UNIVERSITY OF CAMBRIDGE

Automatic Grammatical Error Detection of Non-native Spoken Learner English

Kate Knill[1], Mark Gales[1], Potsawee Manakul[1], Andrew Caines[2]

[1]Engineering Department / ALTA Institute

[2]Computer Science and Technology / ALTA Institute

17 May 2019

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Challenge and Advantages of Spoken Language

- Spoken language consists of

Text + Pronunciation + Prosody + Delivery

- Challenge for feedback on “grammatical” errors in spoken language

Spoken Text ≠ Written Text

  - We don’t speak in sentences, we repeat ourselves, hesitate, mumble etc
  - There is no defined spoken grammar standard

- Advantages of speech
  - There are no spelling or punctuation mistakes
  - We provide additional information within the audio signal

Example of Learner Speech

MANUAL TRANSCRIPTION

flor company is an engineering compa- is is is eng- engineering company %hes% %hes% in the in the poland %hes% we do business the ref- refinery business and the chemical business %hes% the job we can offer is a engineering job %hes% basically this is the job in the office

META-DATA EXTRACTION

// flor company is an [FS engineering {P compa-} [REP is is + is] {P eng-} + engineering company] {F %hes%} {F %hes%} [REP in the + in the} poland // {F %hes%} we do business the {P ref-} refinery business and the chemical business {F %hes%} // the job we can offer is a engineering job // {F %hes%} basically this is the job in the office //

GRAMMATICAL ERRORS

// flor company is an engineering company in the poland
// we [RV do] [RN business] the refinery business and the chemical business
// the job we can offer is [FD a] engineering job
// basically this is [RD the] job in the office

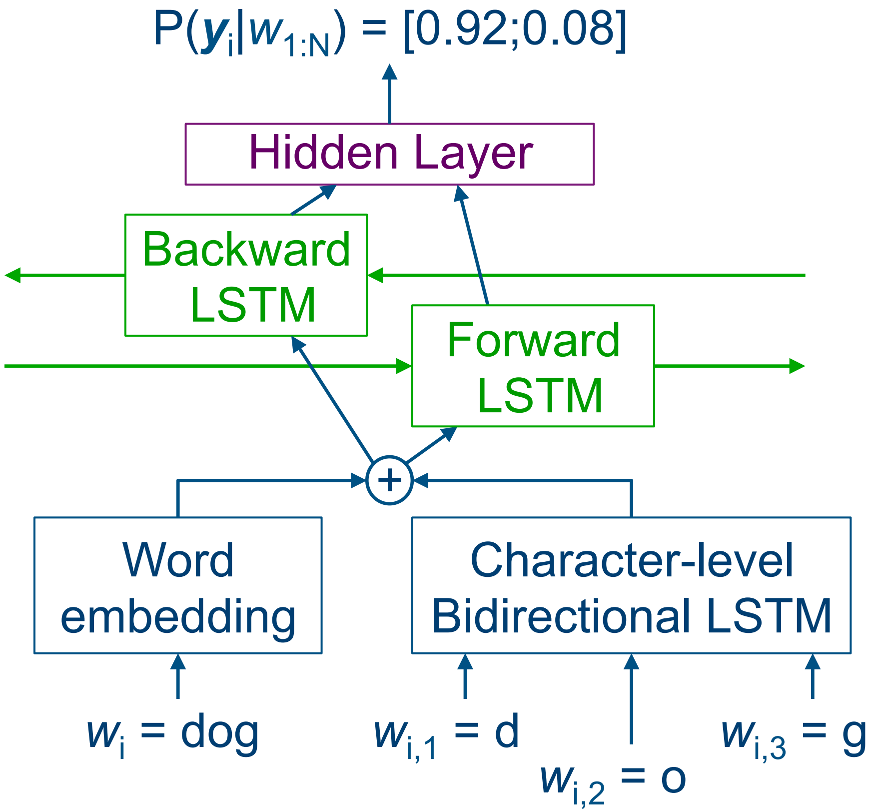
Grammatical Error Detection

- Task: given a sentence automatically label each word with
  - P_{word} (grammar is correct) and P_{word} (grammar is incorrect)
- Example sentence

