Foundation Models in Spoken Language Processing: Time to go home or make hay?
Observations from automated language learning and assessment

Kate Knill

ALTA Institute, Cambridge University Engineering Department

14 June 2023
What about "your manager has asked you for the result of an internal staff survey on travelling to work the first chart shows how staff at your company travelled to work the second chart shows what improvements to company facilities the staff would like to see look at the chart and then talk about the information describing the survey results um by car forty percent by train thirty five thirty five percent by bus ten percent on foot five percent bicycle five percent motor bicycle motorbike five percent free company bus to and from station very many"
Here is a corrected version of the sentence:

"Your manager has asked you for the result of an internal staff survey on travelling to work. The first chart shows how staff at your company travelled to work, and the second chart shows what improvements to company facilities the staff would like to see. Look at the chart and then talk about the information, describing the survey results: by car, 40%; by train, 35%; by bus, 10%; on foot, 5%; by bicycle, 5%; by motorbike, 5%; and a free company bus to and from the station is desired by many."

In this sentence, I made a few changes to improve the grammar and clarity. First, I added the missing article "the" before "first chart" and "second chart". I also added the article "a" before "free company bus" to make it clear that there is only one bus. I also changed the word "very" to "many" to better match the verb "desired". Finally, I added a few commas to make the sentence easier to read. Let me know if you have any other questions.
Time to go home?!?

Or make hay while the sun shines?

Ash Thomas, Tiverton, Creative Commons Attribution 2.0 Generic license
Talk Outline

• Foundation Models
  • What they are
  • Predictive and Generative AI models
• Applications in automated language learning and assessment
  • Neural Text and Speech Representation-based Auto-marking
  • Grammatical Error Correction for Feedback and Assessment
  • Multiple Choice Reading Comprehension: is the model doing what we want it to?
• Conclusions
A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.
Foundation Model: Application Process

LANGUAGE MODEL SIZES TO MAR/2023

- BERT 340M
- GPT-1 117M
- GPT-2 1.5B
- GPT-3 175B
- Jurassic-1 178B
- MT-NLG 530B
- PaLM
  - PaLM-Coder
  - Minerva
  - Med-PaLM
  - Flan-PaLM
  - U-PaLM
  - Flan-U-PaLM
  - Med-PaLM 2
  - 540B
- Luminous 200B
- BLOOM
  - BLOOMZ 176B
- Gopher 280B
- Flamingo 80B*
- Chinchilla 70B*
- LaMDA
  - LaMDA 2
  - Bard 137B
- Anthropic-LM
  - Claude 52B
  - 9.4B
  - RL-CAI 6B
- GPT-J
  - 6B
- GPT-NeoX-20B
- CM3
  - 13B
- Atlas
  - 11B
- Kosmos-1
  - 1.6B*
- Atlas
  - 11B
- Flan-T5
  - 11B
- NLLB
  - ChatGLM-6B 54.5B
- OPT-175B
  - BB3 175B
- MOSS
  - 20B*
- LLaMA
  - Undisclosed 65B*
- YaLM
  - 100B
- Galactica
  - 120B
- GLM-130B
  - ChatGLM-6B
- WeLM
  - 10B
- AlexaTM
  - VIMA 200M
- PaLI
  - 710B*
- Z-Code++
  - 7.7B
- SeeKeR
  - 2.7B
- Gato
  - 1.2B
- FIM
  - 9.8B
- NooR
  - 10B
- Toolformer
  - 75B
- Alpaca
  - Undisclosed
- Nemo
  - 7B
- Cog
  - 7B
- GPT-4
  - Undisclosed
- Megatron-11B
  - 11B
- T5
  - 11B
- Plato-XL
  - 11B
- Macaw
  - 11B
- Cohere
  - 52.4B
- ruGPT-3
  - 20B
- XGLM
  - 7.5B
- Cedille
  - 6B
- Fairseq
  - 13B
- BlenderBot2.0
  - 13B
- BlenderBot
  - 6B
- Fina
  - 20B
- FIM
  - 9.8B
- VIMA
  - 200M
- AlexaTM
  - VIMA 200M
- LifeArchitect.ai/models

What Are Foundation Models?

