UNIVERSITY OF
CAMBRIDGE

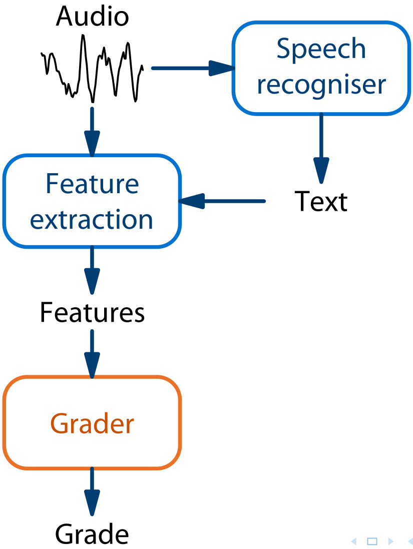
Off-topic spoken response detection for language assessment

Andrey Malinin

25 May 2016

Department of Engineering

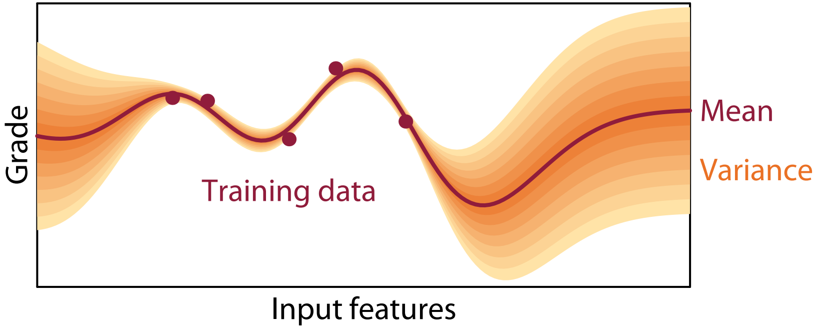
Automatic assessment architecture



Grader features

- Mainly fluency based
- 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).

Gaussian process grader



- GP performance - 0.83 correlation with human experts
- Variance is confidence in prediction → detect outliers
- Interpolating with standard examiners - 0.88 correlation
    - Humans can do contents evaluation.

Current system

- Can accurately assign grades to a speaker
- More consistent and faster than humans
- Can use confidence to detect difficult speakers.
- Cannot detect responses inappropriate to the question
  - Speaker failed to construct coherent response
  - Speaker failed to understand the question
  - Speaker gave a memorized response.

Assessing response relevance

- Q1: "Tell us about how you like to schedule business meetings and conferences?"

Assessing response relevance

- Q1: "Tell us about how you like to schedule business meetings and conferences?"
- R1: "I like to schedule my meetings in the morning, often using video conferences to be able to flexibly adjust for the busy schedules of my partners."

Assessing response relevance

- Q1: "Tell us about how you like to schedule business meetings and conferences?"
- R1: "I like to schedule my meetings in the morning, often using video conferences to be able to flexibly adjust for the busy schedules of my partners."
- R2: "When I travel, I like to buy my tickets and check the schedule ahead of time, so that I do not have to rush to the airport or train station."

Assessing response relevance

- Q1: "Tell us about how you like to schedule business meetings and conferences?"
- R1: "I like to schedule my meetings in the morning, often using video conferences to be able to flexibly adjust for the busy schedules of my partners."
- R2: "When I travel, I like to buy my tickets and check the schedule ahead of time, so that I do not have to rush to the airport or train station."