	Internet	was	something	amazing	for	me	.
P(c)	0.02	0.96	0.97	0.97	0.95	0.98	0.99
P(i)	0.98	0.04	0.03	0.03	0.05	0.02	0.01

- Predict prob. distribution $\mathbf{y}_i$ for each token $\mathbf{w}_{1:N} = \{w_1, \cdots, w_N\}$

Sequence Labeller



marekrei

Corpora

- Non-native English learners with grammatical error annotation
  - BULATS: free speech with up to 1 minute per response
  - NICT-JLE: oral proficiency test interviews
  - CLC: range of written exams at different grade levels

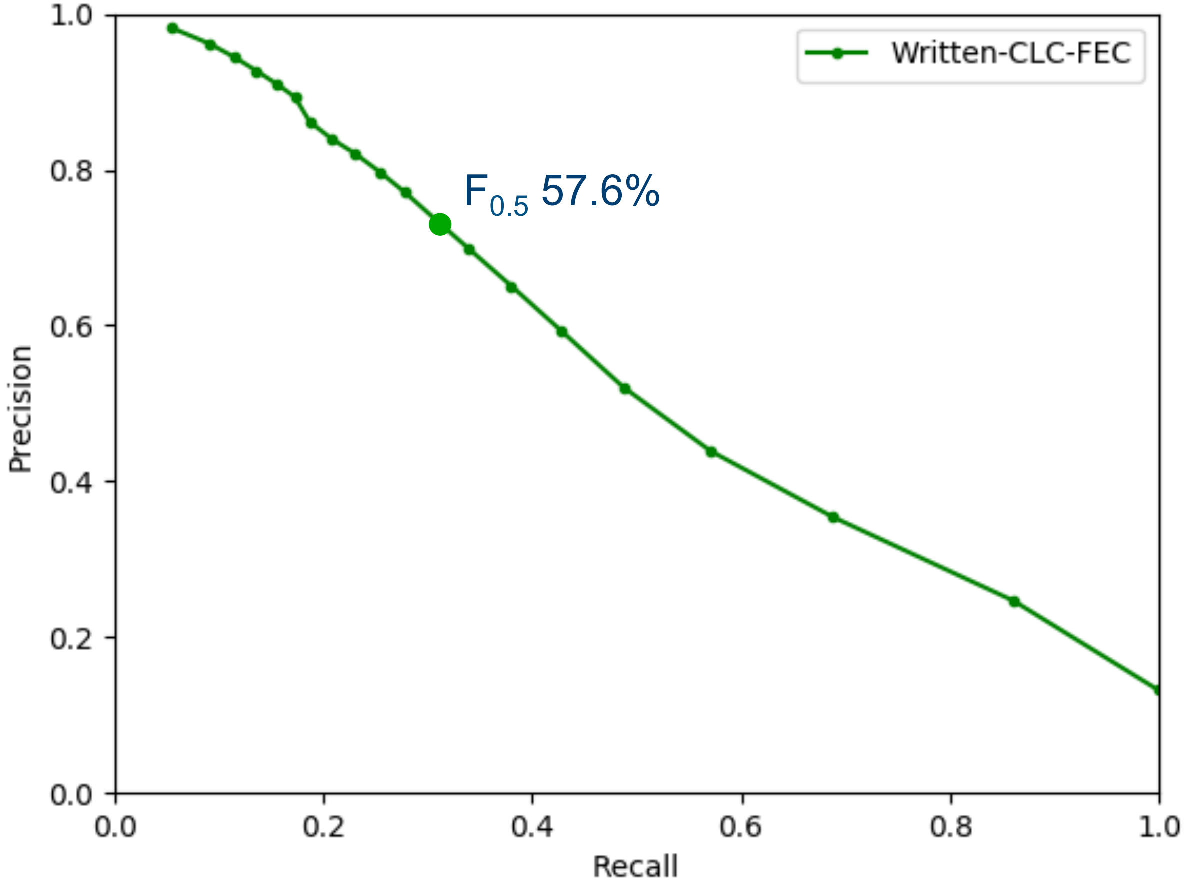
Corpus	Spoken/ Written	# Wds	# Uniq Wds	Audio	L1s	Grades
BULATS	Spoken	61.9K	3.4K	Yes	6	A1-C2
NICT--JLE	Spoken	135.3K	5.6K	No	1	A1-B2
CLC	Written	14.1M	79.1K	No	Many	A1-C2