• **Predictive AI:** systems that make “decisions”
  • foundation models used as key component
  • e.g. wav2vec2.0, BERT, ELECTRA etc etc

• **Generative AI:** systems that generate “data”
  • foundation models can be used in a “zero-shot” fashion
  • e.g. ChatGPT, BARD, DALL-E etc etc
Interesting aspects of (some) foundation models: Homogenization

• Same model can be applied over a wide-range of tasks
  • Spoken Language Processing tasks we’ve tried using ChatGPT (*)
    • Speech recognition output correction
    • Prompt generation (pronunciation/stress) for synthesis
    • Text processing/tidying
    • Grammatical error correction
    • Multiple choice question generation / answering
    • Hallucination detection
    • Triple extraction for knowledge representation
    • …

* Other Generative AI models are available
Interesting aspects of (some) foundation models: Emergence

- Behaviour implicitly induced rather than explicitly trained
  - Prompt engineering and in-context learning

<table>
<thead>
<tr>
<th>Zero-shot</th>
<th>One-shot</th>
<th>Few-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translate English to French:</td>
<td>Translate English to French:</td>
<td>Translate English to French:</td>
</tr>
<tr>
<td>cheese =&gt;</td>
<td>sea otter =&gt; la loutre de mer</td>
<td>sea otter =&gt; la loutre de mer</td>
</tr>
<tr>
<td></td>
<td>cheese =&gt;</td>
<td>raspberries =&gt; les framboises</td>
</tr>
<tr>
<td></td>
<td></td>
<td>red man =&gt; l’homme rouge</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cheese =&gt;</td>
</tr>
</tbody>
</table>
• BERT: BiDirectional Encoder Representations from Transformers

  Pre-trained on English Wikipedia (2500M words) and the Toronto BookCorpus (800M words)
  • Around 110M trainable parameters
Predictive AI: Masked LLMs for Speech Input

Generative AI

- Essentially autoregressive language models trained on lots of data e.g. GPT-3

<table>
<thead>
<tr>
<th>Dataset</th>
<th># tokens</th>
<th>Proportion within training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Crawl</td>
<td>410 billion</td>
<td>60%</td>
</tr>
<tr>
<td>WebText2</td>
<td>19 billion</td>
<td>22%</td>
</tr>
<tr>
<td>Books1</td>
<td>12 billion</td>
<td>8%</td>
</tr>
<tr>
<td>Books2</td>
<td>55 billion</td>
<td>8%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3 billion</td>
<td>3%</td>
</tr>
</tbody>
</table>
InstructGPT: human-in-the-loop training

**Step 1**
Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

**Step 2**
Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

**Step 3**
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

---

Conversational Generative AI

- Generative AI has evolved to support conversations: ChatGPT, BARD, LLaMA, ErnieBot …
  - e.g. can answer followup questions, note own mistakes, challenge premise of discussion

- ChatGPT difference to InstructGPT: dialogue format in training
  - Step 1: Use human AI trainers to provide ‘conversations’ between a user and an AI assistant
  - Step 2: Reward model consists of two or more conversation model responses ranked by quality
    - Data added from conversations that AI trainers had with the chatbot

- GPT-4
  - Multimodal input: images as well as text
  - “System message”: specify tone and task e.g. “to be a 16th century pirate”, “write response in JSON”
2023-2024 OPTIMAL LANGUAGE MODELS

GPT-4
1T
20:1

PaLM 2
340B
11:1

Cerebras-GPT
13B
20:1

StableLM
65B
24:1

MPT
7B
143:1

LLaMA
65B
22:1

Falcon
40B
25:1

InternLM
104B
16:1

MOSS
208B
27:1

Next...

