Applying Deep Learning in Non-native Spoken English Assessment

Kate Knill

APSIPA 21 November 2019

Automated Language Teaching & Assessment Institute

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

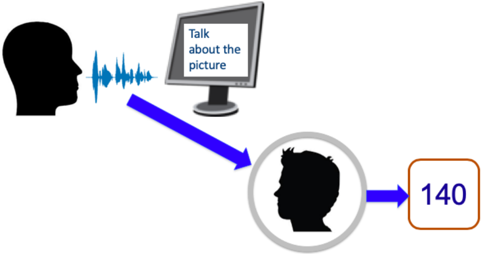
UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

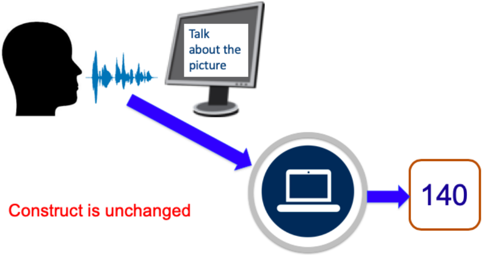
- Virtual Institute for

  cutting-edge research on non-native English assessment
  - Machine Learning and Natural Language Processing
  - Develop technology to enhance assessment and learning
  - Look to benefit learners and teachers worldwide

Spoken Language Assessment & Learning

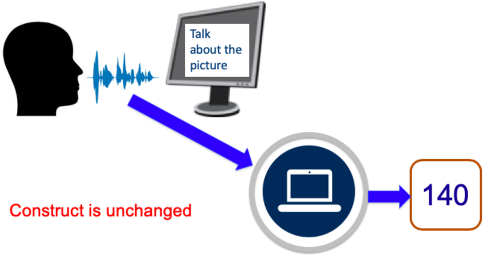


Spoken Language Assessment & Learning



Construct is unchanged

Spoken Language Assessment & Learning



- Automate (English) spoken language assessment & learning
  - without simplifying/limiting form of test: "free speaking"
  - possibility for richer, interactive, tests
  - desire to assess communication skills

CEFR - Levels of Foreign Language (L2) Learning

- Internationally agreed standard for assessing level
  - Common European Framework of Reference (CEFR)

- Basic User

  A1 - breakthrough or beginner
  A2 - way-stage or elementary
- Independent User

  B1 - threshold or intermediate
  B2 - vantage or upper intermediate
- Proficient User

  C1 - effective operational proficiency or advanced
  C2 - mastery or proficiency

Spoken BULATS (Linguaskill Business)

- Business Language Testing Service (BULATS) test
  - includes: Reading and Listening, Speaking and Writing tests
  - low-stakes test - Spoken test recorded and assessed off-line
- Example of a test of communication skills:

A Introductory Questions: your name, where you are from
B Read Aloud: read specific sentences
C Topic Discussion: discuss a company that you admire

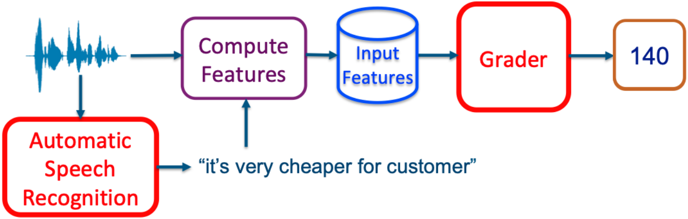
D Interpret and Discuss Chart/Slide: example above
E Answer Topic Questions: 5 questions on organising a meeting

Overview

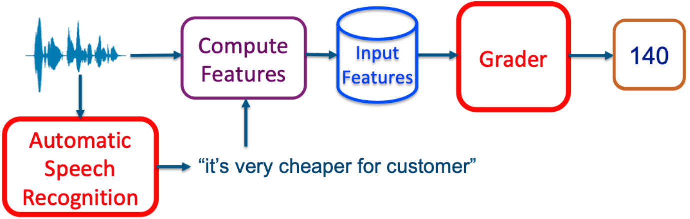
- Assessment: spoken language assessment framework
  - non-native speech recognition
  - features for assessment
  - form of classifier and uncertainty
- Feedback to candidate: integrate assessment and learning
  - spoken “grammatical error” detection/correction
- Malpractice: detecting attempts to “game” the system
  - off-topic response detection

Assessment

Assessment Framework [18]