Assessing response relevance

Two scenarios:

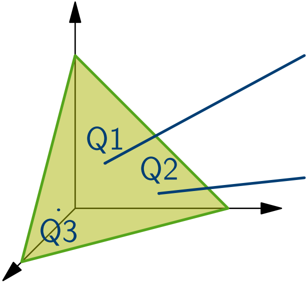
- Assess based on similarity to question prompts.
  - Faster rollout
  - Limited data → "Describe this picture".
  - Augment based on web data (Ex: UKWAC) → In Progress

Assessing response relevance

Two scenarios:

- Assess based on similarity to question prompts.
  - Faster rollout
  - Limited data → "Describe this picture".
  - Augment based on web data (Ex: UKWAC) → In Progress
- Assess based on similarity to example responses.
  - Richer representation of question topic.
  - Need to collect example responses

Example-response based topic space construction

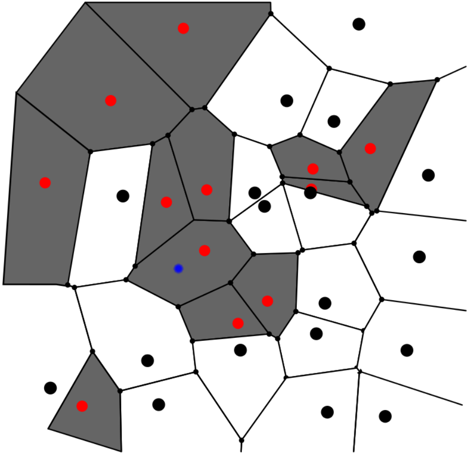


Example Responses to “How do you feel about working on Sundays?”

Example Responses to “Describe this chart.”

- Text → vector space representation
- Commonly used
- Bag-of-words

Nearest neighbour topic classification



Assign class based on class of nearest neighbour.

Summary of standard approach

Standard Approach:

1. Construct topic space using example responses.
2. Classify using a (K) Nearest Neighbour classifier

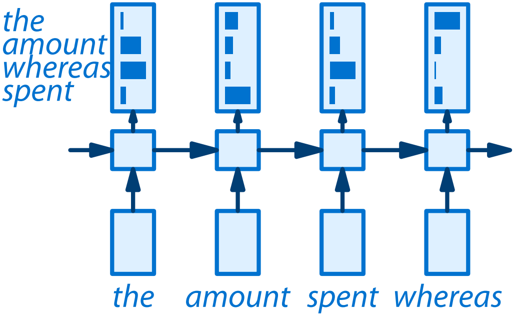
Limitations:

- Doesn't scale with training data size.
- Sequence information is not modelled.

Solution → Deep Learning

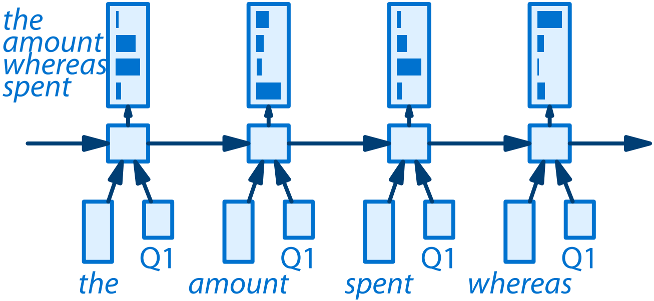
Recurrent neural network language model

- Predicts probability of next word given previous context:
  $\mathrm{P}(w_i|w_{i-1}, \cdots, w_0) \rightarrow \mathrm{P}(\textit{whereas}|\textit{spent},\ \textit{amount},\ \textit{the},\ \ldots)$
- Assign probability to entire sequence:
  $\mathrm{P}(\mathbf{w}) = \prod_1^i \mathrm{P}(w_i|w_{i-1}, \cdots, w_0)$



Topic adapted recurrent neural network LM

- Computes probability of word sequence given the topic: $\mathtt{P}(\mathbf{w}|Q_{\mathbf{w}})$
- Assign topic $\hat{Q}_{\mathbf{w}}$ which maximizes $\mathtt{P}(\mathbf{w}|Q_{\mathbf{w}})$
- Scales with number of topics.

