UNIVERSITY OF CAMBRIDGE

MLMI14: Spoken Document and Meeting Summarisation

Mark Gales

Lent 2021

ALTA Institute - Spoken Language Assessment & Learning

- Automate (English) spoken language assessment &learning
  - without simplifying/limiting form of test
  - desire to assess communication skills
- Detail: `http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html`

CEFR - Levels of Foreign Language (L2) Learning

- Internationally agreed standard for assessing levels
  - Common European Framework of Reference (CEFR)
- Basic User

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

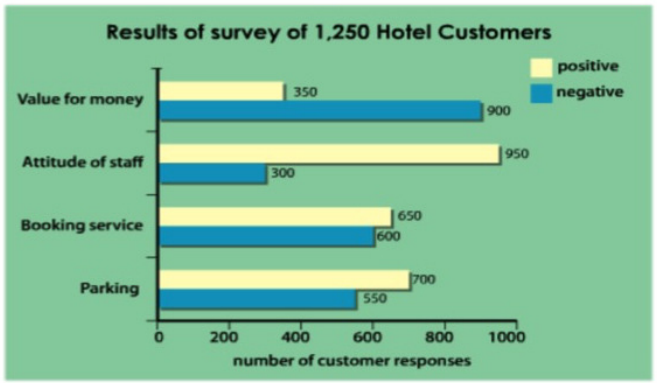
  B1 - threshold or intermediate
  B2 - vantage or upper intermediate
- Proficient User

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

Linguaskill (BULATS) Spoken Tests

- 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



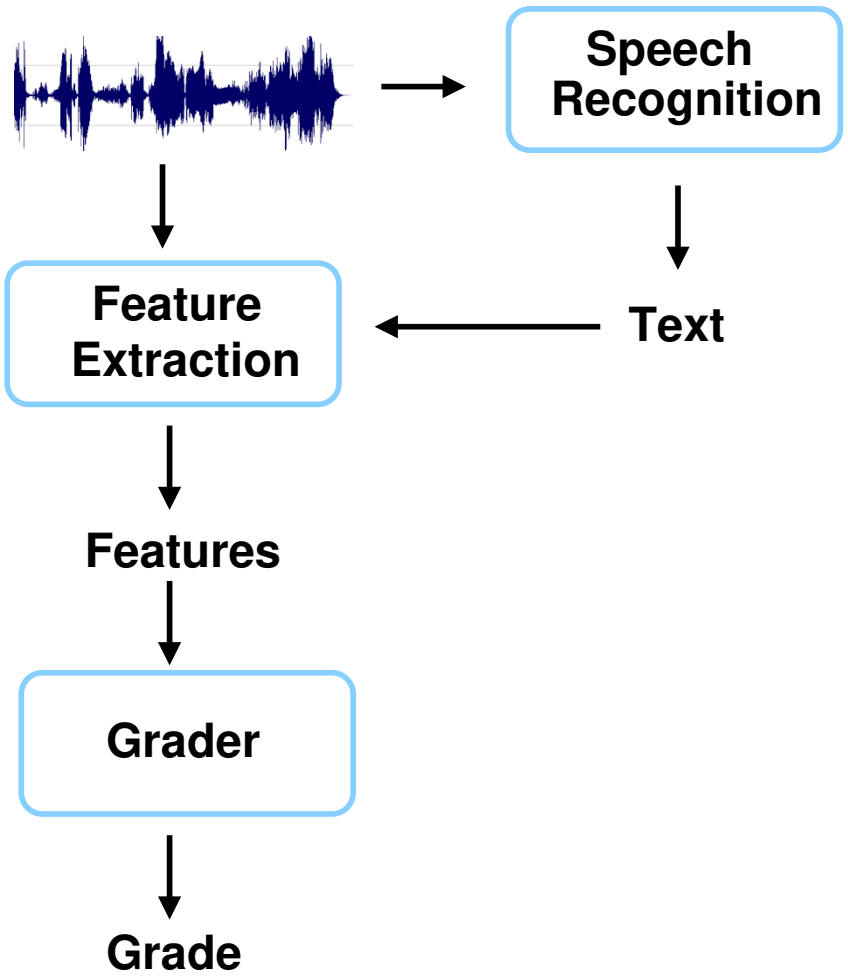
D Interpret and Discuss Chart/Slide: example above
E Answer Topic Questions: 5 questions about organising a meeting

Overview

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

Assessment

Assessment Framework [16]



- Key Challenges:
  - large range of L1s
  - wide-range of ability
  - non-native speech
  - high WERs

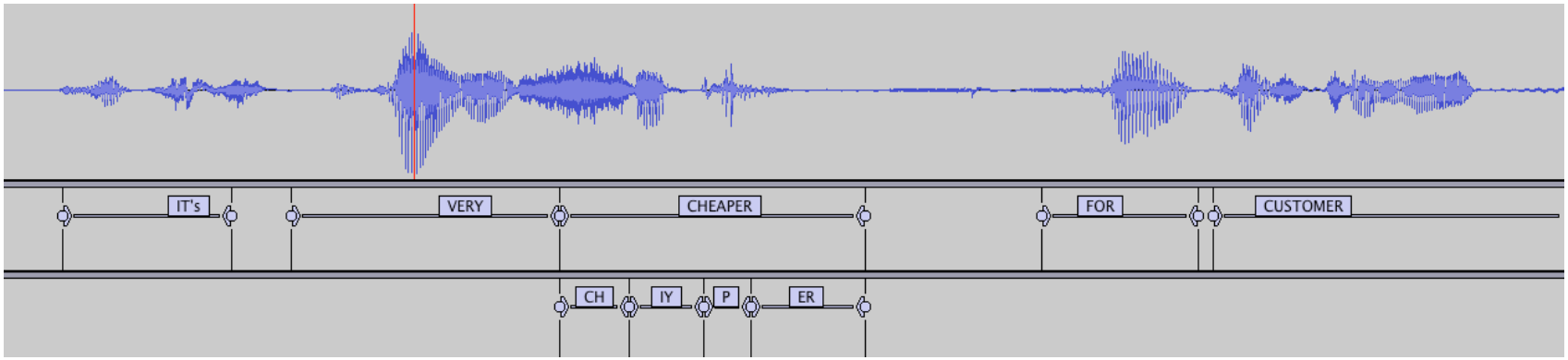
Transcribing Non-Native Speakers [14]

- WERs between pairs of professional transcription services:

Service	compared with	Error rate
Lifeline	Transcription Star	21.8
Lifeline	Lifeline India	26.8
Transcription Star	Lifeline India	25.4

- Hard-to-transcribe data: noise, non-native speakers
- Not possible to get high-quality transcripts - crowd-source
  - cost $< \frac{1}{10}$ as much as professional transcriptions, fast
  - but low-level accuracy of transcripts
  - crowd-turkers may try and cheat ...
- Combine multiple crowd-sourced transcripts together
  - simplest - take the longest, reasonable initial approach
  - refine using ASR-based combination (from initial transcription)

Automatic Speech Recognition [15, 2]

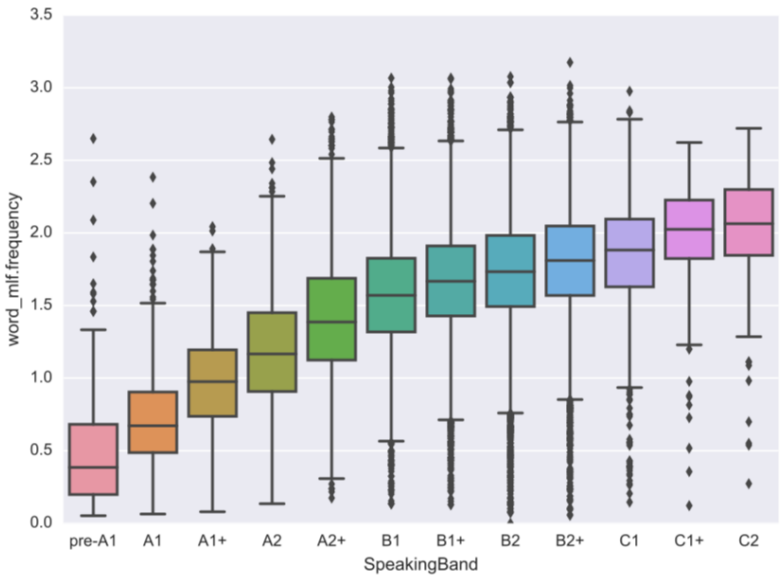


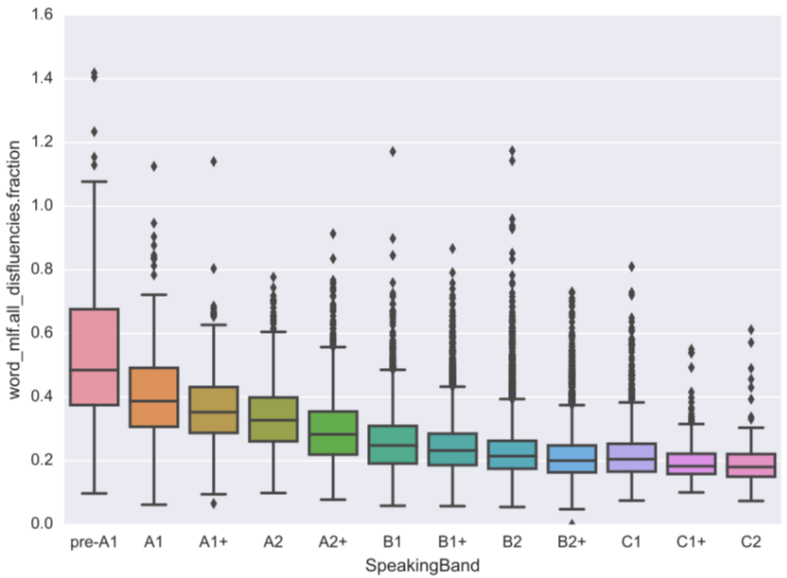
- Baseline Automatic Speech Recognition (ASR) yields:
  - word/disfluencies/partial-word sequence and time-stamps
  - phone-sequences and associated time-stamps
- Deep-learning based ASR systems used
  - Kaldi-based lattice-free MMI acoustic models
  - ensemble combination use sequence teacher-student training
  - rescoring with RNNLM and su-RNNLM based language models

Baseline Grader Features [16]

- Baseline features mainly fluency based, includes:
- 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 (per word)

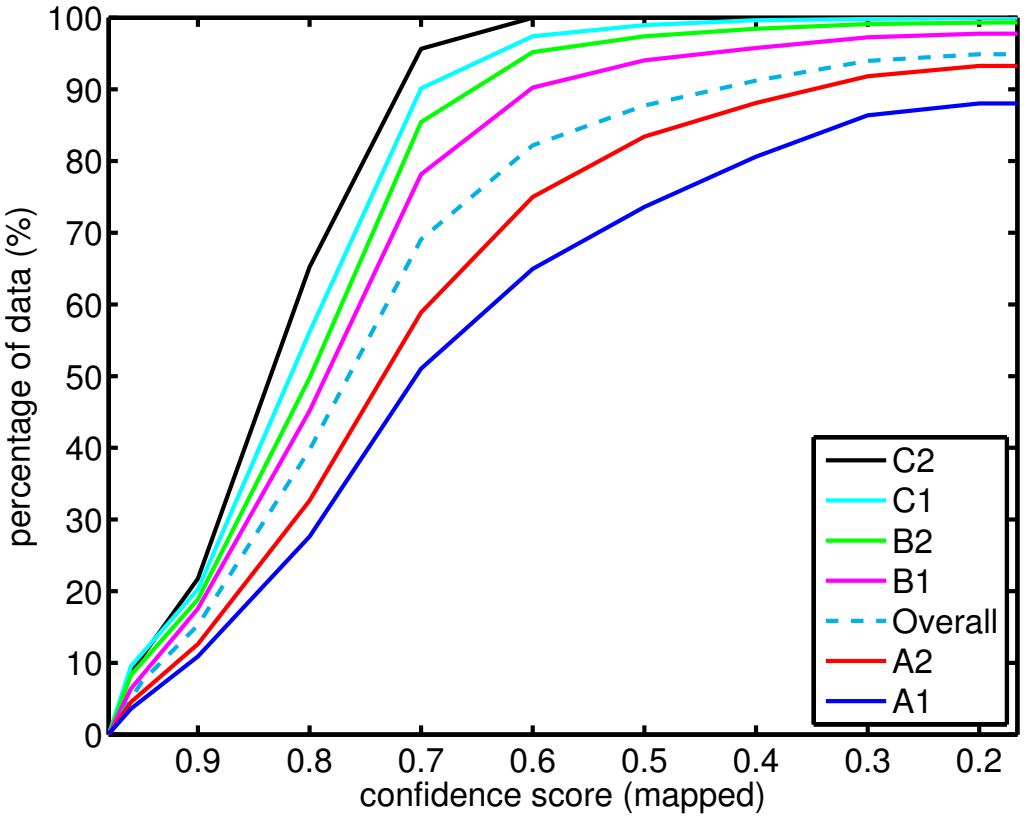
Baseline Features: Correlation with Grades