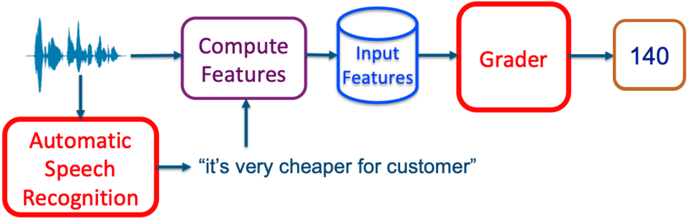
Assessment Framework [18]



Key Challenges:

- Input speech variability
  - Speakers: large range of L1s, non-native speech, wide ability
  - Recordings: varying background noises, channel corruptions

Assessment Framework [18]

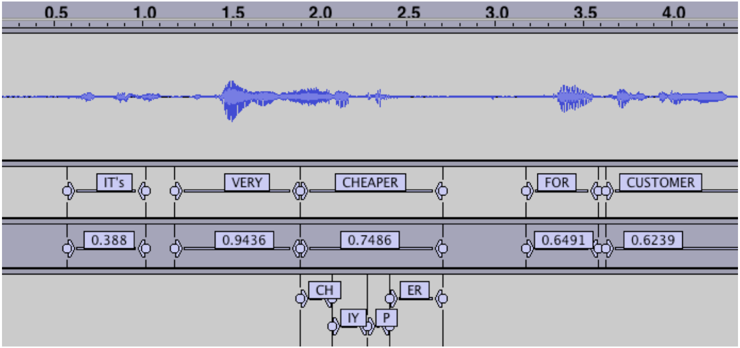


Key Challenges:

- Input speech variability
  - Speakers: large range of L1s, non-native speech, wide ability
  - Recordings: varying background noises, channel corruptions

    ⇒ High word error rate (WER): propagates through system

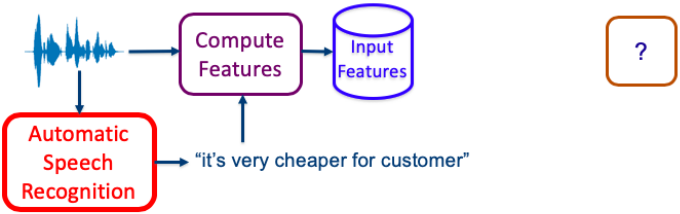
Automatic Speech Recognition [17, 2]



- Baseline Automatic Speech Recognition (ASR) yields:
  - time aligned word/disfluencies/partial-word sequence
  - time aligned phone/grapheme sequence
  - word level confidence scores

ASR System [17, 2]

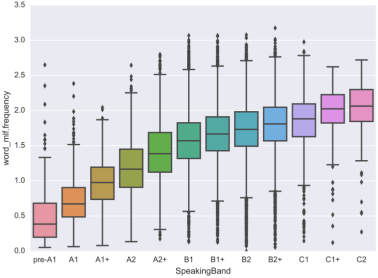
- Deep-learning based ASR systems used:
  - Kaldi-based lattice-free MMI acoustic models
  - ensemble combination uses sequence teacher-student training
  - rescoring with RNNLM and su-RNNLM based language models

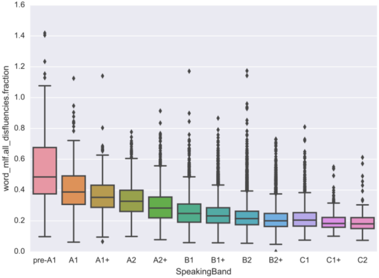

Compute
Features
Input
Features
?
Automatic
Speech
Recognition
"it's very cheaper for customer"

Baseline Grader Features [18]

- Baseline features mainly fluency based, including:
- Audio Features: statistics about
  - fundamental frequency (F0)
  - speech energy and duration
- Aligned Text Features: statistics about
  - silence durations
  - number of disfluencies (um, uh etc)
  - speaking rate
- Text identity features
  - number of repeated words (per word)
  - number of unique word identities

Baseline Features: Correlation with Grades

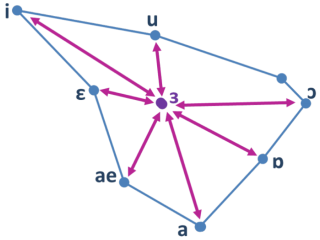




- Examine distribution of extracted features with grade
  - example box-plots for speaking rate and percentage disfluencies