Data Processing for Spoken GED

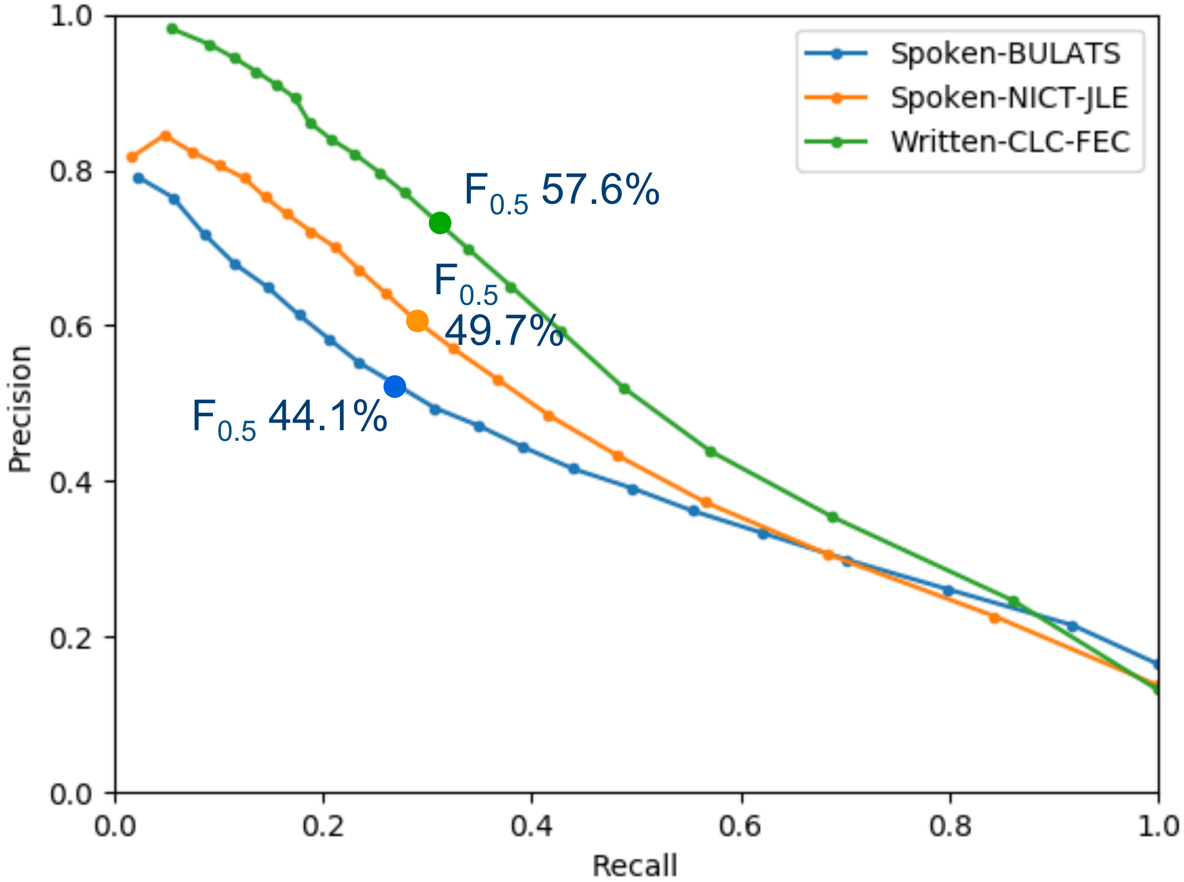
- Match data processing in training and testing
  a. Train: Text data - correct spelling errors and remove punctuation and casing
  b. Test: Speech data - convert speech transcriptions to be “like” text