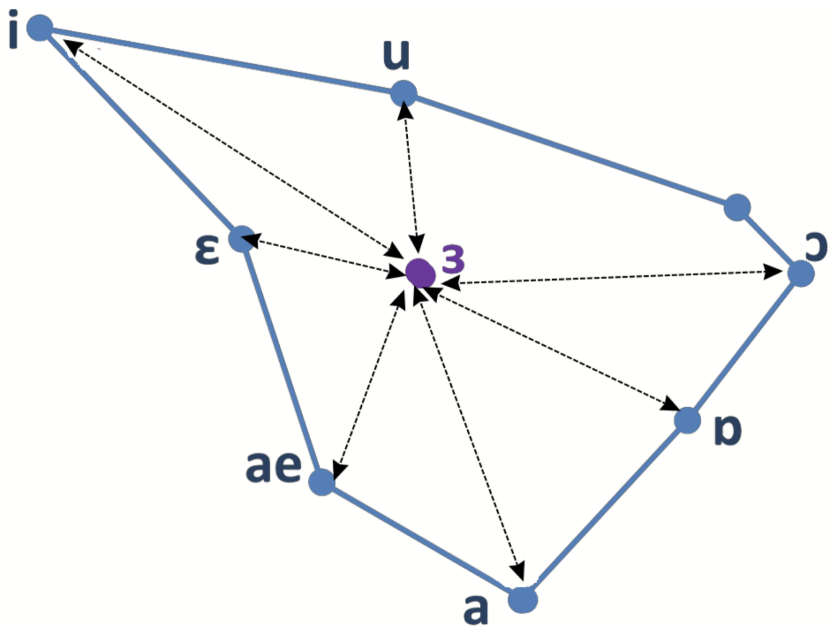
- Examine distribution of extracted features with grade
  - example box-plots for speaking rate and percentage disfluencies

ASR System Features - Confidence Scores



- Compute confidence scores for each candidate
  - plot of cumulative % data for each grade shown

Derived Features: Phone-Distances [9]



- Pronunciation changes as speakers become more proficient
- Incorporate pronunciation features
  - no native speech data in target domain
  - need to remove speaker attributes from pronunciation
- Phone-distance features (relative position) one approach

Model-Based Pronunciation Features

- Build statistical model of each phone, measure 'distance"
  1. each phone p_i, each speaker s estimate Gaussian:

$$p(\mathrm{p}_i|\mathrm{s}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(si)}, \boldsymbol{\Sigma}^{(si)})$$

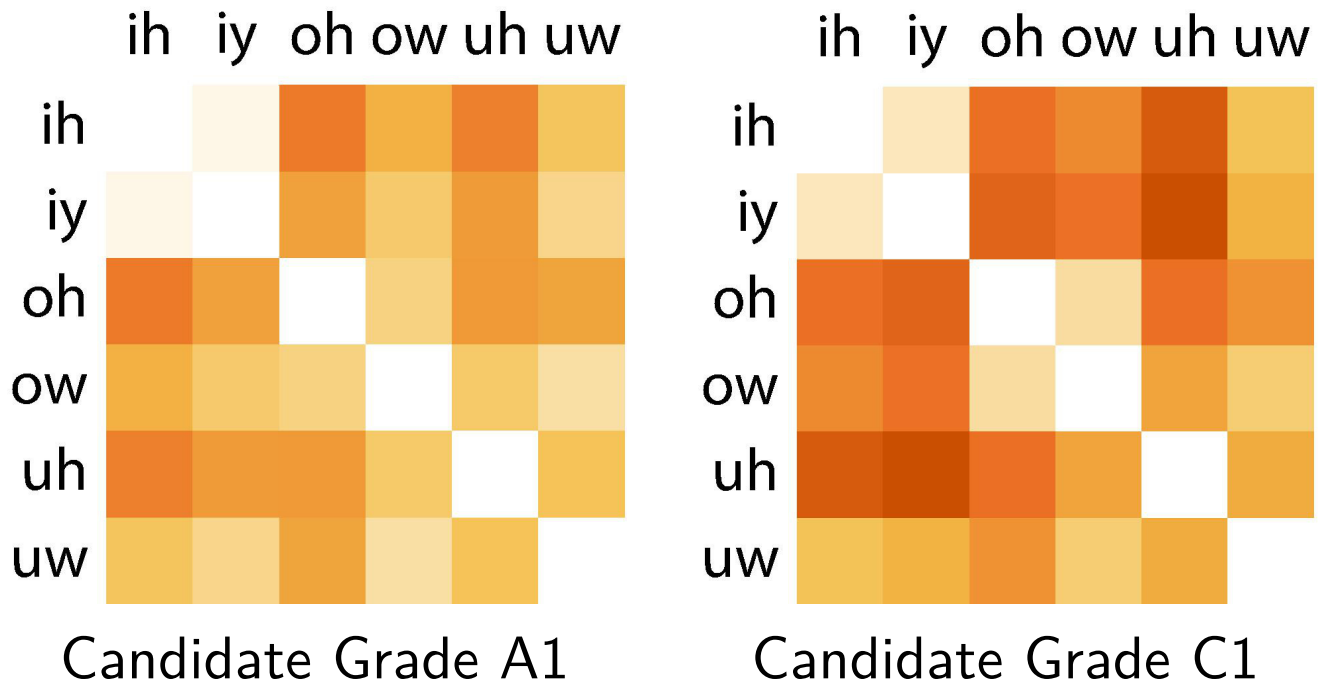
     this yields P Gaussian models
  2. compute $P \times P$ symmetric matrix $\mathbf{M}^{(s)}$

$$m_{ij}^{(s)} = \mathcal{KL}\left(p(\mathrm{p}_i|\mathrm{s}), p(\mathrm{p}_j|\mathrm{s})\right) + \mathcal{KL}\left(p(\mathrm{p}_j|\mathrm{s}), p(\mathrm{p}_i|\mathrm{s})\right)$$

$$\mathcal{KL}\left(p(\mathrm{p}_i|\mathrm{s}), p(\mathrm{p}_j|\mathrm{s})\right) = \int \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(si)}, \boldsymbol{\Sigma}^{(si)}) \log\left(\frac{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(si)}, \boldsymbol{\Sigma}^{(si)})}{\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(sj)}, \boldsymbol{\Sigma}^{(sj)})}\right) d\boldsymbol{x}$$

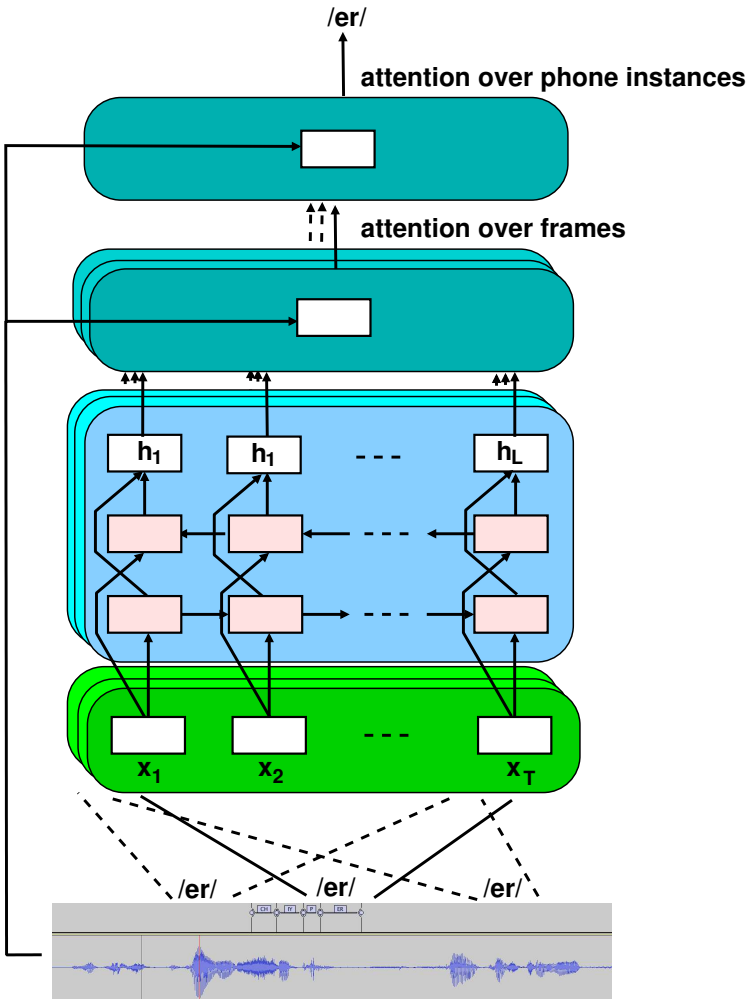
- Pattern in $\mathbf{M}^{(s)}$ expected to be first language (L1) dependent
  - use vector of matrix as features for assessment

Pronunciation Features

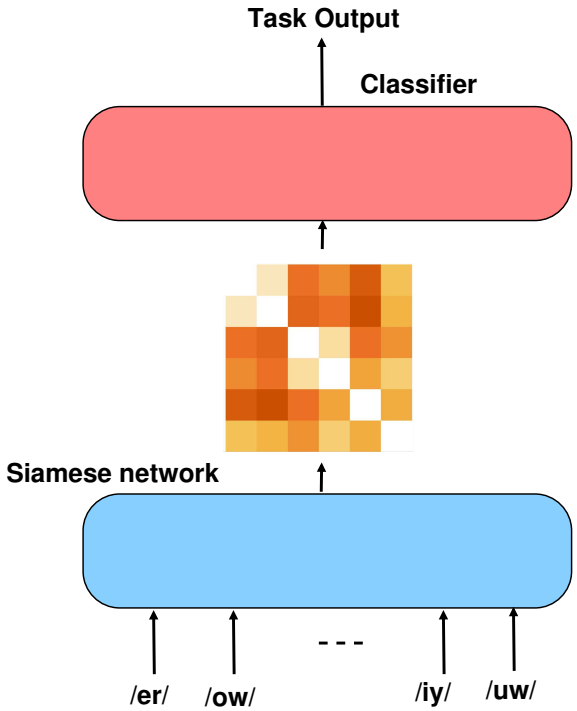


- Pattern of distances different between different levels
  - yields small gains in assessment performance

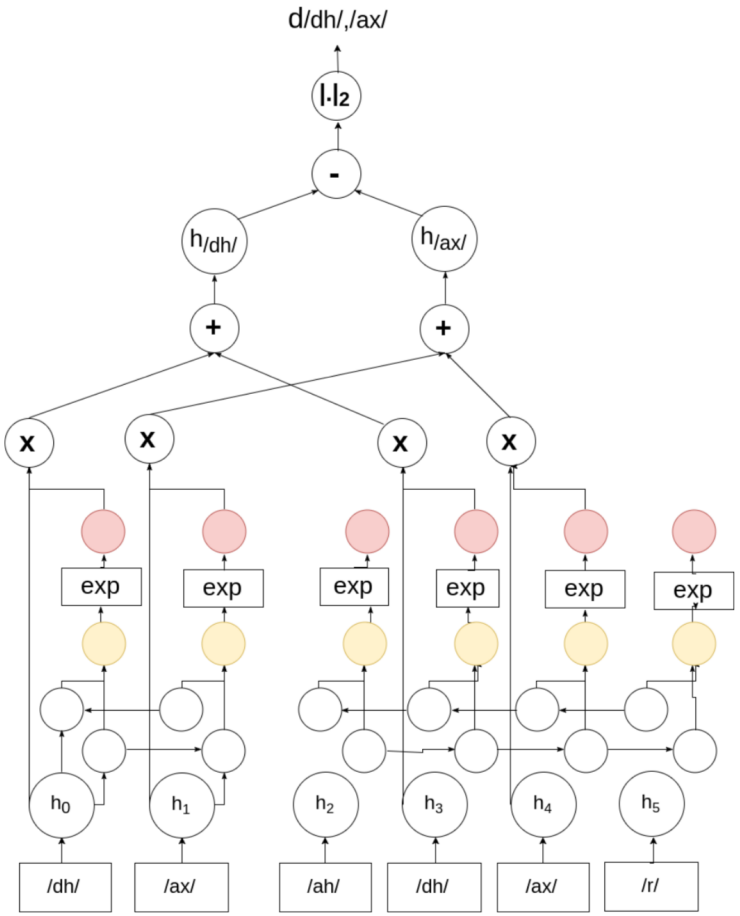
Deep Learning Pronunciation Features [7]