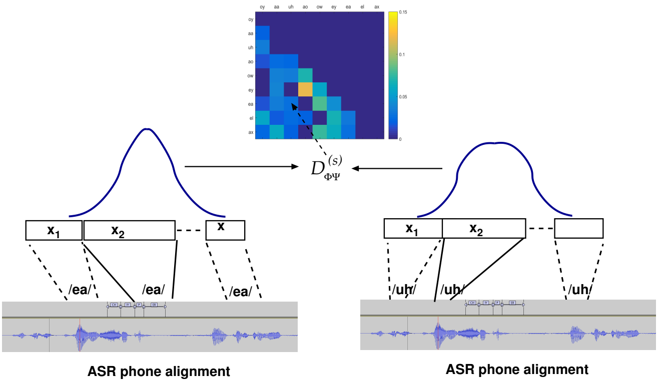
Derived Features: Phone-Distances [13]

- Pronunciation is an important predictor of proficiency
  - but no reference native speech for free speaking tasks
- Phone distance features are one approach



  - each phone characterised relative to others
  - independent of speaker attributes
  - characterise speaker's pronunciation of each phone

Model-based Pronunciation Features [6]

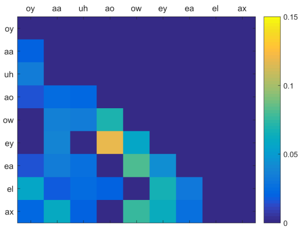


- Train Gaussian model for each phone $\boldsymbol{x}^{(i)}$ and speaker s:

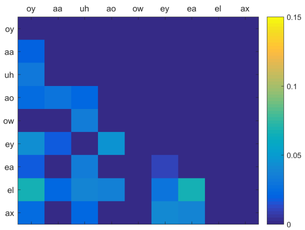
$$p(\boldsymbol{x}^{(i)}|\omega_{\phi}) = \mathcal{N}(\boldsymbol{x}^{(i)}; \boldsymbol{\mu}_{\phi}{}^{(s)}, \boldsymbol{\Sigma}_{\phi}{}^{(s)})$$

- Compute relative entropy between each phone-pair $\mathcal{D}_{\phi,\psi}{}^{(s)}$

Model-based Pronunciation Features



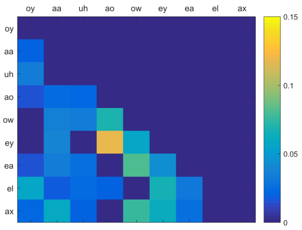
Candidate Grade A1



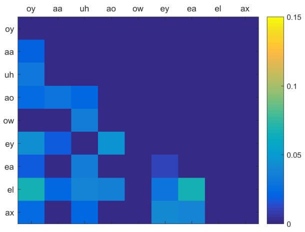
Candidate Grade C1

- Pair-wise entropies used as features in grader
  - yields small gains in assessment performance
  - pattern is first language (L1) dependent

Model-based Pronunciation Features



Candidate Grade A1

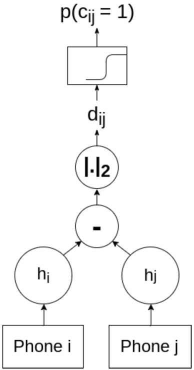


Candidate Grade C1

- Pair-wise entropies used as features in grader
  - yields small gains in assessment performance
  - pattern is first language (L1) dependent
- General approach ⇒ tunable approach based on deep learning

Siamese Network-based Phone Distances [3]

- Siamese networks map features to a meaningful distance space



- Train distances for classification

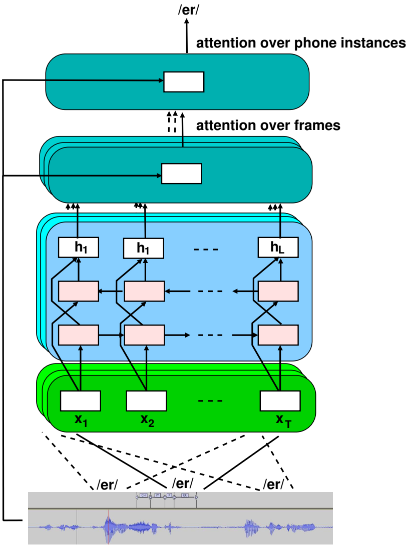
$$y = \mathcal{F}\left(\left\|\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{x}_j;\boldsymbol{\theta})\right\|\right)$$

