.1
Johnson et al. (2004) †≀	-	-	81.0
Rasooli et al. (2013) ⋄	85.1	77.9	81.4
Zwarts et al. (2011) ⋈ ≀	-	-	83.8
Qian et al. (2013) ⋈	-	-	84.1
Honnibal et al. (2014) ⋄	-	-	84.1
ACNN	**89.5**	**80.0**	**84.5**
Ferguson et al. (2015) ⋆	90.0	81.2	85.4
Zayats et al. (2016) ⊗†	91.8	80.6	85.9
Jamshid Lou et al. (2017) ⋈ ≀	-	-	86.8

Results - Qualitative analysis



Results: +ves

Detects:

- repetition:

  Uh, I have never even I have never even looked at one closely.

Results: +ves

- correction:

  But I know that in some I know in a lot of rural areas they’re not that good.

Results: +ves

- multiple/nested disfluencies

  They’re handy uh they they come in handy at most unusual times.

Results: +ves

- stutter-like repetitions

Well I I I think we did I think we did learn some lessons that we weren’t uh we weren’t prepared for.

Results: -ves

- fluent repetitions

  But uh I’m afraid I’m I’m probably in the minority

Results: -ves

- long distance between reparandum and repair words

My point was that there is for people who don’t want to do the military service it would be neat if there were an alternative . . .

Results: -ves

- disfluent words which cause ungrammaticality

  Did you you framed it in uh on on you framed in new square footage

Results: ACNN vs CNN -- highlight pred

- ACNN better at detecting “rough copies”
  - Repetition

Uh well **I** actually my dad’s my dad’s almost ninety

  - Correction

Not a man not a repair man but just a friend

  - Stutter-like disfluency

So **we’re** we’re part we’re actually part of MIT

Questions?

References

[1] Paria Jamshid Lou and Mark Johnson. 2017. Disfluency detection using a noisy channel model and a deep neural language model. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17), pages 547–553.

[2] Mohammad Sadegh Rasooli and Joel Tetreault. 2013. Joint parsing and disfluency detection in linear time. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13), pages 124–129, Seattle, USA.

[3] Victoria Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional LSTM. In Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH'16), pages 2523– 2527, San Francisco, USA