Parsing transcripts of speech

Andrew Caines[1], Michael McCarthy[2] & Paula Buttery[1]

[1]University of Cambridge

[2]University of Nottingham

Speech-Centric NLP, 7 September 2017

Background

- Speech (can be) very different from writing
- Put phonetics & prosody aside for now
- Focus on the transcribed form: lexis, morphology, syntax
- Most NLP tools trained on (newswire) written language
- How well do they cope with spoken data?

Background

- Speech (can be) very different from writing
- Put phonetics & prosody aside for now
- Focus on the transcribed form: lexis, morphology, syntax
- Most NLP tools trained on (newswire) written language
- How well do they cope with spoken data?

Background

- Speech (can be) very different from writing
- Put phonetics & prosody aside for now
- Focus on the transcribed form: lexis, morphology, syntax
- Most NLP tools trained on (newswire) written language
- How well do they cope with spoken data?

Background

- Speech (can be) very different from writing
- Put phonetics & prosody aside for now
- Focus on the transcribed form: lexis, morphology, syntax
- Most NLP tools trained on (newswire) written language
- How well do they cope with spoken data?

Background

- Speech (can be) very different from writing
- Put phonetics & prosody aside for now
- Focus on the transcribed form: lexis, morphology, syntax
- Most NLP tools trained on (newswire) written language
- How well do they cope with spoken data?

Speech versus Writing

- Fundamental difference: lack of sentence unit as used in writing; instead speech-units (SUs)

  (Moore et al. 2016 *COLING*)
- And disfluencies –
  - filled pauses: um he's a closet yuppie is what he is
  - repetitions: I played, I played against um
  - false starts: You're happy to – welcome to include it

  (Moore et al. 2015 *TSD*)
- Features of conversation: turn-taking, overlap, co-construction, *etc*

Speech versus Writing

- Fundamental difference: lack of sentence unit as used in writing; instead SPEECH-UNITS (SUs)

  (Moore et al. 2016 *COLING*)
- And disfluencies –
  - FILLED PAUSES: um he's a closet yuppie is what he is
  - REPETITIONS: I played, I played against um
  - FALSE STARTS: You're happy to – welcome to include it

  (Moore et al. 2015 *TSD*)
- Features of conversation: turn-taking, overlap, co-construction, *etc*

Speech versus Writing

- Fundamental difference: lack of sentence unit as used in writing; instead SPEECH-UNITS (SUs)

  (Moore et al. 2016 *COLING*)
- And disfluencies –
  - FILLED PAUSES: um he's a closet yuppie is what he is
  - REPETITIONS: I played, I played against um
  - FALSE STARTS: You're happy to – welcome to include it

  (Moore et al. 2015 *TSD*)
- Features of conversation: turn-taking, overlap, co-construction, *etc*

Speech versus Writing

- Fundamental difference: lack of sentence unit as used in writing; instead SPEECH-UNITS (SUs)

  (Moore et al. 2016 *COLING*)
- And disfluencies –
  - FILLED PAUSES: um he's a closet yuppie is what he is
  - REPETITIONS: I played, I played against um
  - FALSE STARTS: You're happy to – welcome to include it

  (Moore et al. 2015 *TSD*)
- Features of conversation: turn-taking, overlap, co-construction, *etc*

Speech versus Writing

- Fundamental difference: lack of sentence unit as used in writing; instead SPEECH-UNITS (SUs)

  (Moore et al. 2016 *COLING*)
- And disfluencies –
  - FILLED PAUSES: um he's a closet yuppie is what he is
  - REPETITIONS: I played, I played against um
  - FALSE STARTS: You're happy to – welcome to include it

  (Moore et al. 2015 *TSD*)
- Features of conversation: turn-taking, overlap, co-construction, *etc*

Speech versus Writing

- Fundamental difference: lack of sentence unit as used in writing; instead speech-units (SUs)

  (Moore et al. 2016 *COLING*)
- And disfluencies –
  - filled pauses: um he's a closet yuppie is what he is
  - repetitions: I played, I played against um
  - false starts: You're happy to – welcome to include it

  (Moore et al. 2015 *TSD*)
- Features of conversation: turn-taking, overlap, co-construction, *etc*

Speech versus Writing

- In this work we compare 4 English corpora from Universal Dependencies 2.0 and Penn Treebank 3
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE

Speech versus Writing

- In this work we compare 4 English corpora from Universal Dependencies 2.0 and Penn Treebank 3
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE

Speech versus Writing

- In this work we compare 4 English corpora from Universal Dependencies 2.0 and Penn Treebank 3
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE