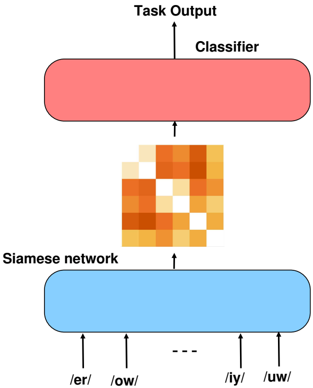
  - maps features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to new space
  - parameters of mapping network the same $\boldsymbol{\theta}$

- Easy to define training targets
  - $y = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ different classes
  - $y = 0$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ same class

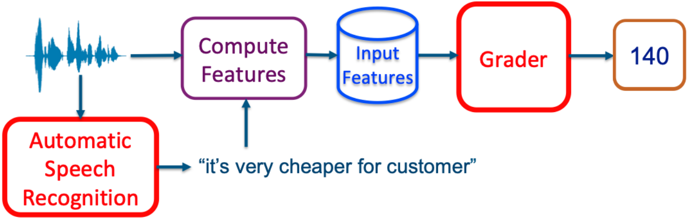
- For phone-distance system
  - can use KL-divergence targets

Deep Learning Pronunciation Features [7]





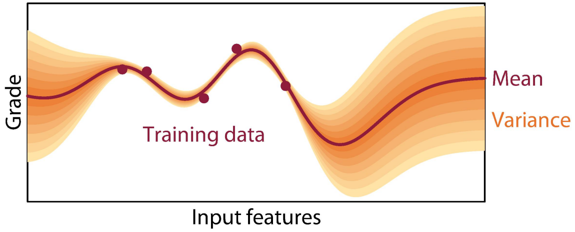
Grader



- Supervision data assessment is a score (0-6)
  - assessment run as a regression task: $p(y|\boldsymbol{x}^{\star};\boldsymbol{\theta})$

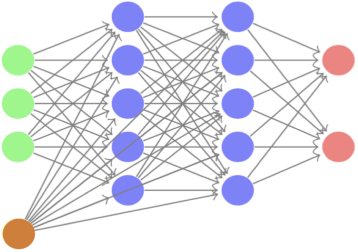
Assessment: Gaussian Process [14, 16]

- Gaussian process
  - non-parametric model based on joint-Gaussian assumption



  - GP mean is used as the score prediction
  - GP variance is a standard aspect of the model
    - gives measure of confidence in assessment

Deep Learning: Deep Density Networks [1, 9]

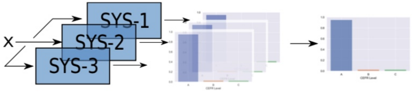


- Deep Density Networks predict parameters of a distribution

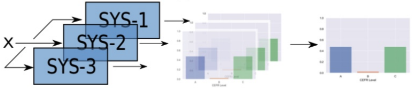
$$p(y|\boldsymbol{x}^\star;\boldsymbol{\theta}) = \mathcal{N}(y; f_\mu(\boldsymbol{x}^\star;\boldsymbol{\theta}), f_\sigma(\boldsymbol{x}^\star;\boldsymbol{\theta}))$$

  - flexible framework for any form of distribution
  - distribution variance gives measure of confidence in assessment

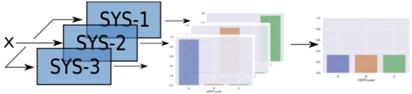
Grader Uncertainty: Ensembles of DDNs [10]



(a) Confident



(b) Uncertain on decision boundary



(c) Uncertain far from training data

- Generate distribution over distributions
  - Ensemble diversity yields more reliable uncertainty estimates
  - Sources of uncertainty can be split ⇒ better decision making

Assessment System Performance

- Accurately annotated corpus for system development
  - 220 speakers over 6 L1 languages (3 Asian, 3 European)
  - accurate manual transcriptions, ASR evaluation (WER%)
  - expert (CA) CEFR grading, grader evaluation

Assessment System Performance

- Accurately annotated corpus for system development
  - 220 speakers over 6 L1 languages (3 Asian, 3 European)
  - accurate manual transcriptions, ASR evaluation (WER%)
  - expert (CA) CEFR grading, grader evaluation

- Non-Native ASR: real-time decoding (non-RNNLM)