ASR phone alignemnt



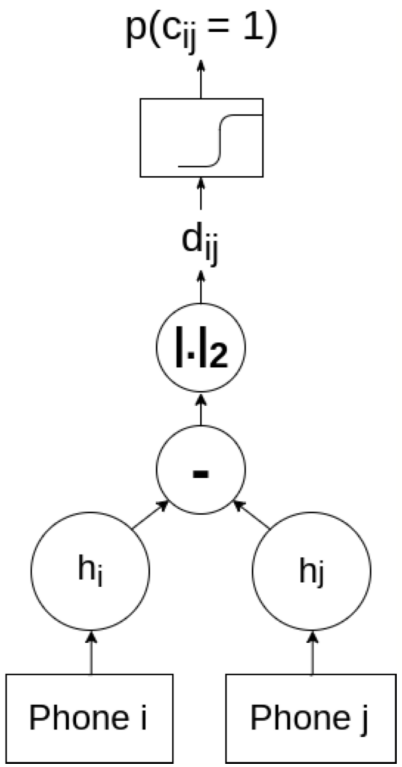
Deep Learning Pronunciation Features [7]



- Bi-directional encoding of frames
- Attention over encodings
  - maps to fixed length
  - attention uses embeddings $\boldsymbol{h}$, $\tilde{\boldsymbol{h}}$
- Attention over phone instances
  - maps number of phones to one
  - attention uses word phone sequence
- Siamese network for phone distance
- Complete system tuned to task
  - grader performance
  - language identification
- Distance tunable to task

Siamese Networks [3]

- Siamese networks map features to meaningful distance space



- Train distances for classification

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

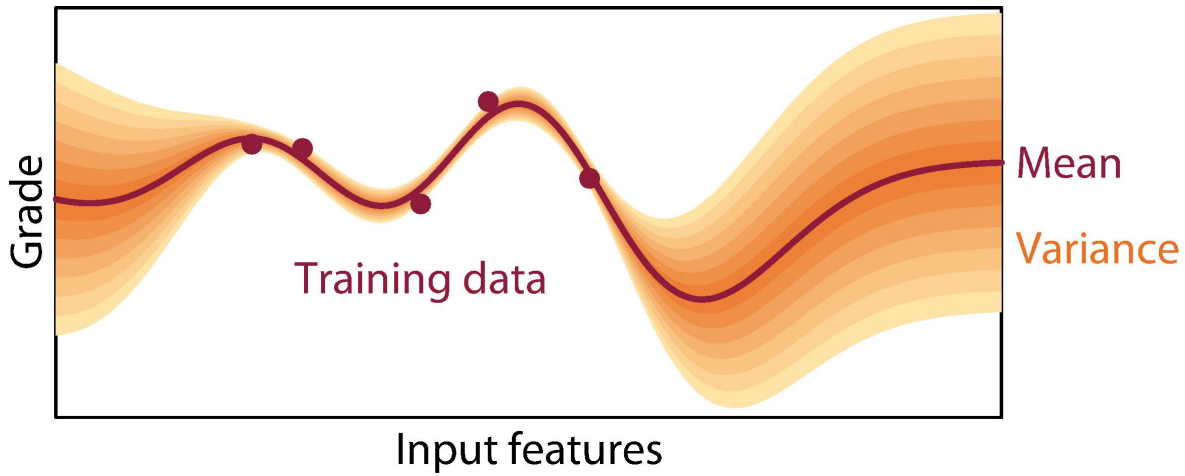
  - maps features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to new space
  - parameters of mapping network the same $\boldsymbol{\theta}$

- 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

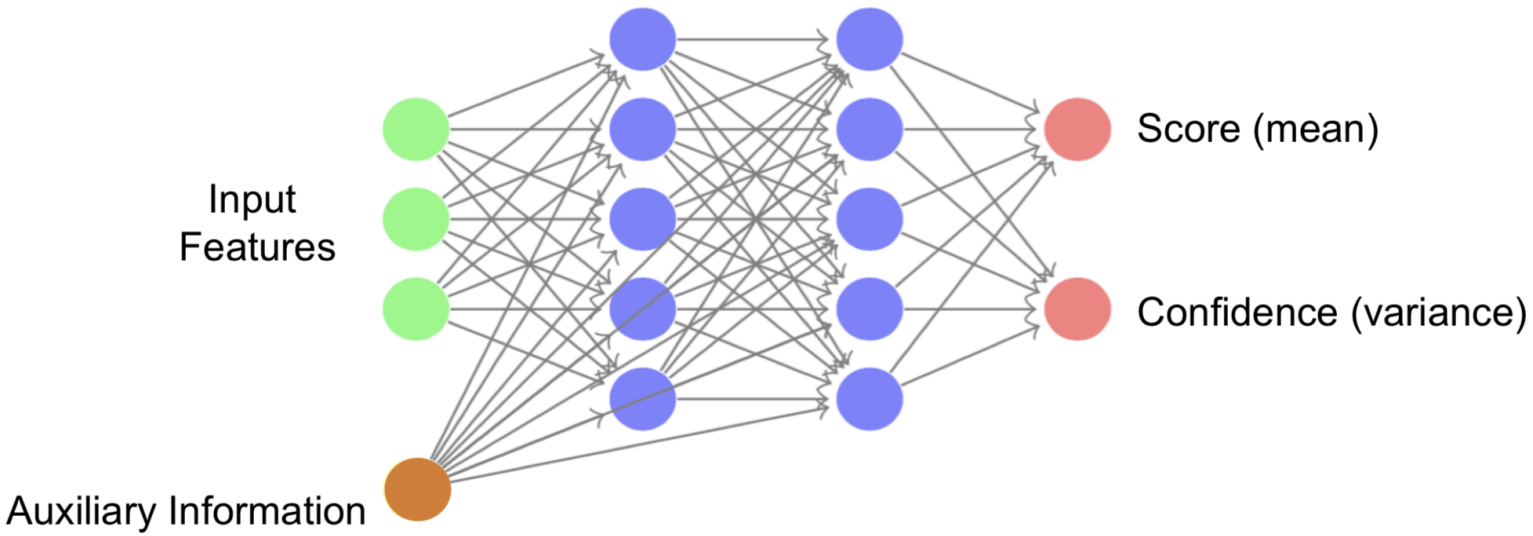
Assessment: Gaussian Process

- Supervision data assessment is a score (0-6)
  - assessment run as a regression task: $p(y|\boldsymbol{x}^\star;\boldsymbol{\theta})$
- Gaussian process standard approach
  - non-parametric model based on joint-Gaussian assumption



  - mean is used as the score prediction
  - confidence (variance) standard aspect of model
  - approximations required as quantity of data increases

Deep Learning: Deep Density Networks



- Deep Density Networks predict distribution parameters

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

  - network predicts parameters of distribution
  - flexible framework for any form of distribution
  - simple to include additional (auxiliary) information, e.g. L1

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

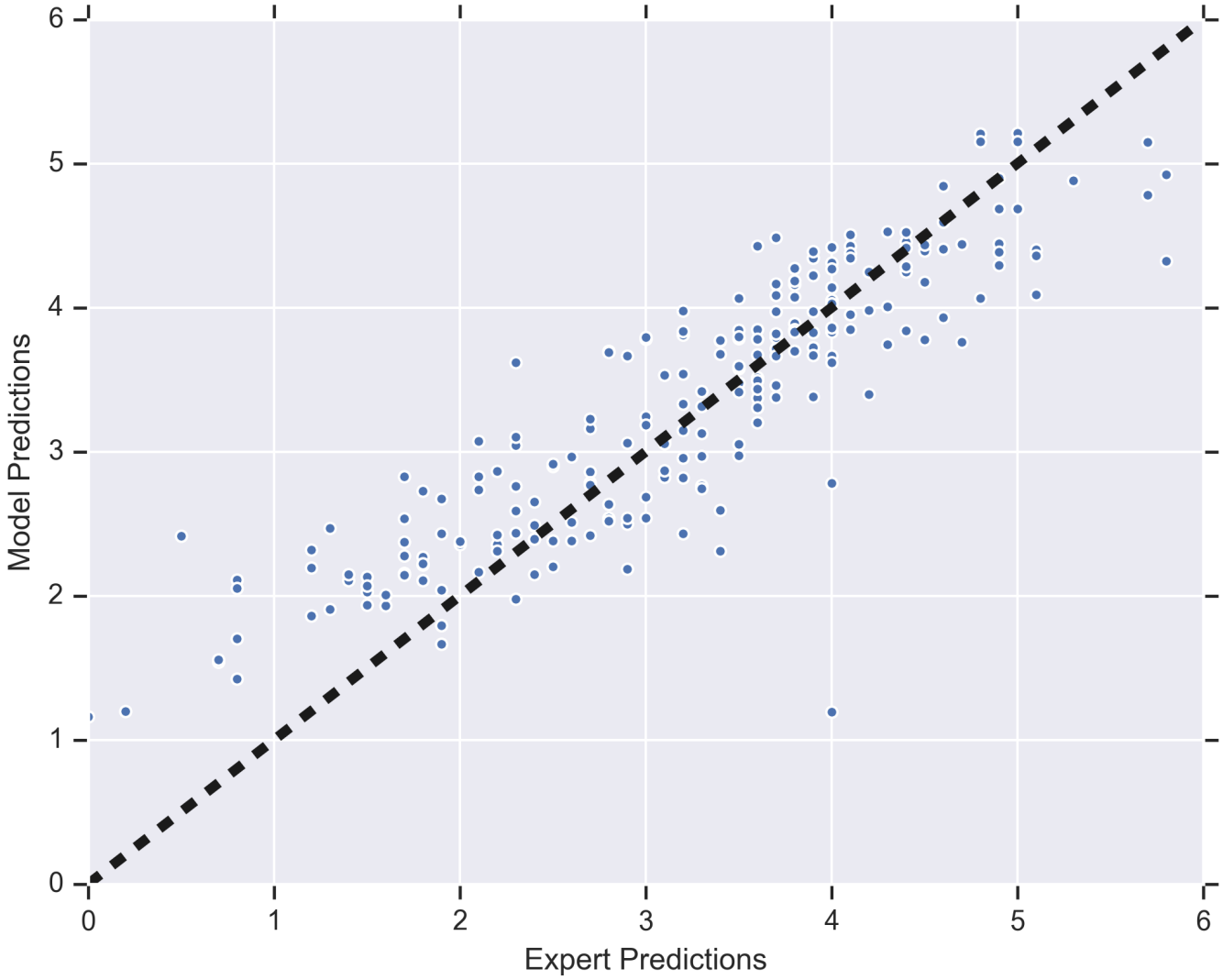
  - "basic users" (A1/A2) highly challenging data

- Assessment: using complete test

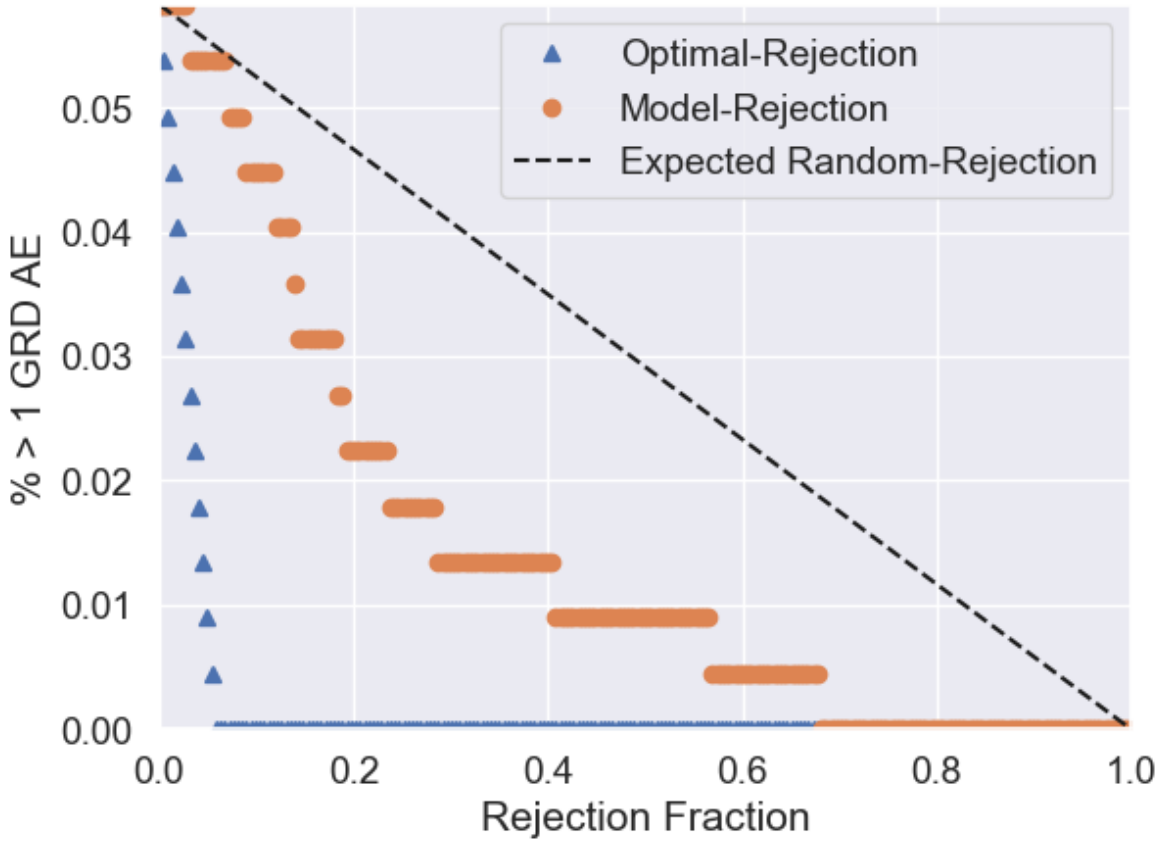
PCC	MSE	%≤ 0.5	%≤ 1.0
0.888	0.31	68.2	94.2