Speech versus Writing

- In this work we compare 4 English corpora from Universal Dependencies 2.0 and Penn Treebank 3
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE

Speech versus Writing

- In this work we compare 4 English corpora from Universal Dependencies 2.0 and Penn Treebank 3
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE

Speech versus Writing

Medium	**Tokens**	**Types**
speech	394,611*	11,326**
writing	394,611	27,126

*sampled from 766,650 total
**mean of 100 samples (st.dev=45.5)

Speech versus Writing

Speech	Freq.	Rank	Writing	Freq.
I	46,382	1	the	41,423
and	33,080	2	to	26,459
the	29,870	3	and	22,977
you	27,142	4	I	20,048
that	27,038	5	a	18,289
it	26,600	6	of	18,112
to	22,666	7	in	14,490
a	22,513	8	is	10,020
uh	20,695	9	you	10,002
's	20,494	10	that	9952
of	17,112	11	for	8578
yeah	14,805	12	it	8238
know	14,723	13	was	8195
they	13,147	14	have	6604
in	12,548	15	on	5821

Speech versus Writing

Speech	**Freq.**	**Rank**	**Writing**	**Freq.**
you know	11,165	1	of the	4313
it's	8531	2	in the	3702
that's	6708	3	to the	2352
don't	5680	4	I have	1655
I do	4390	5	on the	1607
I think	4142	6	I am	1500
and I	3790	7	for the	1475
I'm	3716	8	I would	1427
I I	3000	9	and the	1389
in the	2972	10	and I	1361
and uh	2780	11	to be	1318
a lot	2714	12	I was	1140

Speech versus Writing

Speech	**Freq.**	**Rank**	**Writing**	**Freq.**
VBP_PRP	51,845	1	NN_DT	48,846
NN_DT	47,469	2	NN_IN	36,274
ROOT_UH	39,067	3	NN_NN	27,490
IN_NN	26,868	4	NN_JJ	21,566
VB_PRP	24,321	5	VB_NN	19,584
ROOT_VBP	24,156	6	VB_PRP	16,320

Parsing experiments

- Used Stanford CoreNLP toolkit to parse CoNLL format treebanks
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE
- We report unlabelled attachment scores (% tokens with correct heads)

Parsing experiments

- Used Stanford CoreNLP toolkit to parse CoNLL format treebanks
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE
- We report unlabelled attachment scores (% tokens with correct heads)