// flor company is an engineering company in the poland
// we do business the refinery business and the chemical business
// the job we can offer is a engineering job
// basically this is the job in the office

GED Using CLC Trained Model



GED Using CLC Trained Model



(Small Scale) System Error Analysis

- True precision higher for Spoken BULATS than scores suggest
  - System error (~27%)

    .. and i have to practice more because I have ..

  - Unmarked error (~40%)

    .. so I think you need taxi

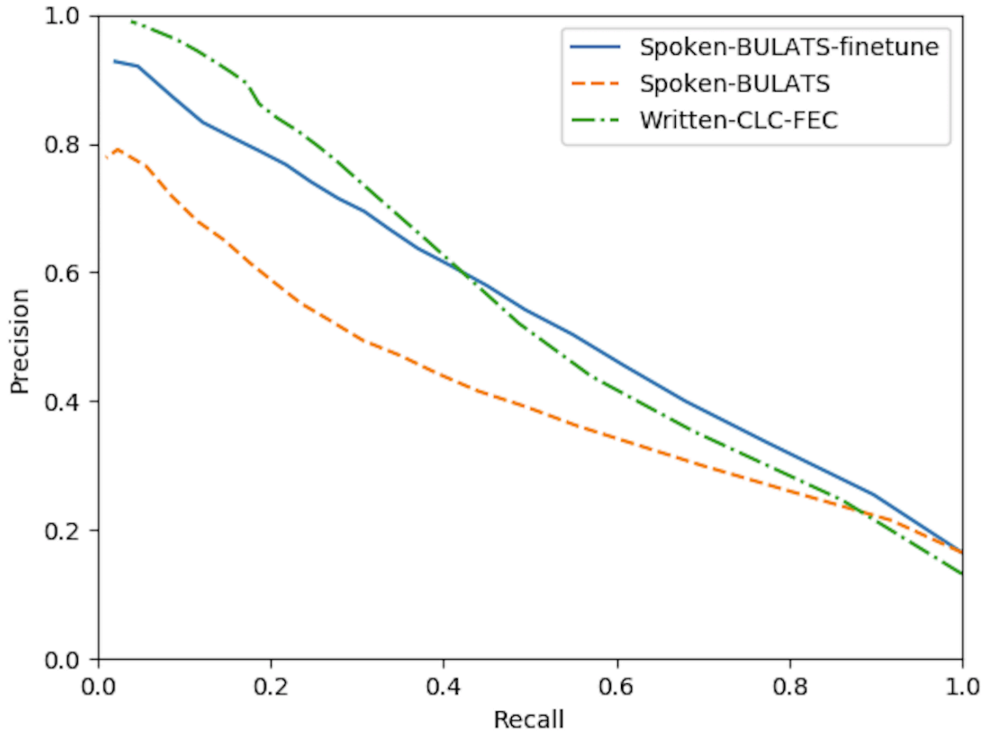
  - Next to error tagged word(s) (~27%)

    .. and continue to inform with customer when we have ..

- To provide feedback we need to boost recall of high precision items
  - Issue: lack of labelled learner speech corpora
    - Adapt/“fine-tune” CLC trained system to subset of target speech data

Boosting GED Performance on Spoken BULATS

- Fine-tune CLC system with 80% data, dev 10%, test 10% x10



- Fine-tuning produces significant boost in performance
  - has also learnt some annotator bias e.g. “two thousand eight”

Example of Learner Speech: ASR Transcription

MANUAL GRAMMATICAL ERRORS

flor company is an engineering compa- is is is eng- engineering company %hes% %hes% in the in the poland %hes% we do business the ref- refinery business and the chemical business %hes% the job we can offer is a engineering job %hes% basically this is the job in the office