  - ≤ 1.0 indicates within one CEFR grade-level

Performance Analysis



Incorporating Ensemble-Based Uncertainty [17]



- 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

Malpractice:

Off-Topic Response Detection and Adversarial Attacks

Assessment Malpractice

- Applying systems to high-stakes tests challenging, requires:
  - highly accurate predictions
  - confidence in predictions (grades) generated
- Need approaches to address malpractice
  - candidates use "pre-recorded" responses
  - different candidate takes test after enrolment
  - need to detect candidates speaking their L1 (native tongue)
  - adversarially attack spoken deep-learning systems
- Here off-topic detection and adversarial attacks discussed
- Note: off-topic not necessarily "malpractice"

Off-Topic Responses

Title Dissection

QUESTION:
(aka *prompt*)

Describe your role at your workplace.

ANSWER 1:
(aka *response 1*)

"Well, I primarily work on Excel sheets with lots of manual number crunching. A lot of time is spent discussing potential projects with clients too."

ANSWER 2:
(aka *response 2*)

"[Ahem] The age-old rivalry between Cambridge and Oxford is now outdated due to the superiority of the former in nearly all respects. "

- Range of possible causes:
  - candidate unable to formulate valid response
  - candidate does not understand the prompt/question
  - candidate deliberately "cheats"

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

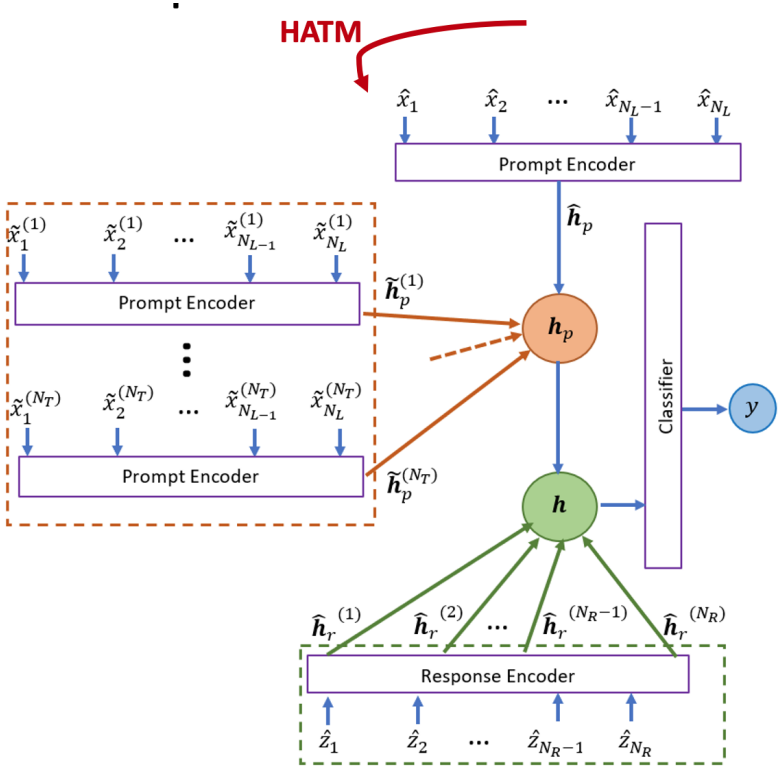
    $$\boldsymbol{w}^r = \{\texttt{Cambridge Assessment is wonderful, it ...}\}$$

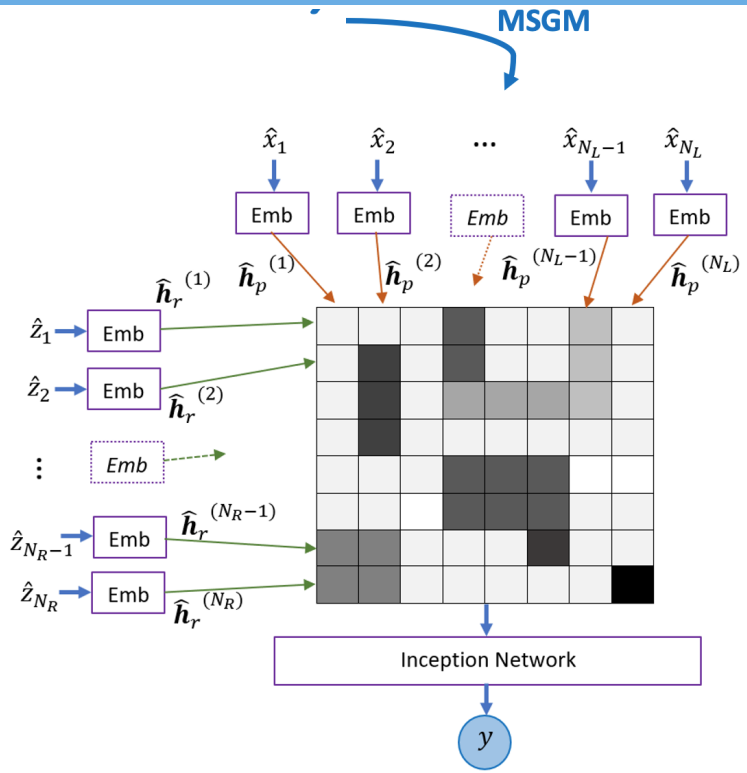
  and derives probability of relevance

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

- Need mechanism to map prompt/response to fixed length vector

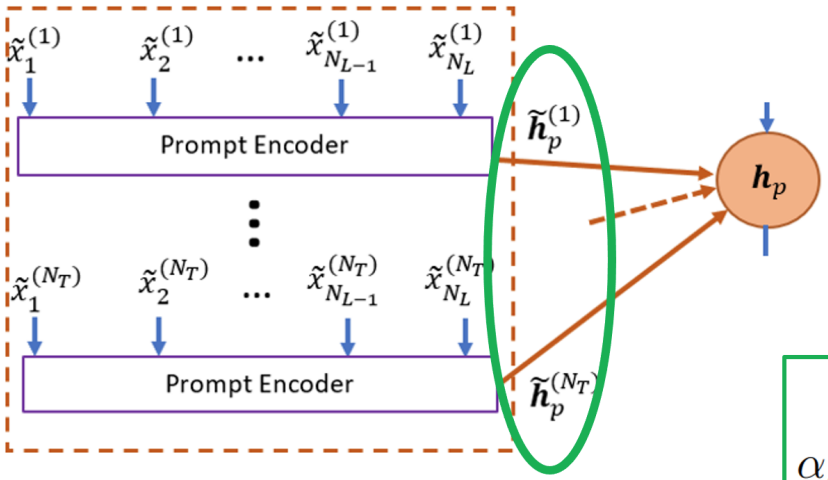
"Bespoke" Detection Architectures [11]





- HATM: use attention mechanisms over prompt/response
  - attention over training prompts, attention over response
- MSGM: treat as a vision classification task
  - cosine-distance between (embedded) prompt/response words
  - scale matrix to standard grid (image) size

Generalisation: Prompt-Dropout



$$\mathbf{h}_p = \sum_{k=1}^{N_T} \alpha_k \tilde{\mathbf{h}}_p^{(k)}$$

$$\alpha_k = \begin{cases} 0 & \text{w.p. } 1-\kappa \\ \texttt{Softmax}\left[\hat{\mathbf{h}}_p, \tilde{\mathbf{h}}_p^{(k)}, \boldsymbol{\theta}_{pa}\right] & \text{w.p. } \kappa \end{cases}$$

- Only limited number of prompts available to train system
  - large number of responses for each of the prompts
  - challenge to generalise well to unseen prompts
- Use modified version of dropout that "drops" training prompts

Data-Augmentation: "Easy Data Augmentation" and Translation

STRUCTURED TECHNIQUES – *Easy Data Augmentation (EDA)*

Describe your role at your workplace.

1) *Describe your position at your workplace.*

2) *Describe your typical role at your workplace.*

3) *Describe your workplace at your role.*

4) *Describe _ role at your workplace.*

UNSTRUCTURED TECHNIQUES – *Back Translation*

Describe your role at your workplace.

صف دورك في مكان عملك.

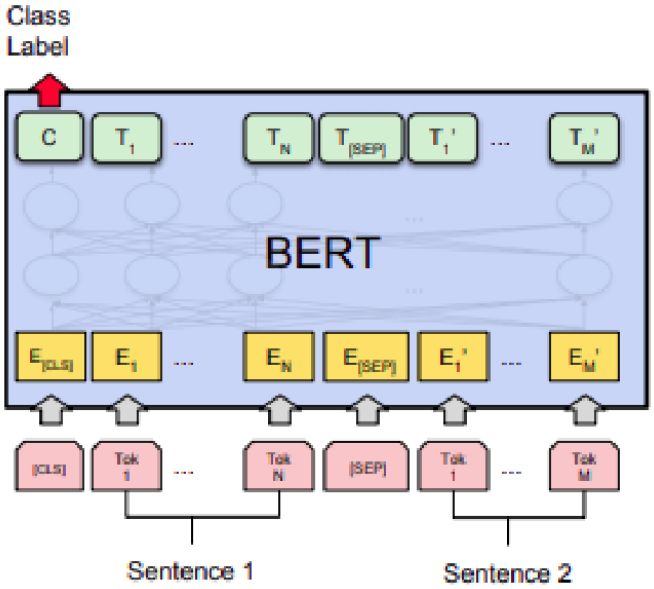
Talk about what your job entails while you are at work.

"Well, I primarily work on Excel sheets with lots of manual number crunching. A lot of time is spent discussing potential projects with clients too."

- Generate variants of the prompts - data augmentation
  - structured approach: modify words using expert knowledge
  - unstructured approach: use back-translation:

$$\texttt{English} \rightarrow \texttt{French} \rightarrow \texttt{English}$$

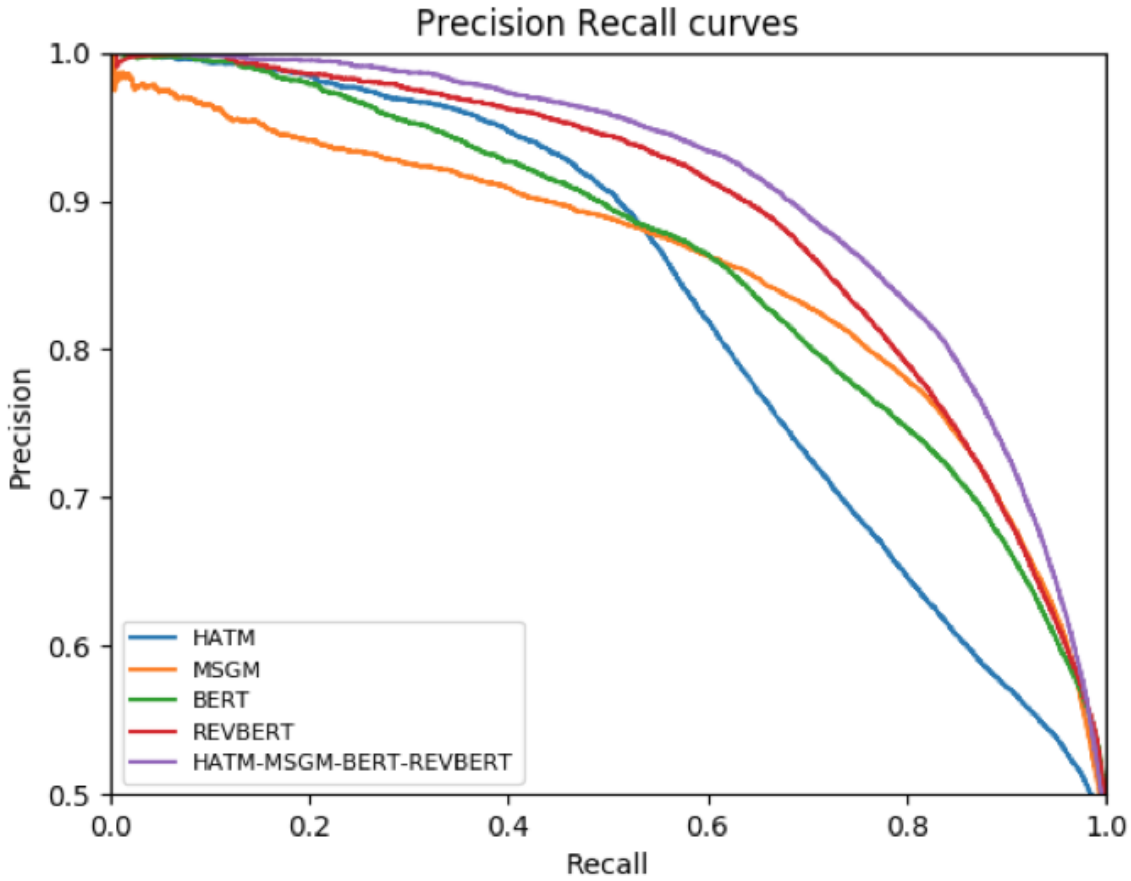
Pre-Trained Model: BERT



- Used BERT for word-embeddings
- Also trained for next sentence-prediction
  - uses sentence-pairs with [SEP]
- Use class label for classification