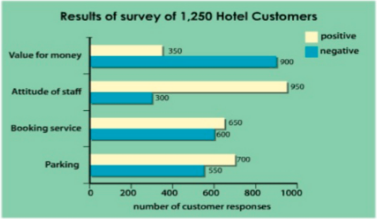


Business Language Testing Service (BULATS) Spoken Tests

A. Introductory Questions: 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

Experimental setup

Training Data - BLXXXtrn03+04 + BLXXXuns00

- 214 question topics.
- 10.5K crowd-sourced responses. 50 per topic.
- + 210K ASR transcribed at 30% WER. 1000 per topic.
- Gujarati L1

Evaluation Data - BLXXXeval1

- 1560 responses
- ASR transcribed at 30% WER.
- Gujarati L1

Topic classification experiment

- Goal: Assess performance in response topic classification.

System	Trn. data (resp)	% Error
KNN	10.5K	20.8
RNNLM-1		17.5
KNN-2	220.5K	-
RNNLM-2		9.3

- Increasing quantity improves RNNLM performance.
- KNN is impractical for large datasets.

Example error

- RES: I am working as a colour teacher in city I am my college time from nine pm nine am to the five pm as work as hard work and I all my colleges we have attend for the lectures and also we also give the all types of activity in our college there are celebrations are going on also we participate in our celebrations I also like to work with my colleagues there are we work in the group and we are the best place and our president and decide to make we have going to making out the work place teacher then we are make a work place teachers so we are useful to our society and the student.

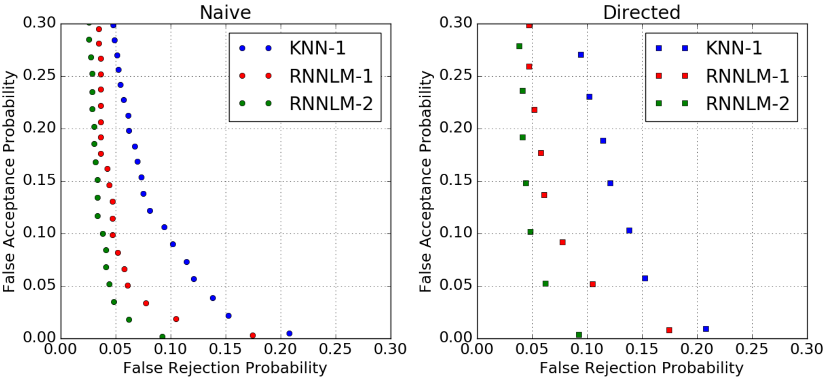
Example error

- RES: I am working as a colour teacher in city I am my college time from nine pm nine am to the five pm as work as hard work and I all my colleges we have attend for the lectures and also we also give the all types of activity in our college there are celebrations are going on also we participate in our celebrations I also like to work with my colleagues there are we work in the group and we are the best place and our president and decide to make we have going to making out the work place teacher then we are make a work place teachers so we are useful to our society and the student.
- Q1: Talk about a successful day you had at work. You should say what work you did on that day; how you dealt with any difficulties; why the day was successful.
- Q2: Talk about a colleague you like working with. You should say how long you have worked with this colleague; what work you do together; why you like working with him/her.

Off-topic response detection

- Data only contains valid responses. →
  - Randomly select responses to other questions
- How to select off-topic responses to a question?
  - Naive: responses from any BULATS section
  - Directed: responses only from the same BULATS section.
- Construct ROC curve → false rejection vs. false acceptance:
  - Accept response if in top - N results.
  - Very operating point with the value of N.

Off-topic response detection results



Off-topic response detection results

% Equal Error Rate			
System	Trn. data (resp)	Naive	Directed
KNN	10.5K	10.0	13.0
RNNLM-1	10.5K	6.0	7.5
RNNLM-2	220.5K	4.5	6.0

Summary of RNNLM system

- Outperforms standard approaches
- Is practical for large training datasets
- Architecture is extensible
- Benefits from work on RNN language models.

Conclusions

- Can accurately and consistently grade a speaker
- Can use confidence to detect difficult speakers.
- Can robustly detect inappropriate responses.

Acknowledgements

- Kate Knill, Rogier Van Dalen, Yu Wang and Mark Gales
- Ronan Cummins for providing Query Retrieval System and UKWAC corpus.
- This work will be presented at ACL 2016 in Berlin.

Questions?