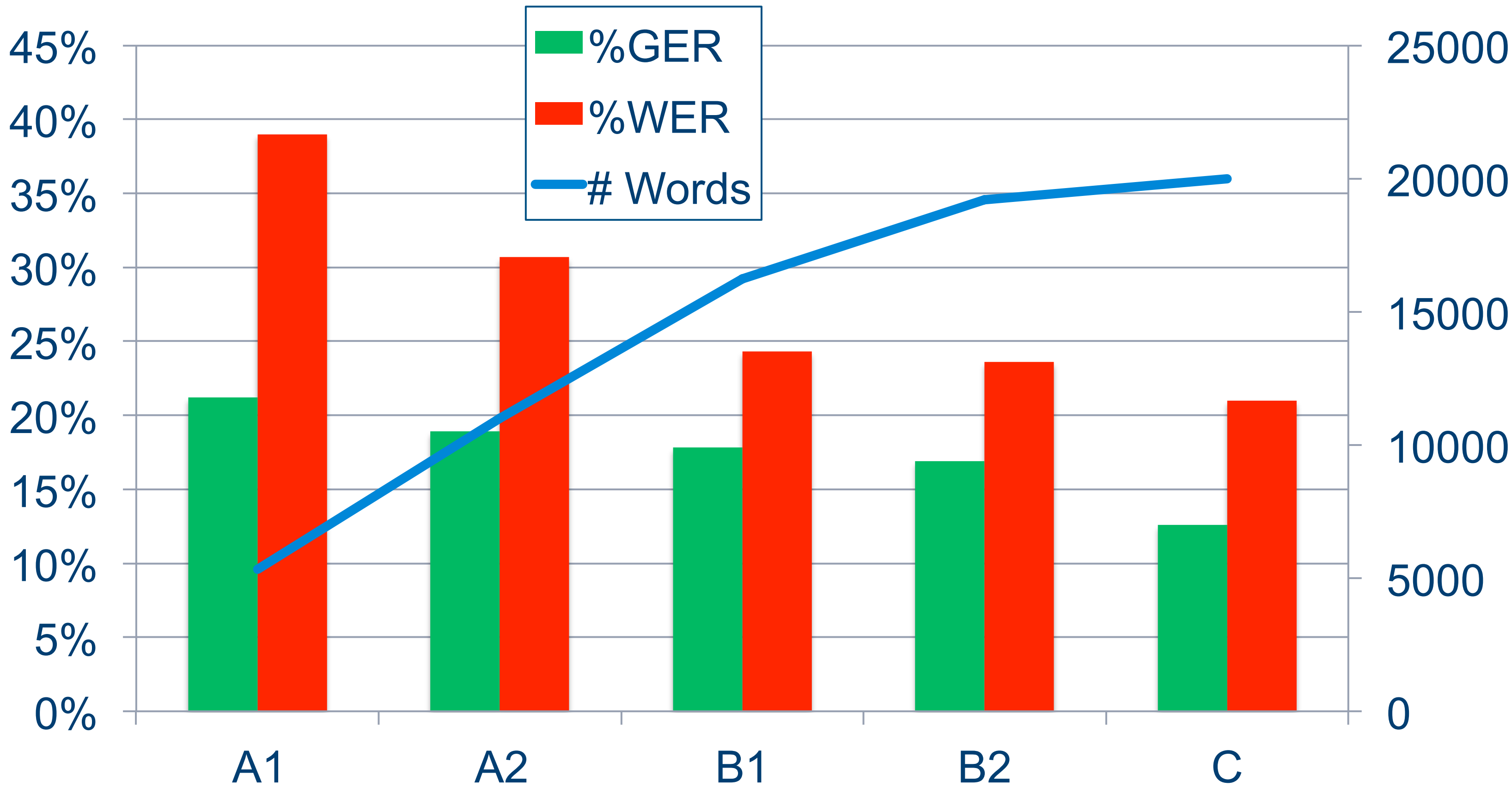
ASR TRANSCRIPTION ERRORS

flower companies in joining company is is is engineer engineering company in the in the poll one %hes% we do business there are refinery business in a chemical business %hes% the job we can offer is [del] engineering job %hes% basically this is the job in the office