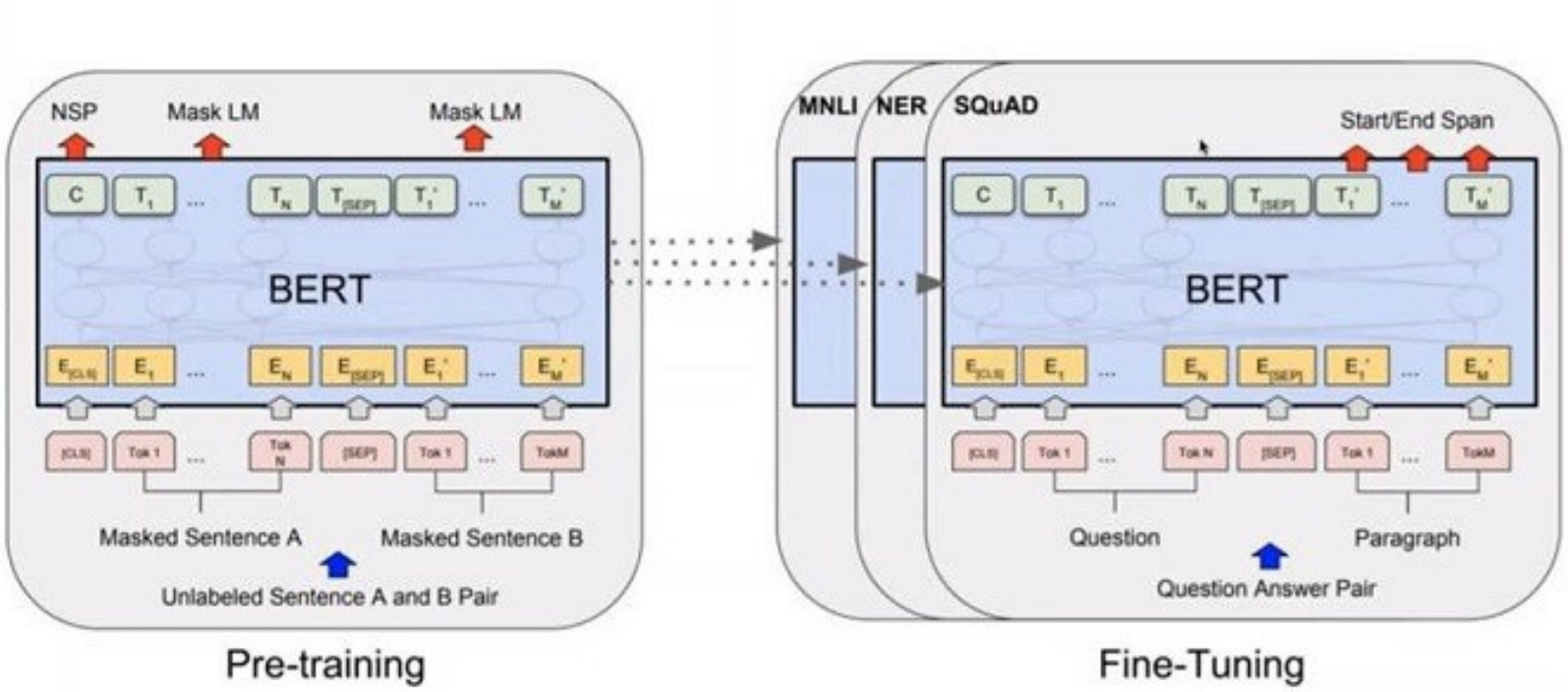
- Sentence 2 is predicted from sentence 1: choice
  - BERT: is the response valid for the prompt
  - REVBERT: is the prompt appropriate for the response

Performance: Linguaskill



- Prompts and responses not seen in training data
  - significantly better performance on prompts seen in training

Beyond On/Off Topic: Question Answering



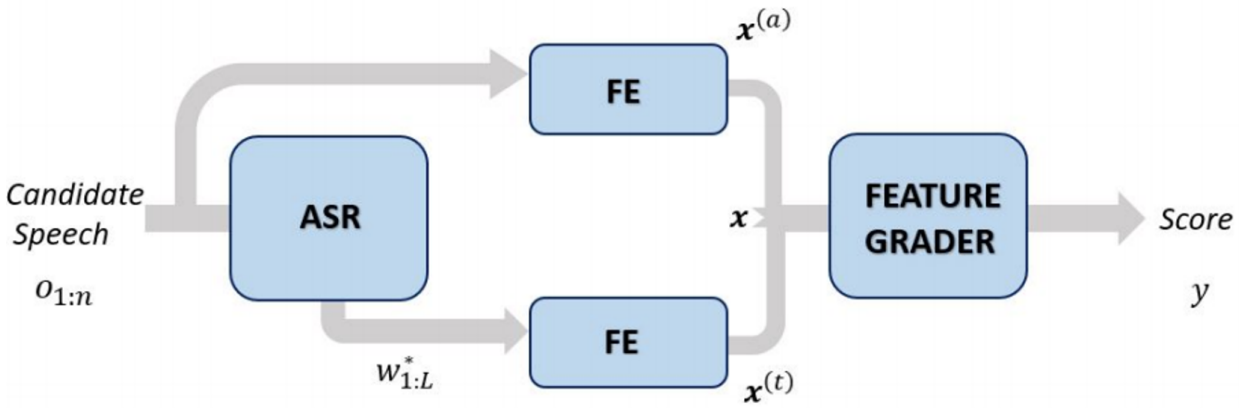
- If response relevant, where is the answer in the response:

  What is your favourite company?

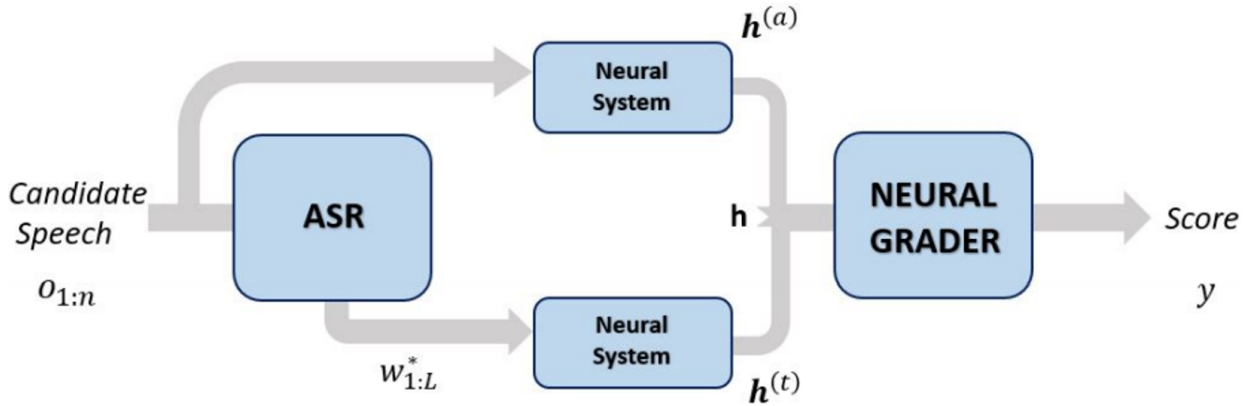
  My favourite company is Cambridge Assessment

Neural Spoken Language Assessment

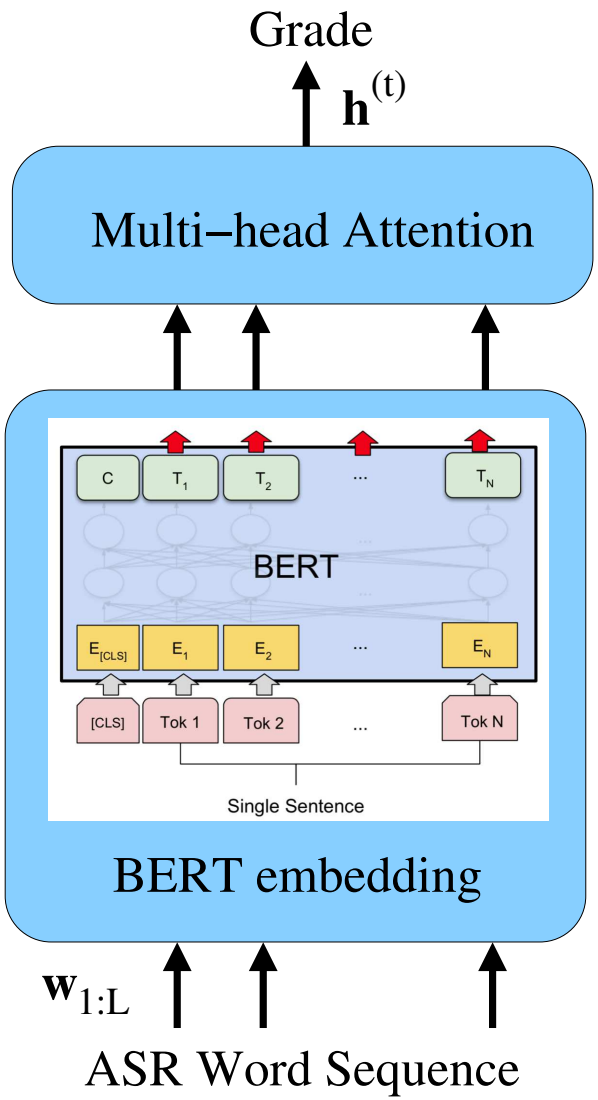
- System used expert features: $\{\boldsymbol{o}_{1:n}, \boldsymbol{w}_{1:L}\} \rightarrow \{\boldsymbol{x}^{(a)}, \boldsymbol{x}^{(t)}\} \rightarrow y$



- Replace by neural features: $\{\boldsymbol{o}_{1:n}, \boldsymbol{w}_{1:L}\} \rightarrow \{\boldsymbol{h}^{(a)}, \boldsymbol{h}^{(t)}\} \rightarrow y$



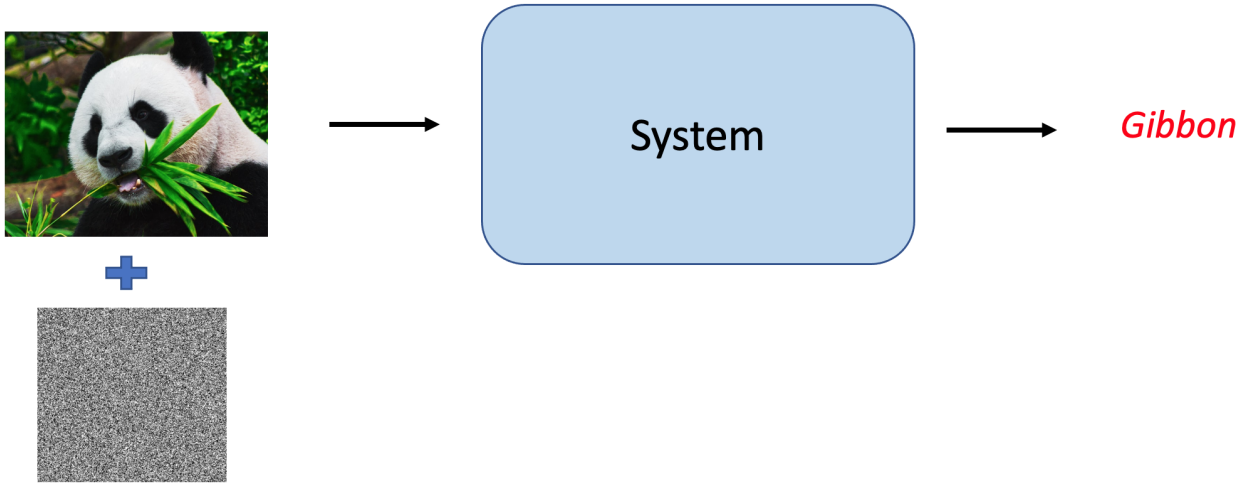
Neural Text Assessment



- Use BERT for word-embeddings
- Multi-head attention
  - maps embeddings to $\boldsymbol{h}^{(t)}$
- Feed-forward network predicts score
- Performance