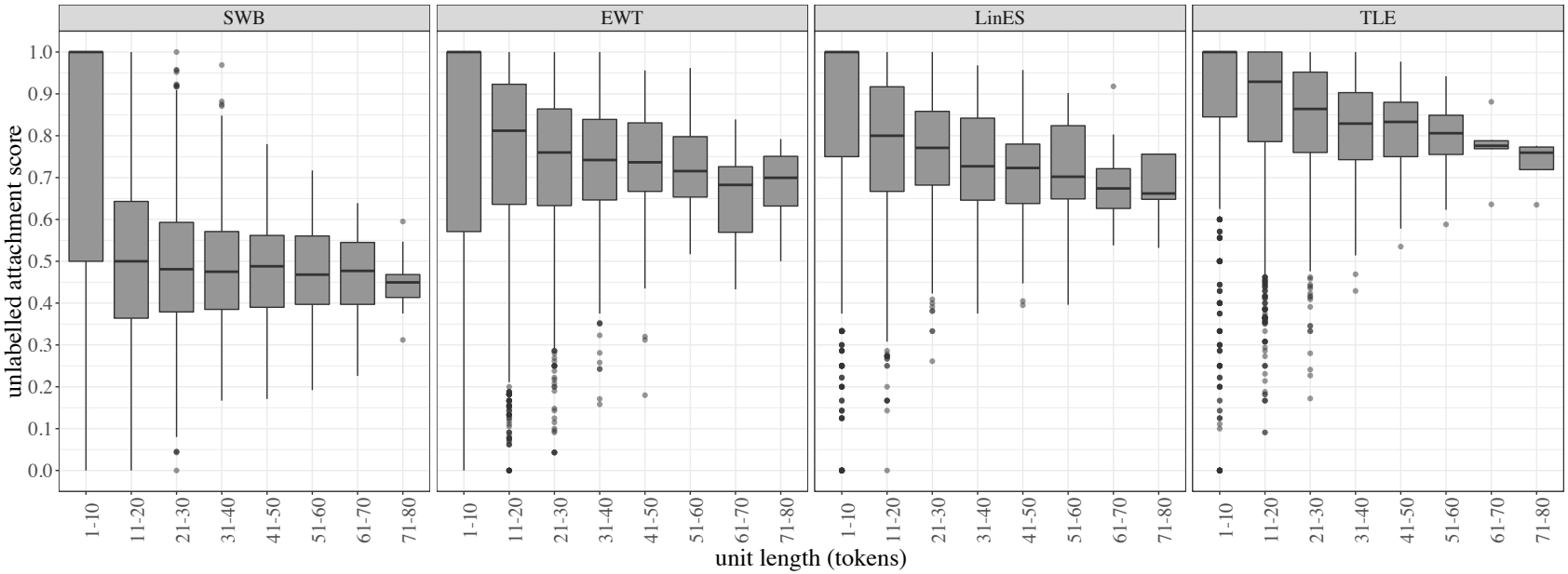
Parsing experiments

- Used Stanford CoreNLP toolkit to parse CoNLL format treebanks
  - PTB Switchboard Corpus of transcribed telephone conversations (SWB)
  - UD English Web Treebank (EWT)
  - UD English LinES (LinES), parallel corpus of English novels and Swedish translations
  - UD Treebank of Learner English (TLE), subset of CLC-FCE
- We report unlabelled attachment scores (% tokens with correct heads)

Parsing experiments

Corpus	**Medium**	**Units**	**Tokens**	**UAS**
SWB	speech	102,900	766,560	.540
EWT	writing	14,545	218,159	.744
LinES	writing	3650	64,188	.758
TLE	writing	5124	96,180	.845

Parsing experiments



Parsing experiments

- What if we train instead on the *Wall Street Journal* + Switchboard?
- We used Stanford Parser to train PCFGs with max.40 and 80 token SUs
- And make these models available (future baselines?)

Parsing experiments

- What if we train instead on the *Wall Street Journal* + Switchboard?
- We used Stanford Parser to train PCFGs with max.40 and 80 token SUs
- And make these models available (future baselines?)

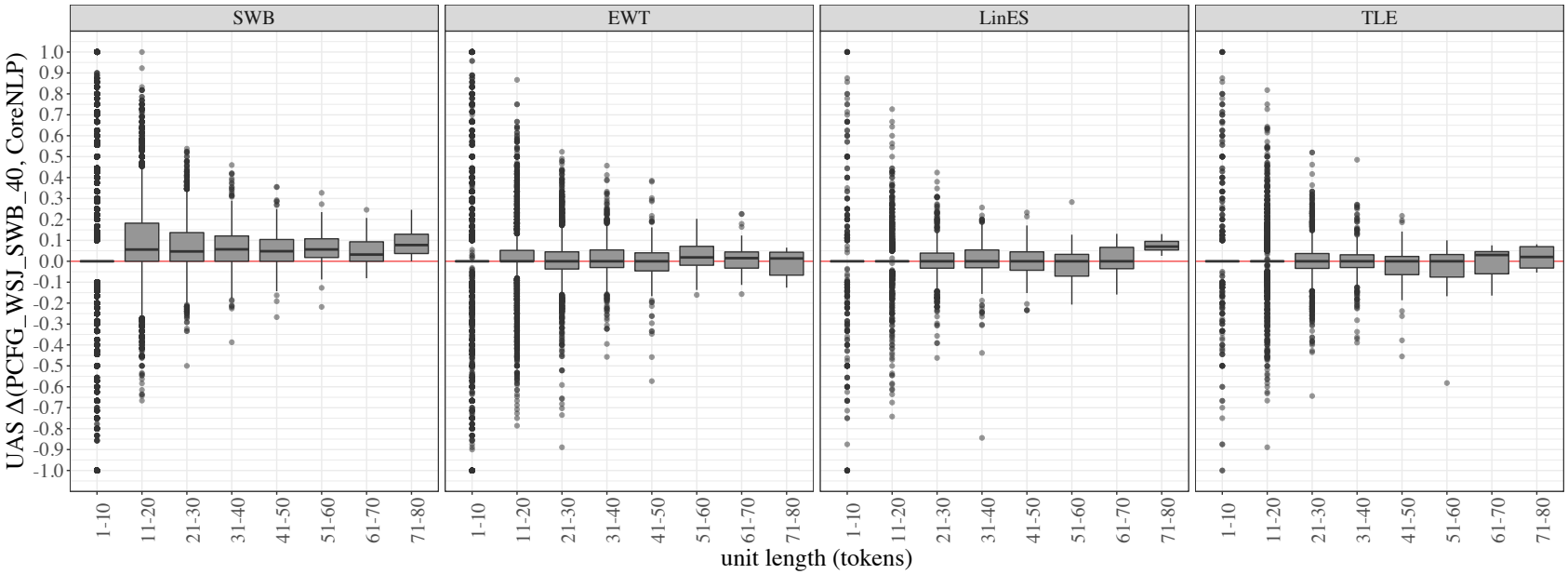
Parsing experiments

- What if we train instead on the *Wall Street Journal* + Switchboard?
- We used Stanford Parser to train PCFGs with max.40 and 80 token SUs
- And make these models available (future baselines?)