ASR "GRAMMATICAL ERRORS" – feedback focused

// flower companies in engineering company in the poll one
// we do business there are refinery business in a chemical business
// the job we can offer is engineering job
// basically this is the job in the office

BULATS ASR Annotation Error Rates



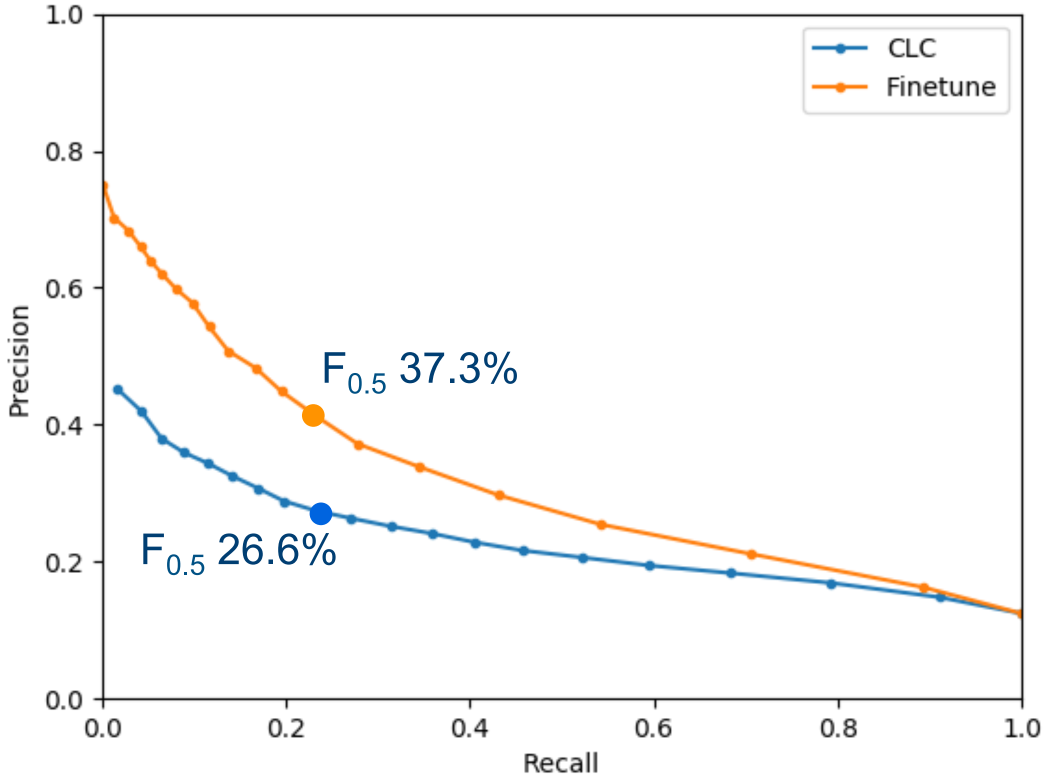
- Overall: 71751 words   16.5% GER   25.2% WER

BULATS ASR Word Error Rate

	#	%WER
Overall	71751	25.2
“Fluent”	52698	19.3
Grammatical Error	10348	29.1
Disfluency	2524	36.4

GED on BULATS ASR Transcriptions

- Manual transcriptions used for GE marking and meta-data extraction



- Significantly lower performance than manual transcriptions

Conclusions

- Detecting “grammatical” errors in learner speech is hard!
  - As is annotating the errors
- Focus on high precision region for feedback
  - Testing if regions where errors detected are sufficient to provide useful help
- More research required into:
  - Meta-data extraction
  - Boosting training data by mimicing learner speech errors
  - Detecting portions of ASR transcription the system is confident in

Questions?

Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.