System	Feat.	PCC	%<0.5	%<1.0
GP	All	0.881	60.5	91.4
	Text	0.855	60.4	87.7
Neural	Text	0.878	66.8	91.4

Adversarial Attacks



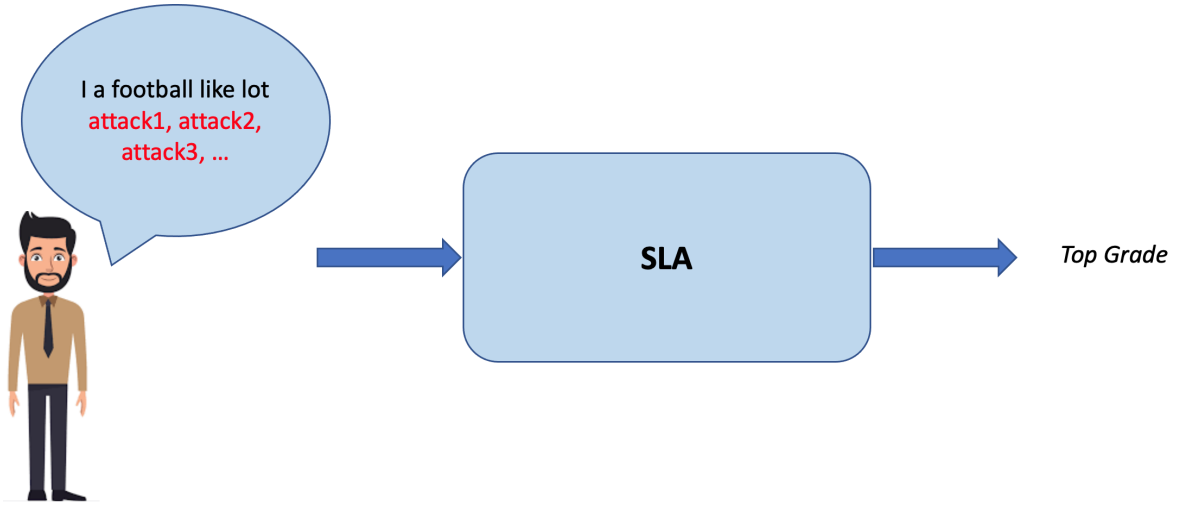
- Imperceptible perturbation, $\boldsymbol{\delta}^{(i)}$, changes system classification
  - un-targeted - simply change prediction $(t \neq \mathcal{F}(\boldsymbol{x}))$
  - targeted - swap to specific class t

$$\boldsymbol{\delta}^{(i)} = \arg\min_{\boldsymbol{\delta}} \left\{ \mathcal{F}(\boldsymbol{x}^{(i)} + \boldsymbol{\delta}) = t \right\} \quad \text{s.t. } \mathcal{H}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i)} + \boldsymbol{\delta}) < \epsilon$$

  - $\mathcal{H}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(i)} + \boldsymbol{\delta})$ is the measures of change in input

Universal Spoken Language Adversarial Attacks [12]

- Adversarially attack the ASR system output
  - add words (from vocabulary) to end of a response
  - examines universal attacks: same words for all candidates

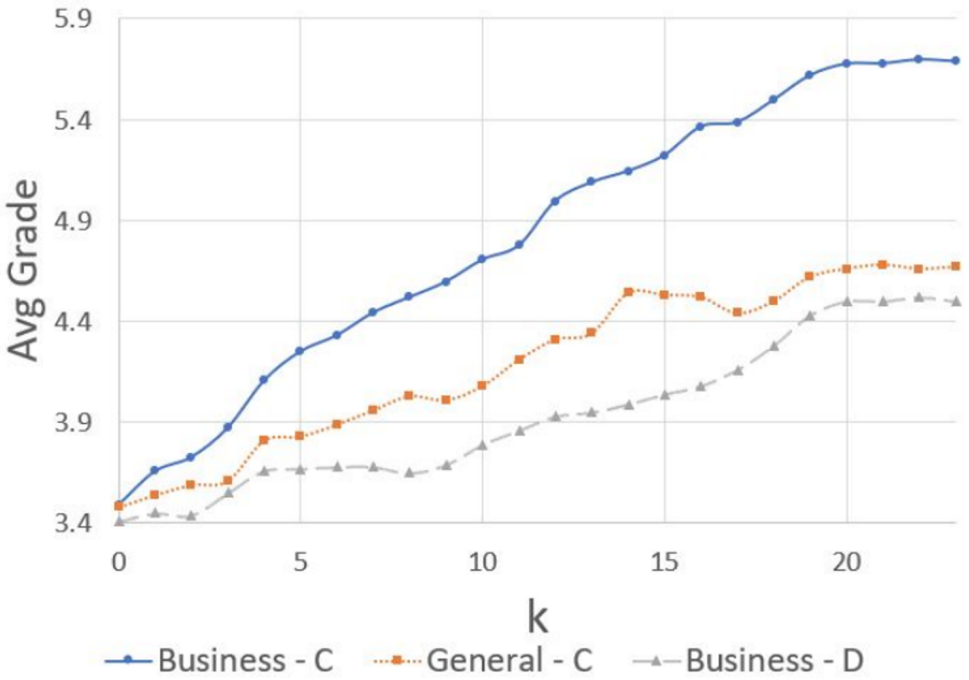


- Iteratively add word (from vocab. $\mathcal{V}$) to maximise score

$$\delta_k = \arg\max_{\delta \in \mathcal{V}} \left\{ \sum_{s=1}^{S} \mathcal{F}\left(\boldsymbol{w}^{(s)} \oplus \boldsymbol{\delta}_{1:k-1} \oplus \delta; \boldsymbol{\theta}\right) \right\}$$

$$\boldsymbol{w}_{1:n}^{(s)} \oplus \boldsymbol{\delta}_{1:k} = w_1^{(s)}, \ldots, w_n^{(s)}, \delta_1, \ldots, \delta_k$$

Impact of Adversarial Attacks



- For a 6 word attack
  - Business data section C

System	Attack	Avg. Score
GP	–	3.88
	GP	4.27
	Neural	4.02
Neural	—	3.49
	GP	3.54
	Neural	4.33

- Graph shows impact of number of words and transferability
  - adding 20 words yields almost perfect (6.0) average scores
  - same words transfers to different part of test (section D)
  - same words transfers to differing sets of prompts/task General-C

Spoken Language Adversarial Attack Detection

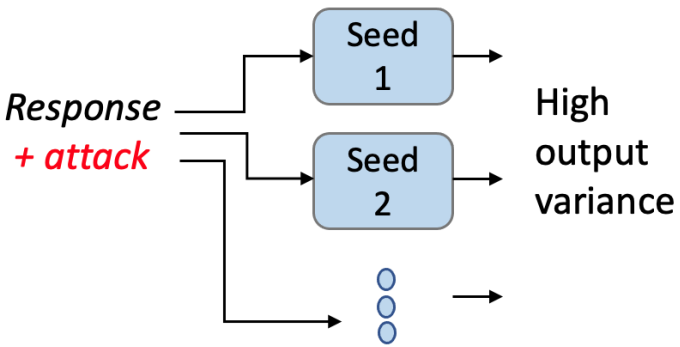
Perplexity

Response + attack
= doesn't conform to language model
= high perplexity

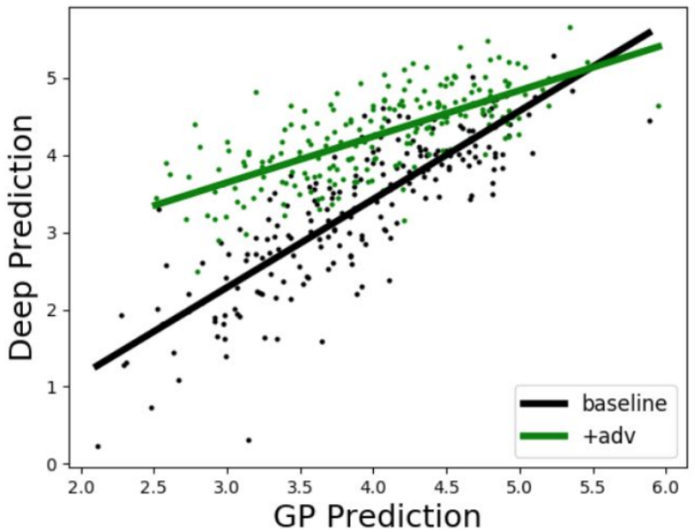
Topic Relevance

Response + attack
= off-topic to the prompt
= detected by separate system

Ensemble Variance

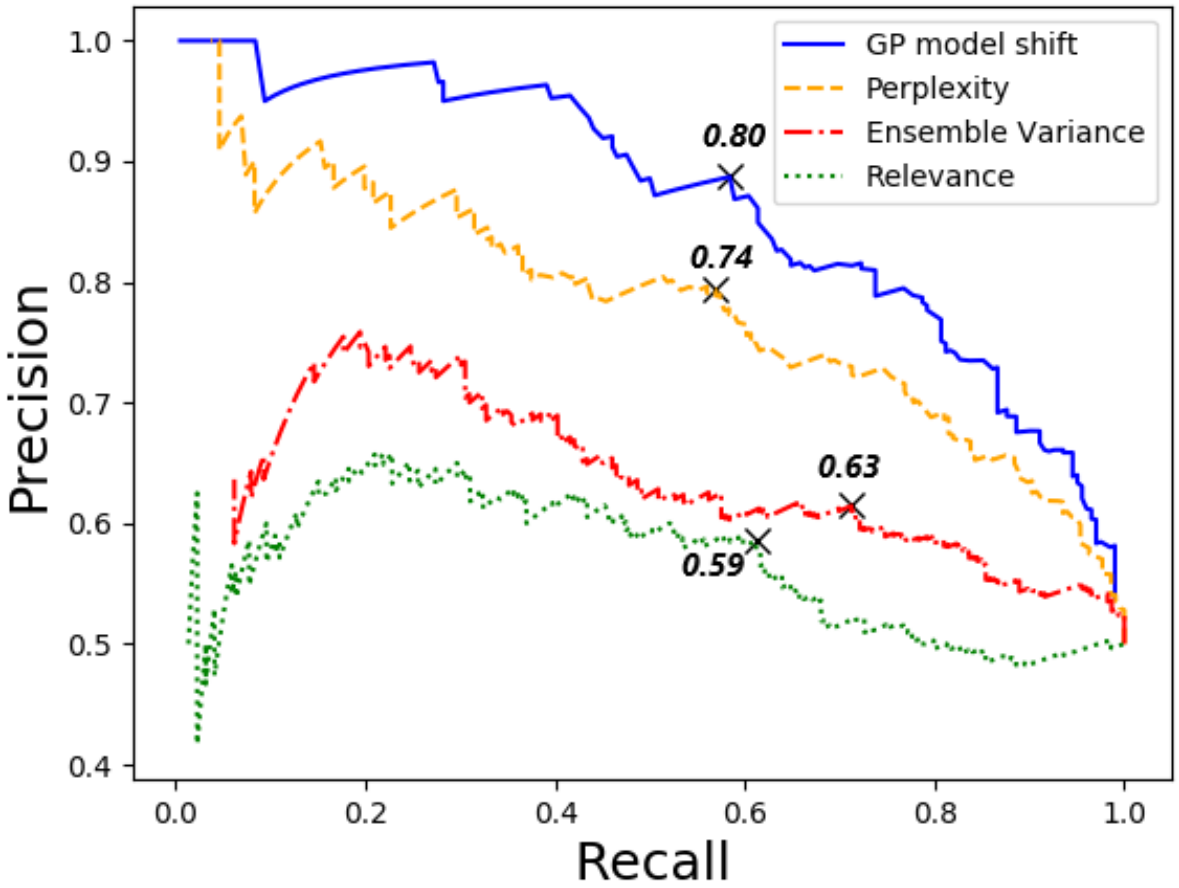


GP model shift



- Attempt to detect attacks - is the word sequence unusual

Spoken Language Adversarial Attack Detection



- Marked operating points $F_{0.5}$ - higher weight to precision

$$\mathtt{F-Score:}\quad F_\alpha = (1+\alpha^2)\frac{\mathtt{Pr}\times\mathtt{Re}}{\alpha^2\mathtt{Pr}+\mathtt{Re}}$$