	A1	A2	B1	B2	C	Avg
Baseline ASR	33.8	27.7	21.2	19.9	16.5	21.3
+RNNLM	31.8	25.4	19.6	18.0	14.7	19.5

  - "basic users" (A1/A2) highly challenging data

Assessment System Performance

- Accurately annotated corpus for system development
  - 220 speakers over 6 L1 languages (3 Asian, 3 European)
  - accurate manual transcriptions, ASR evaluation (WER%)
  - expert (CA) CEFR grading, grader evaluation

- Non-Native ASR: real-time decoding (non-RNNLM)

	A1	A2	B1	B2	C	Avg
Baseline ASR	33.8	27.7	21.2	19.9	16.5	21.3
+RNNLM	31.8	25.4	19.6	18.0	14.7	19.5

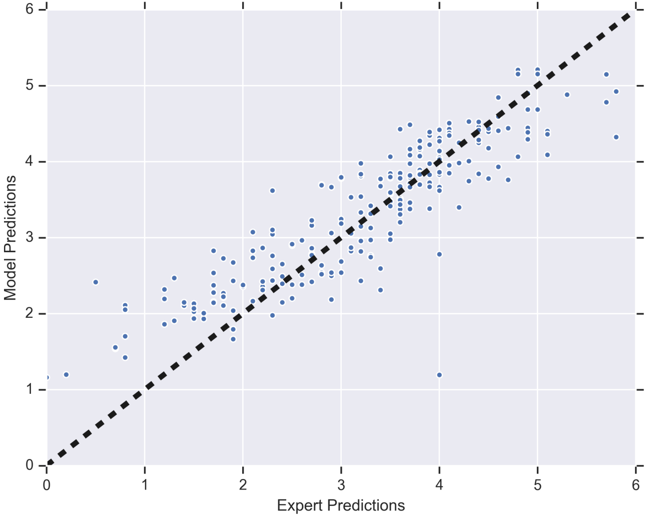
  - "basic users" (A1/A2) highly challenging data

- Assessment: using complete test

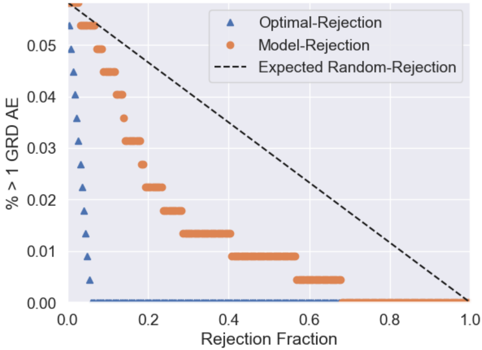
PCC	MSE	%$\leq$ 0.5	%$\leq$ 1.0
0.888	0.31	68.2	94.2

  - $\leq$ 1.0 indicates within one CEFR grade-level

Performance Analysis



Incorporating Assessment Uncertainty



- Use uncertainty measures to detect “high” error predcitions
  - these can be tagged for manual checking

Speak and Improve: https:speakandimprove.com

- Current beta of free speaking web-application
  - collaboration between ALTA, Cambridge Assessment and Industrial partners

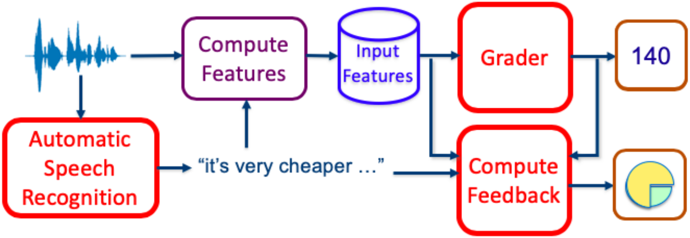
Feedback:

Spoken Learner 'Grammatical' Errors

Candidate Feedback

- Feedback to the candidate is important for language learning
  - many aspects of spoken language contribute to overall grade
  - performance on each aspect varies between candidates
- Message Realisation (Fluency):
  - is the pronunciation correct?
  - is the correct intonation pattern used?
  - is the speech delivered in a coherent fashion?
- Message Construction:
  - is the response relevant to the prompt?
  - is the message grammatically correct (in speech context)?
  - is the message using the appropriate vocabulary?