Google DeepMind Gemini
Anthropic Claude-Next
OpenAI GPT-5
Other stealth projects
2T
20:1

LifeArchitect.ai/models
Talk Outline

• Foundation Models
  • What they are
  • Predictive and Generative AI models

• Applications in automated language learning and assessment
  • Neural Text and Speech Representation-based Auto-marking
  • Grammatical Error Correction for Feedback and Assessment
  • Multiple Choice Reading Comprehension: is the model doing what we want it to?

• Conclusions
Automated Learning and Assessment for L2 English
L2 learner speech data is challenging!

No punctuation/sentences

ASR Errors

Hesitations

Information encoded in how we speak not just what we say

Disfluencies
Spoken Language Assessment and Feedback Pipeline

Free speaking prompt-response tests
– speak for up to 1 minute

Mark over 1 or several responses

Speech Processing

Grader

Score: 3.5
Conf: 90%

Feedback

Analytic – holistic feedback across all speech
Fine-grained – feedback on specific errors in words/phrases
Construct: assess core speaking skills

- Pronunciation
- Hesitation / Extent
- Coherence / Discourse Management
- Language Resource
- Task Achievement
Automatic Spoken Language Assessment
Feature-based Auto-marking System

- Very effective with good construct coverage
  - Features selected to model different assessment aspects
  - Deployed in range of low-medium stakes tests and practice tests
- Limitations
  - Many features hand-crafted so may not be optimal
  - Difficult to know what are best features for new auto-marking scenarios e.g. conversational assessment
Applying Foundation Models to Auto-marking: Neural Text Grader

- BERT word embeddings form input features to grader
  - Train LSTM with attention to regression head grader on in-domain data
  - Applicable to both monologic and dialogic (conversational) tests
- Limitations
  - Limited ability to assess all aspects of the construct: pronunciation, fluency
  - Less information on ‘why’ auto-marker predicted a particular score
Applying Foundation Models to Auto-marking: Neural Speech Grader

- Wav2vec2.0 speech representations form input features to grader
  - Trained mean pooling (monologic tests) or attention (dialogic tests) models to regression head grader
  - Applicable to both monologic and dialogic tests
- Limitations
  - Limited ability to assess all aspects of the construct: language resource, coherence/discourse
  - Less information on ‘why’ auto-marker predicted a particular score

S. McKnight et al, “Automatic Assessment of Conversational Speaking Tests”, UK Speech C
Auto-marking performance comparison: monologic free-speaking test

<table>
<thead>
<tr>
<th>Grader</th>
<th>↑ PCC</th>
<th>↓ RMSE</th>
<th>% ≤ 0.5</th>
<th>% ≤ 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>0.932</td>
<td>0.382</td>
<td>82.3</td>
<td>98.7</td>
</tr>
<tr>
<td>Text</td>
<td>0.930</td>
<td>0.393</td>
<td>80.3</td>
<td>98.6</td>
</tr>
<tr>
<td>Speech</td>
<td>0.933</td>
<td>0.393</td>
<td>79.7</td>
<td>99.0</td>
</tr>
<tr>
<td>Std ⊕ Text ⊕ Speech</td>
<td>0.943</td>
<td>0.356</td>
<td>85.0</td>
<td>99.1</td>
</tr>
</tbody>
</table>

- Neural auto-markers have similar overall level of performance to standard grader
  - Wav2vec2.0 currently inconsistent across different parts of the test
- Complementary models – ensemble of 3 graders yields best results
  - See posters by Stefano Bannò and Simon McKnight in poster session C for more details
Spoken Grammatical Error Correction
Grammatical Error Correction (GEC)

- Aim of GEC is to produce grammatically correct sentence
  
  Original: The dog *eated* from the bowl.
  Corrected: The dog *ate* from the bowl.

- Speech adds additional challenge

  Spoken Original: the dog *ea-* eated from um the bowl
  Corrected: the dog *ate* from the bowl
### Spoken GEC – End2end?