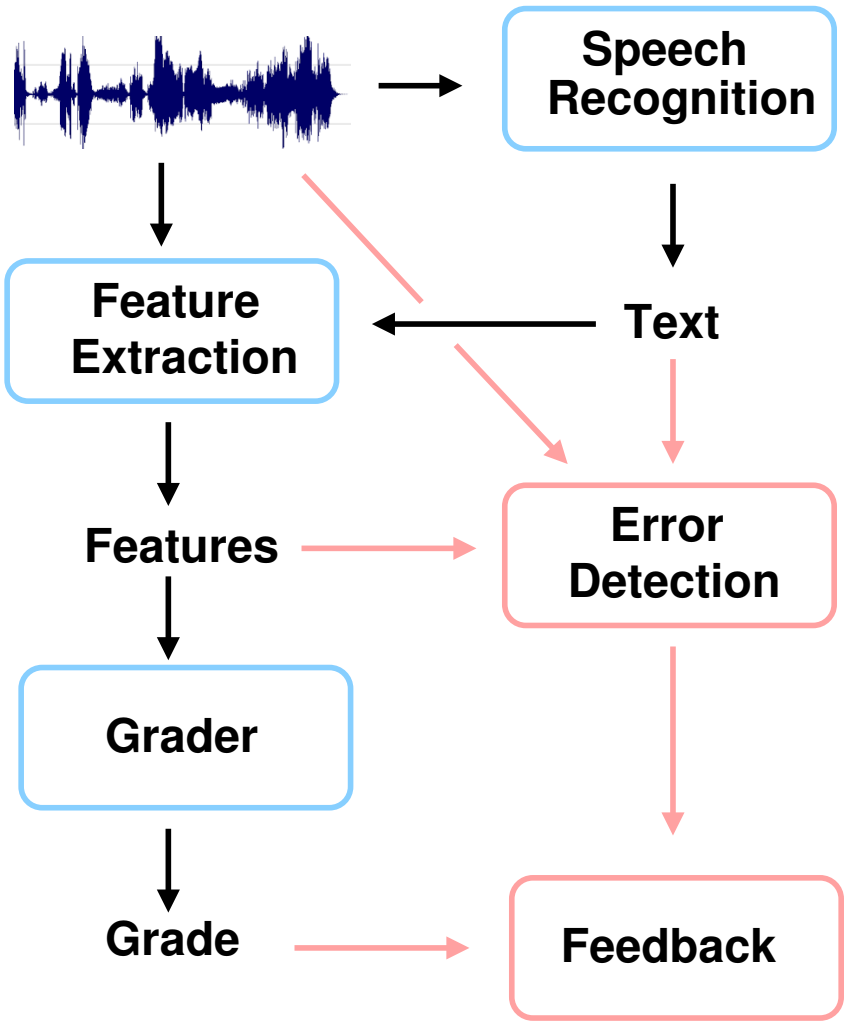
Feedback:

Grammatical Error Detection and Correction

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 candidate
- Message Construction:
  - is the response relevant to the prompt
  - is the message grammatically correct (in speech context)
  - is the message using the appropriate vocabulary
- Fluency:
  - is the pronunciation correct
  - is the correct intonation pattern used
  - is the speech delivered in a coherent fashion

Feedback Framework



- Key Challenges:
  - high WERs
  - wide-range of abilities
  - L1-specific errors
  - fluency feedback
  - grammatical feedback
  - requires high precision

Written Grammatical Error Correction

- Long history of research and products available e.g.
  - Grammarly
  - Write & Improve
- Aimed at feedback to improve writing skills
- Large(ish) English corpora available for training /evaluation
  - Cambridge Learner Corpus (CLC)
  - Lang-8 Corpus
  - CoNLL-2013/CoNLL-2014, JFLEG, FCE-Public

Written Errors

- Multiple forms of error possible:

She	see	Tom	,	is	catched	by	—	policman	at	—	park	.
She	saw	Tom	—		caught	by	a	policeman	in	the	park	.

  - missing articles (`the`/`a`)
  - incorrect preposition (`at`)
  - incorrect verb - declining/tense (`see`/`is catched`)
  - spelling mistakes (`policman`)
  - incorrect punctuation (`,`)

- Multiple possible corrections possible
  - may depend on context of sentence
  - may depend on writer's intent

Grammatical Error Detection (GED)

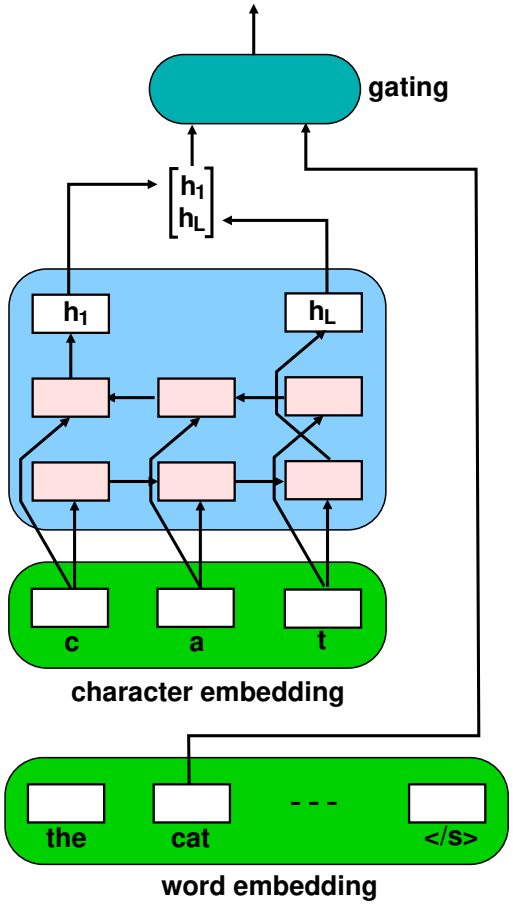
- Task is to label all tokens that are grammatically incorrect:

She	see	Tom	,	is	catched	by	—	policman	at	—	park	.
0	1	0	1	1	1	0	1	1	1	1	0	0

- Standard sequence labelling problem
  - use labelled corpora for supervised learning
  - deep-learning current state-of-the-art e.g. bidirectional LSTMs
- Assessment uses standard precision/recall evaluation
  - better than ROC when classes unbalanced
- How to handle words not in the training corpus? (OOVs)
  - rare words in English
  - spelling mistakes, e.g. `policman`

Character Embedding [13]

- Spelling mistakes/rare words may not be in vocabulary
  - use character $\boldsymbol{m}_i$ and word embedding $\boldsymbol{x}_i$ for word i



- bi-directional character encoding

$$\boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_L \end{bmatrix}; \quad \boldsymbol{m} = \boldsymbol{f}(\boldsymbol{h})$$

- gating function used to combine embeddings
  - "optimal" character/word combination

$$\tilde{\boldsymbol{x}} = \boldsymbol{z} \odot \boldsymbol{x} + (\mathbf{1} - \boldsymbol{z}) \odot \boldsymbol{m}$$

  - gating weight: $\boldsymbol{z} = \sigma(\mathbf{W}\boldsymbol{m} + \mathbf{U}\boldsymbol{h} + \mathbf{b})$
  - $\odot$ element wise multiplication
- Handles spelling mistakes/rare words
  - word not in word-embedding

Grammatical Error Correction

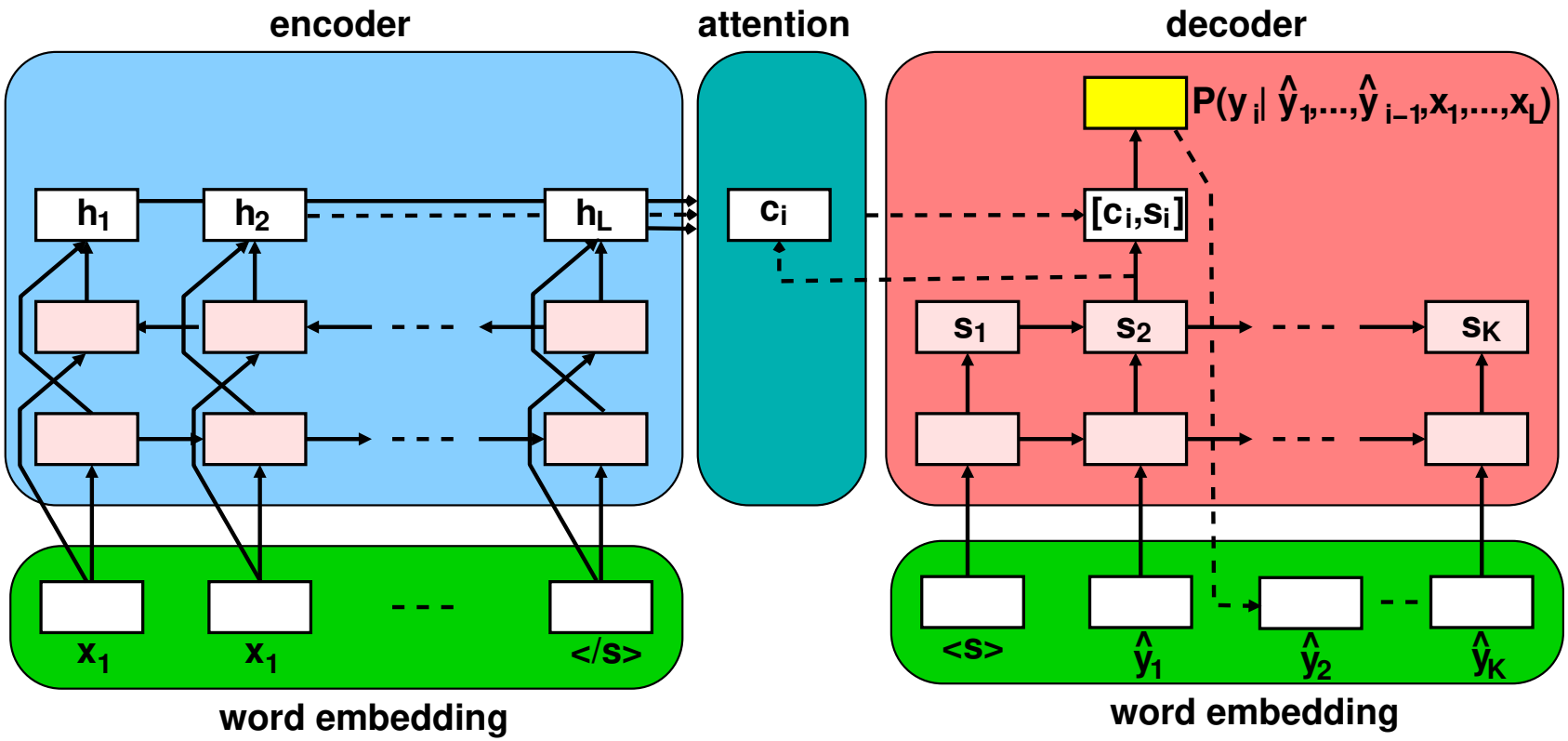
- Supply corrected form back to learner:

She	see	Tom	,	is	catched	by	—	policman	at	—	park	.
She	saw	Tom	—		caught	by	a	policeman	in	the	park	.