Feedback Framework



- Key Challenges:
  - speaker and speech variability
    - wide range of abilities, L1-specific errors
  - requires high precision but WER is high
    - don't want to give feedback on system errors
  - lack of annotated data

Grammatical Error Detection and Correction

Learner	she	say	me	what	i	should	do	it	...
GED	c	i	c	i	c	c	c	c	...
GEC	she	told	me	how	i	should	do	it	...

- Grammatical Error Detection (GED)
  - standard sequence labelling problem
- Grammatical Error Correction (GEC)
  - standard sequence-to-sequence translation problem
  - no unique solution

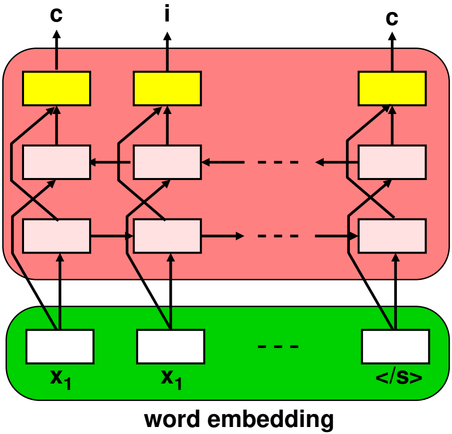
Grammatical Error Detection and Correction

Learner	she	say	me	what	i	should	do	it	...
GED	c	i	c	i	c	c	c	c	...
GEC	she	told	me	how	i	should	do	it	...

- Grammatical Error Detection (GED)
  - standard sequence labelling problem
- Grammatical Error Correction (GEC)
  - standard sequence-to-sequence translation problem
  - no unique solution
- Lots of data for training GED/GEC systems for writing

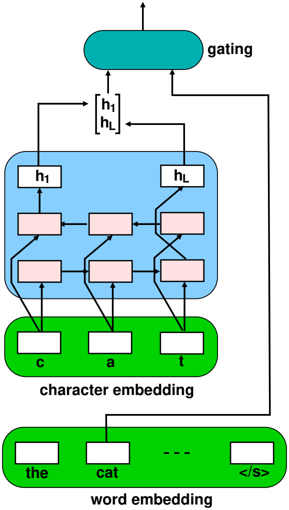
  $\Rightarrow$ fine-tune writing models to speech data

Grammatical Error Detection (GED)



- Predict whether word is correct (`c`) or incorrect (`i`)
  - initial word embedding followed by classifier

Handling Rare/Missing Words [15]



Spoken "Grammatical Error" Detection [5, 8]

- Problem for speech: no agreed grammar
  - native speakers use non-grammatical constructs
  - native speakers hesitate, repeat, false start etc
- Redefine task as

  ⇒ "feedback that is useful for spoken message construction"

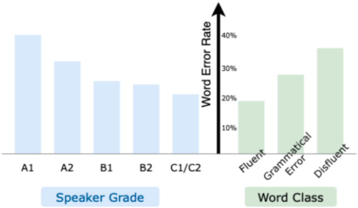
Spoken "Grammatical Error" Detection [5, 8]

- Problem for speech: no agreed grammar
  - native speakers use non-grammatical constructs
  - native speakers hesitate, repeat, false start etc
- Redefine task as

  ⇒ "feedback that is useful for spoken message construction"
- Some overlap with written GEC and GED, but not the same

Modified Spoken GED Criterion [5, 8]

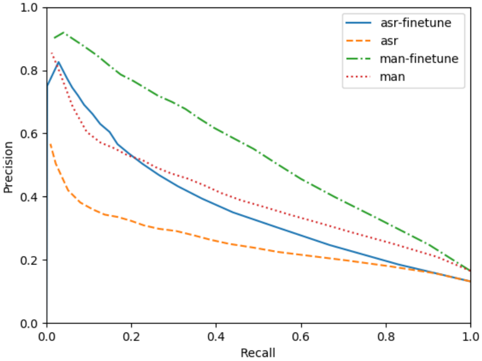
- Have to take impact of ASR into account



Learner	she	say	me	what	i	should	do	it	...
ASR	she	may	me	what	i	should	do	it	...
GED	c	i	c	i	c	c	c	c	...
$\mathrm{GED_f}$	c	c	c	i	c	c	c	c	...