Parsing experiments

Model	**SWB**	**EWT**	**LinES**	**TLE**
CoreNLP	.540	.744	.758	.845
PCFG_WSJ_SWB_40	.624	.748	.760	.847
PCFG_WSJ_SWB_80	.624	.748	.760	.847

Parsing experiments



Conclusions

- Characterised speech vs writing differences
- Showed how unit length affects parsing of speech more than writing
- Demonstrated how much improvement can be made with a domain-appropriate parsing model
- +Speech parsing models available for other researchers: `https://goo.gl/iQMu9w`
- Call for more development of speech transcript treebanks.

Conclusions

- Characterised speech vs writing differences
- Showed how unit length affects parsing of speech more than writing
- Demonstrated how much improvement can be made with a domain-appropriate parsing model
- +Speech parsing models available for other researchers: `https://goo.gl/iQMu9w`
- Call for more development of speech transcript treebanks.

Conclusions

- Characterised speech vs writing differences
- Showed how unit length affects parsing of speech more than writing
- Demonstrated how much improvement can be made with a domain-appropriate parsing model
- +Speech parsing models available for other researchers: `https://goo.gl/iQMu9w`
- Call for more development of speech transcript treebanks.

Conclusions

- Characterised speech vs writing differences
- Showed how unit length affects parsing of speech more than writing
- Demonstrated how much improvement can be made with a domain-appropriate parsing model
- +Speech parsing models available for other researchers: `https://goo.gl/iQMu9w`
- Call for more development of speech transcript treebanks.

Conclusions

- Characterised speech vs writing differences
- Showed how unit length affects parsing of speech more than writing
- Demonstrated how much improvement can be made with a domain-appropriate parsing model
- +Speech parsing models available for other researchers: `https://goo.gl/iQMu9w`
- Call for more development of speech transcript treebanks.

Future work

- Analyse SUs with low UAS: what are the causes?
- Redefine grammar and grammaticality?
- Extra pre-processing: e.g. semantic chunking (Muszynska 2016 *ACL*)
- Or joint SU delimitation, disfluency detection, parsing (e.g. Honnibal & Johnson 2014 *TACL*; Yoshikawa et al 2016 *EMNLP*)
- Other metrics: e.g. SParseval (Roark et al 2006 *LREC*)

Future work

- Analyse SUs with low UAS: what are the causes?
- Redefine grammar and grammaticality?
- Extra pre-processing: e.g. semantic chunking (Muszynska 2016 *ACL*)
- Or joint SU delimitation, disfluency detection, parsing (e.g. Honnibal & Johnson 2014 *TACL*; Yoshikawa et al 2016 *EMNLP*)
- Other metrics: e.g. SParseval (Roark et al 2006 *LREC*)

Future work

- Analyse SUs with low UAS: what are the causes?
- Redefine grammar and grammaticality?
- Extra pre-processing: e.g. semantic chunking (Muszynska 2016 *ACL*)
- Or joint SU delimitation, disfluency detection, parsing (e.g. Honnibal & Johnson 2014 *TACL*; Yoshikawa et al 2016 *EMNLP*)
- Other metrics: e.g. SParseval (Roark et al 2006 *LREC*)

Future work

- Analyse SUs with low UAS: what are the causes?
- Redefine grammar and grammaticality?
- Extra pre-processing: e.g. semantic chunking (Muszynska 2016 *ACL*)
- Or joint SU delimitation, disfluency detection, parsing (e.g. Honnibal & Johnson 2014 *TACL*; Yoshikawa et al 2016 *EMNLP*)
- Other metrics: e.g. SParseval (Roark et al 2006 *LREC*)

Future work

- Analyse SUs with low UAS: what are the causes?
- Redefine grammar and grammaticality?
- Extra pre-processing: e.g. semantic chunking (Muszynska 2016 *ACL*)
- Or joint SU delimitation, disfluency detection, parsing (e.g. Honnibal & Johnson 2014 *TACL*; Yoshikawa et al 2016 *EMNLP*)
- Other metrics: e.g. SParseval (Roark et al 2006 *LREC*)

The End

- Acknowledgements:
  - Cambridge English Language Assessment
  - Sebastian Schuster & Chris Manning re UD corpora
  - SCNLP organisers
  - 3 anonymous reviewers

- Thank you! `apc38@cam.ac.uk`