  - subsumes GED (can align to detect changes in output)

- Standard sequence-to-sequence translation problem
  - use machine translation approaches
  - originally statistical machine translation used
  - currently neural machine translation approaches popular
- Still need to handle out-of-vocabulary terms appropriately

(Simplified) Neural Machine Translation



- Encoder-decoder framework with attention
  - word-embedding maps word (`cat`) to vector (`emb(cat)`)
- Current state-of-the-art more complicated see MT papers

Neural Machine Translation - Attention

- Possible correction generated in an autoregressive fashion

$$\hat{y}_i \sim P(y_i|\hat{\boldsymbol{y}}_{1:i-1}, \boldsymbol{x}_{1:L}) \approx P(y_i|\boldsymbol{s}_i, \boldsymbol{c}_i)$$

where

$$\hat{y}_i \sim \mathcal{F}_{\texttt{softmax}}(\boldsymbol{s}_i, \boldsymbol{c}_i); \quad \boldsymbol{s}_i = \mathcal{F}_{\texttt{rnn}}(\boldsymbol{s}_{i-1}, \tilde{\boldsymbol{h}}_i); \quad \tilde{\boldsymbol{h}}_i = \mathcal{F}_{\texttt{rnn}}(\tilde{\boldsymbol{h}}_{i-1}, \texttt{emb}(\hat{y}_{i-1}))$$

$$\boldsymbol{c}_i = \mathcal{F}_{\texttt{att}}(\boldsymbol{s}_i, \boldsymbol{h}_{1:L}) = \sum_{\tau=1}^{L} \alpha_{i\tau} \boldsymbol{h}_\tau; \quad \alpha_{i\tau} = \frac{\exp(e_{i\tau})}{\sum_{j=1}^{L} \exp(e_{ij})}, \quad e_{i\tau} = \mathcal{F}(\boldsymbol{s}_i, \boldsymbol{h}_\tau)$$

  - $e_{i\tau}$ how well position τ in input predicts position i in output
  - $\boldsymbol{h}_\tau$ is representation (bi-RNN) for the input at position τ
- Current state-of-the-art more complicated see MT lectures
  - iuse transformer-based sequence-to-sequence models

NMT for Grammatical Error Correction [18, 5]

- Developments in GEC following NMT developments
- New forms of sequence to sequence models
  - convolutional sequence-to-sequence models
  - transformer (only attention) sequence-to-sequence models
- Data augmentation using e.g. back-translation approaches
- OOV terms handled using
  - publicly available spelling-correction approaches
  - byte pair encoding (BPE) to handle rare words
- Re-ordering of the N-best output from NMT GEC often used
  - use (correct text) language models to select "best" output

Language Model-Based GEC [1]

- Translation based approaches rely on labelled corpora
  - need large quantities of data for NMT
  - not true for most languages for grammatical error tasks
- BUT "native" data is available for many languages
  - build GEC-based systems using language models
- Rely on access to confusion set data, e.g. for English
  - morphology: Automatically Generated Inflection Database AGID example entry:

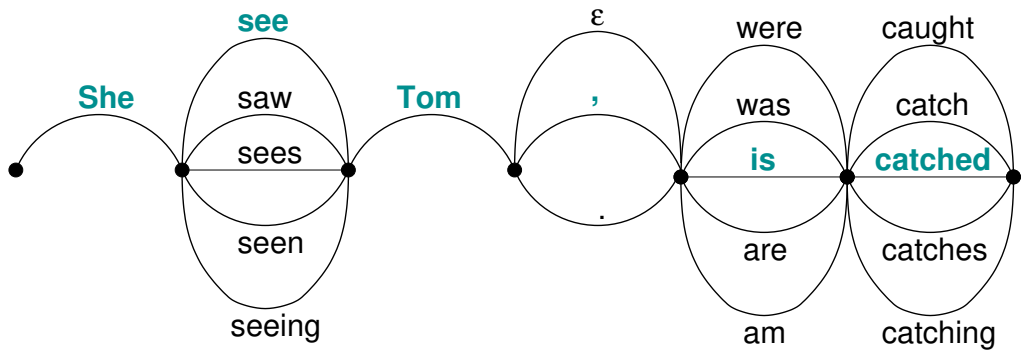
    `see V: saw | seen | seeing | sees`
  - additionally manually generate confusions: {ϵ, a, an, the}

Language Model GEC - example

- Consider first few words of example

She	see	Tom	,	is	catched	...
She	saw	Tom	—		caught	...

- Generate lattice from confusion sets:



  - original source words marked in bold
  - penalty applied to changing word from source
    penalty size controls number of words "corrected"
  - able to apply any form of LM - currently using an RNNLM
- Cannot correct all errors (even with ideal LM)
  - must have appropriate confusion set for error
  - challenging to handle insertions/deletions well

Assessment Criteria [4, 10]

- Maximally Matching Edit Sequence (M^2)
  - Levenshtein distance: source (input) to hypothesis (output)
  - yields a series of system edits made by the GEC system
  - maximum overlap system edits and gold-standard edits (ref)
  - final result given in terms of precision and recall
- Generalized Language Understanding Evaluation (GLEU)
  - related to MT N-gram based scoring BLEU
  - (clipped) counts based on (λ either 0 or 0.1)

$$p_n = \frac{\sum_{\mathrm{ng}\in\mathcal{C}_n} \mathrm{cnt}_{\mathcal{R}_n\setminus\mathcal{S}_n}(\mathrm{ng}) - \lambda(\mathrm{cnt}_{\mathcal{S}_n\setminus\mathcal{R}_n}(\mathrm{ng})) + \mathrm{cnt}_{\mathcal{R}_n}(\mathrm{ng})}{\sum_{\mathrm{ng}\in\mathcal{C}_n} \mathrm{cnt}_{\mathcal{C}_n}(\mathrm{ng}) + \sum_{\mathrm{ng}\in\mathcal{R}_n\setminus\mathcal{S}_n} \mathrm{cnt}_{\mathcal{R}_n\setminus\mathcal{S}_n}(\mathrm{ng})}$$

$$\mathtt{GLEU}(\mathcal{C},\mathcal{S},\mathcal{R}) = \mathtt{BP}\exp\left(\sum_{n=1}^{N}\frac{1}{N}\log(p_n)\right);\quad \mathtt{BP} = \begin{cases} 1 & \text{if } |\mathcal{C}| > |\mathcal{R}| \\ \exp(1-|\mathcal{C}|/|\mathcal{R}|) & \text{otherwise} \end{cases}$$

- $\mathcal{R}_n$ n-grams in reference set $\mathcal{R}$, similarly $\mathcal{S}_n$ source, $\mathcal{C}_n$ corrections
- $\mathcal{R}_n\setminus\mathcal{S}_n$ n-grams in reference set and not in source
- $\mathrm{cnt}_{\mathcal{S}}(\mathrm{ng})$: counts number occurrences of n-gram ng in set $\mathcal{S}$

Spoken "Grammatical Error" Correction [6]

- No agreement as to whether grammar is even appropriate for spoken language
  - native speakers use non-grammatical constructs
  - native speakers often have repetitions
  - native speakers often have false starts
- Broader definition of feedback from spoken English

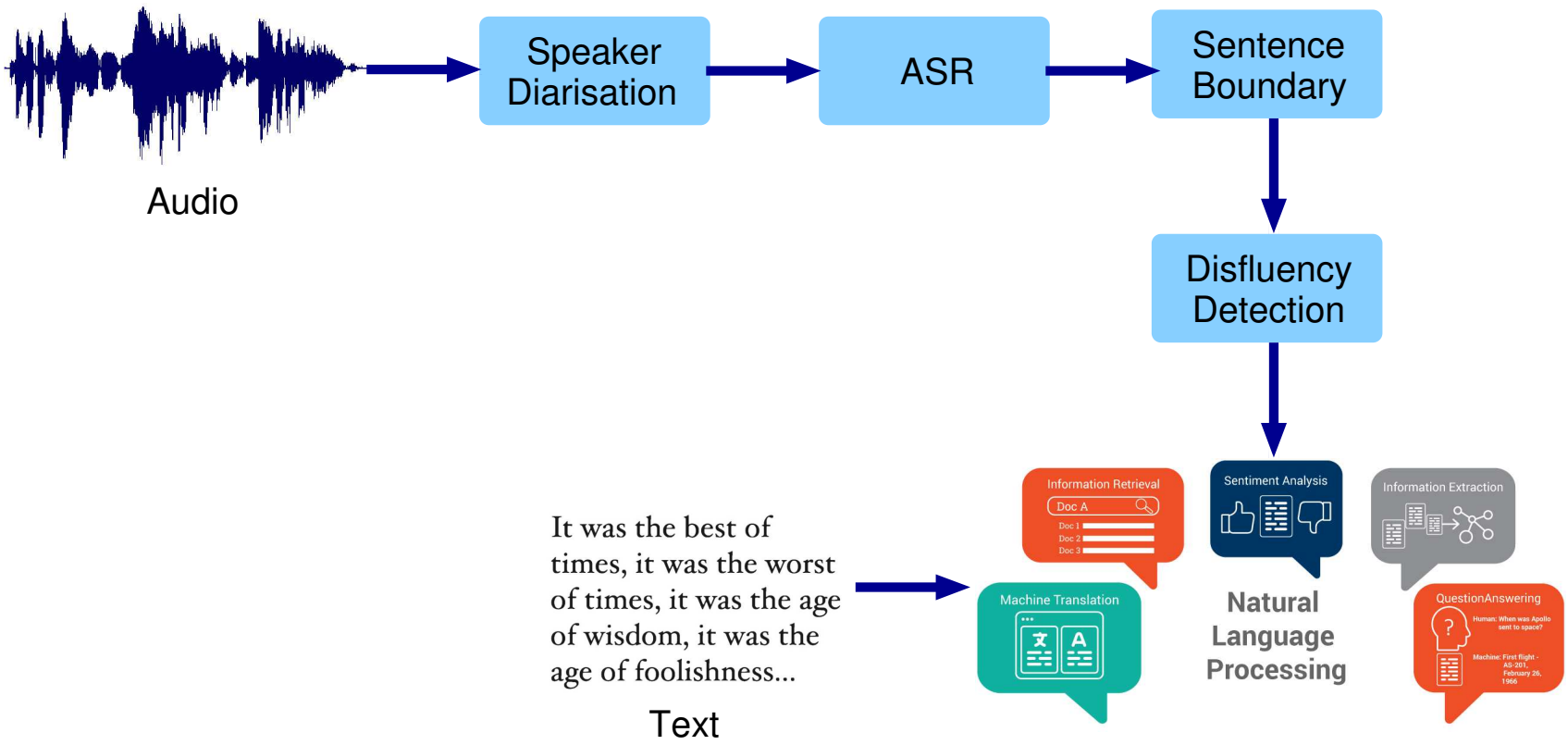
  Feedback that is useful for spoken message construction

  - some overlap with written GEC and GED, but not the same
  - need to also deal with ASR errors

Written vs Spoken GED/GEC

- Advantages: written vs spoken
  - grammar is clearly defined for written
  - no speech recognition errors
  - large(ish) data available (standard text sets)
  - no fillers, dialogue markers, false starts and repetitions
  - sentences normally clearly defined by punctuation
- Disadvantages: written vs spoken
  - spelling mistakes and out-of-vocabulary words
  - audio available (as well as text)
- Currently applying, adapting, written GEC to spoken data
  - ASR errors are a major issue (improved confidence scores)
  - consistency of data annotation a major problem ...

Spoken Language Processing Pipeline



- Convert the audio into form closer to written text
  - incorporate sentence boundary detection
  - incorporate disfluency detection removes:
    - repetitions, false-starts, hesitations, dialogue markers

Spoken "Grammatical Erropr Detection" [8]

System	Written	Spoken	
	CLC	MAN	ASR
Baseline	65.33	48.95	29.82
+ Disfluency Det.	64.13	51.36	31.21
+ Fine-Tune	—	53.45	—

- GEC NMT system trained on written text from CLC corpus
- GLEU Performance on Written (CLC) Spoken (Linguaskill)
  - manual (MAN) and ASR (19.5% WER) transcripts used for spoken data
  - N-fold cross-validation used for fine-tuning experiments

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 stages from ASR to GEC to feedback make use of approaches
- Annotated data is currently limited for many areas
  - challenging to annotate (and agree) spoken learner data

[1] C. Bryant and T. Briscoe, "Language model based grammatical error correction without annotated training data," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2018, pp. 247–253.

[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] D. Dahlmeier and H. T. Ng, "Better evaluation for grammatical error correction," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 568–572.

[5] 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.

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

[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, M. J. F. Gales, K. M. Knill, P. P. Manakul, L. Wang, and Y. Wang, "Impact of ASR performance on spoken grammatical error detection," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1876–1880. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1706

[9] 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.

[10] C. Napoles, K. Sakaguchi, M. Post, and J. Tetreault, "Ground truth for grammatical error correction metrics," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2015, pp. 588–593.

[11] V. Raina, M. J. F. Gales, and K. Knill, "Complementary systems for off-topic spoken response detection," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2020, Online, July 10, 2020*, J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, Eds. Association for Computational Linguistics, 2020, pp. 41–51. [Online]. Available: https://doi.org/10.18653/v1/2020.bea-1.4

[12] V. Raina, M. J. F. Gales, and K. M. Knill, "Universal adversarial attacks on spoken language assessment systems," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3855–3859. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-1890

[13] 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.

[14] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 4709–4713. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178864

[15] 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

[16] 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.

[17] X. Wu, K. M. Knill, M. J. F. Gales, and A. Malinin, "Ensemble approaches for uncertainty in spoken language assessment," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication*

Association, Virtual Event, Shanghai, China, 25-29 October 2020, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3860–3864. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2238

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