- Modified GED criterion ($\mathrm{GED_f}$) - more challenging

BULATS GED Performance [8]



- Significant drop from manual (MAN) to ASR transcriptions
  - even after fine-tuning to limited spoken language data
- Can use ASR confidence to select high precision GED:
  - useful information for feedback eg > 90% missed determiners

Malpractice:

Off-Topic Response Detection

Relevance Detection

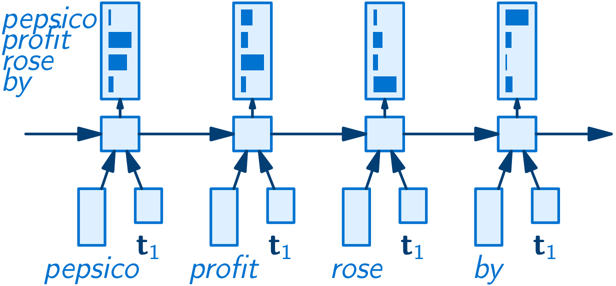
- Off-topic response (relevance) takes:
  - $\boldsymbol{w}^p$: prompt (question) from script
    $$\boldsymbol{w}^p = \{\texttt{Discuss a company that you admire}\}$$
  - $\boldsymbol{w}^r$: response from candidate derived from speech recognition
    $$\boldsymbol{w}^r = \{\texttt{Cambridge Assessment is wonderful, it ...}\}$$

  and derives probability of relevance

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$$

- Two standard options for model:
  - Generative Model of Responses
  - Discriminative Model of Relevance

Generative Model of Responses



- Prompt topic-adapted RNN Language Model
- Probability of relevance derived from:

$$\mathrm{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p) \approx \mathrm{P}(\boldsymbol{w}^p|\boldsymbol{w}^r) \approx \mathrm{P}(\mathbf{t}_p|\boldsymbol{w}^r) = \frac{\mathrm{P}(\boldsymbol{w}^r|\mathbf{t}_p)\mathrm{P}(\mathbf{t}_p)}{\sum_i \mathrm{P}(\boldsymbol{w}^r|\mathbf{t}_i)\mathrm{P}(\mathbf{t}_i)}$$

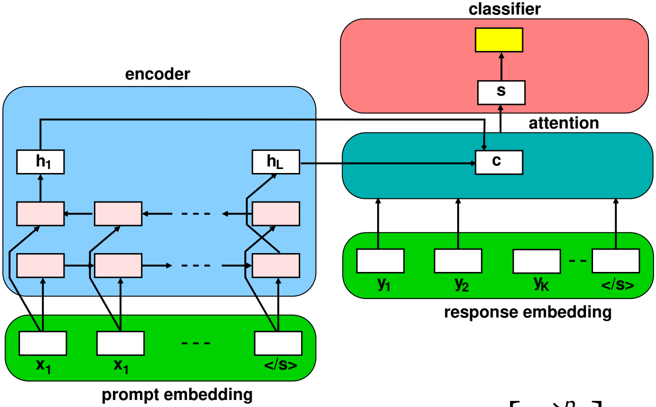
Discriminative Response Model [12]

- Directly model the probability of relevance

$$\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p)$$

- Split the process into sequence of steps:
  1. $\boldsymbol{w}^p \rightarrow \tilde{\boldsymbol{h}}^p$: prompt embedding
  2. $\boldsymbol{w}^r|\tilde{\boldsymbol{h}}^p \rightarrow \boldsymbol{c}^r$: response encoding (given prompt encoding)
  3. $\mathtt{P}(\mathtt{rel}|\boldsymbol{w}^r, \boldsymbol{w}^p) = \mathtt{P}(\mathtt{rel}|\boldsymbol{c}^r) = f(\boldsymbol{c}^r)$: probability of relevance

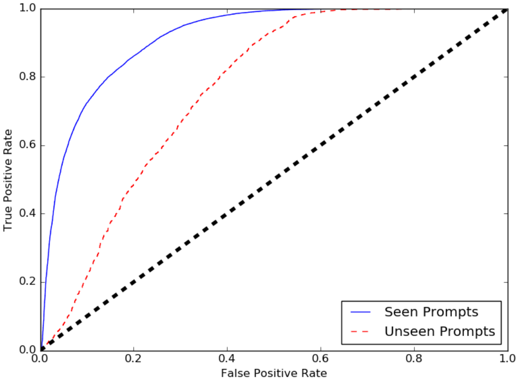
Attention-Based Model



$$\boldsymbol{c} = \sum_{\tau=1}^{L} \alpha_\tau \boldsymbol{h}_\tau^r; \qquad \alpha_\tau = f(\tilde{\boldsymbol{h}}^p, \boldsymbol{h}_\tau^r); \qquad \tilde{\boldsymbol{h}}^p = \begin{bmatrix} \overrightarrow{\boldsymbol{h}}_L^p \\ \overleftarrow{\boldsymbol{h}}_1^p \end{bmatrix}$$

- The prompt embedding can be applied to any prompt
  - naturally handles unseen (in training data) prompts

Results: Seen & Unseen Prompts ROC Curves



- ROC curve for performance with Seen and Unseen prompts
  - against balanced set of seen/unseen prompt responses

Conclusions

Conclusions

- Spoken language learning and assessment important
  - increasing need for automated (and validated) systems
- Deep learning is central to current state-of-the-art systems
  - all assessment and feedback stages make use of approaches
- The lack of annotated data is a big challenge
  - very hard to annotate (and agree) spoken learner data

Thanks

- Thanks to Cambridge Assessment, University of Cambridge, for supporting this research
- Thanks to the CUED ALTA Speech Team for their contributions: Prof. Mark Gales, Rogier van Dalen, Kostas Kyriakopoulos, Yiting Lu, Andrey Malinin, Potsawee Manakul, Anton Ragni, Linlin Wang, Yu Wang
- http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[2] X. Chen, X. Liu, Y. Wang, A. Ragni, J. H. M. Wong, and M. J. F. Gales, "Exploiting future word contexts in neural network language models for speech recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 9, pp. 1444–1454, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2922048

[3] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 539–546 vol. 1.

[4] T. Ge, F. Wei, and M. Zhou, "Reaching human-level performance in automatic grammatical error correction: An empirical study," *CoRR*, vol. abs/1807.01270, 2018.

[5] K. Knill, M. Gales, P. Manakul, and A. Caines, "Automatic grammatical error detection of non-native spoken learner English," in *Proc. ICASSP*, 2019.

[6] K. Kyriakopoulos, M. Gales, and K. Knill, "Automatic characterisation of the pronunciation of non-native English speakers using phone distance features," in *Proceedings of Workshop on Speech and Language Technology for Education (SLaTE)*, 2017.

[7] K. Kyriakopoulos, K. Knill, and M. J. F. Gales, "A deep learning approach to assessing non-native pronunciation of english using phone distances," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 1626–1630. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1087

[8] Y. Lu, K. Knill, M. J. F. Gales, P. Manakul, LinlinWang, and Y. Wang, "Impact of asr performance on spoken grammatical error detection," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association.*, 2019.

[9] A. Malinin, A. Ragni, M. Gales, and K. Knill, "Incorporating uncertainty into deep learning for spoken language assessment," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[10] A. Malinin and M. J. F. Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems*

(NeurIPS), 2018, pp. 7047–7058. [Online]. Available: http://papers.nips.cc/paper/7936-predictive-uncertainty-estimation-via-prior-networks

[11] A. Malinin, K. Knill, and M. J. F. Gales, "A hierarchical attention based model for off-topic spontaneous spoken response detection," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, 2017, pp. 397–403. [Online]. Available: https://doi.org/10.1109/ASRU.2017.8268963

[12] A. Malinin, R. C. van Dalen, K. Knill, Y. Wang, and M. J. F. Gales, "Off-topic response detection for spontaneous spoken english assessment," in *ACL*, 2016.

[13] N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for call," in *2006 IEEE Spoken Language Technology Workshop*, Dec 2006, pp. 126–129.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MITPress, 2006.

[15] M. Rei, G. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 2016, pp. 309–318.

[16] R. van Dalen, K. Knill, and M. Gales, "Automatically grading learners' English using a Gaussian Process," in *Proc. ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 2015.

[17] Y. Wang, J. H. M. Wong, M. J. F. Gales, K. M. Knill, and A. Ragni, "Sequence teacher-student training of acoustic models for automatic free speaking language assessment," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, 2018, pp. 994–1000. [Online]. Available: https://doi.org/10.1109/SLT.2018.8639557

[18] Y. Wang, M. J. F. Gales, K. M. Knill, K. Kyriakopoulos, A. Malinin, R. C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken english," *Speech Communication*, vol. 104, pp. 47–56, 2018.

[19] Z. Yuan and T. Briscoe, "Grammatical error correction using neural machine translation," in *HLT-NAACL*, 2016.