The diagram shows the process of removing disfluency and grammar errors from non-native disfluent speech to produce a grammatically correct fluent text. The text explains that end-to-end (E2E) is not feasible currently due to:

- **No paired training data**
- **Hard to give feedback to learners**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Audio</th>
<th>Text</th>
<th>DSF</th>
<th>GEC</th>
<th>L2?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR-Train¹</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Switchboard²</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CLC³ + BEA⁴</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**E2E not feasible (currently)**

1. Y. Lu et al, "Impact of ASR Performance on Spoken Grammatical Error Detection", INTERSPEECH 2019
Spoken GEC – Cascade pipeline

- Train/fine-tune each module on different data sets
- Various sources of information for feedback
Spoken GEC – Cascade pipeline challenges

- Error propagation
- Domain mismatch
- Prosody information loss

Non-native disfluent speech

ASR → DSF → GEC → Grammatically correct fluent text

the cat sit on the um mat → the cat sit on the mat → the cat sat on the mat

ASR → DSF → GEC

the cat see um the um mat → the cat see the mat → the cat saw the mat

GEC

lemme in → <unk> in

GEC

i feel well pretty bad → i feel well pretty bad

DD
Gramformer seq2seq Spoken GEC

- Treat GEC as a machine translation problem
  - Transformer-based encoder-decoder models
    - Large pre-trained e.g. T5 ‘Text-to-Text Transfer Transformer’

- **Gramformer**: T5-based fine-tuned for GEC on various corpora
  - Can further fine-tune to target domain
  - 223M parameters

---

Spoken GEC Performance

• Manual, Fluent speech transcripts → Gramformer

<table>
<thead>
<tr>
<th></th>
<th>Writing</th>
<th>Speaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M^2 F_{0.5}$↑</td>
<td>56.6</td>
<td>53.6</td>
</tr>
</tbody>
</table>

- Spoken GEC can match performance of Written GEC L2 English learner task

• Cascade GEC system: TDNN-F ASR → BERT DSF → Gramformer

<table>
<thead>
<tr>
<th>Input to Gramformer</th>
<th>Speaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>ASR+DD+GEC</td>
</tr>
<tr>
<td>SER ↓</td>
<td>43.3</td>
</tr>
<tr>
<td>TER ↓</td>
<td>8.3</td>
</tr>
</tbody>
</table>

- Need to mitigate for ASR error propagation and domain mismatch

ASR Errors An Issue for ChatGPT Too

Last one (for now ;-)?): "I think my stall language skill is have fast range rate and experiences for quality control as we working for more than ten years"

Here is a corrected version of the sentence:

"I think my language skills have a fast rate of improvement, and I have a lot of experience in quality control. I have been working in this field for more than ten years."

- ASR error: “I think my strongest skill…” → “I think my stall language skill …”
Are models doing what we want them to?


[2] Badgeley et Al. 2018
Context: My friends like different clothes. Sue likes red clothes. She is often in a red skirt and red shoes. Mina likes white clothes. She is in a white shirt. Her sister Emma likes to wear a green skirt. She looks nice. David often wears a white cap and black pants. Peter often wears a white coat and black pants.

Question: Mina’s sister’s name is ___

Options:  
A) Sue  
B) Emma  
C) Jenny  
D) David  

*real example
Is Reading Comprehension important?

**Question:** Mina’s sister’s name is ___.

Options:
- A) Sue
- B) Emma
- C) Jenny
- D) David

**Question:** The word jolting in line 5 is closest in meaning to

Options:
- A) Predicted
- B) Shocking
- C) Unknown
- D) Illuminating

**Question:** Harry is ___ years older than Yue.

Options:
- A) 11
- B) 12
- C) 13
- D) 14
Mina's sister's name is _

My friends like different...

A) Sue  
B) Emma  
C) Jenny  
D) David

predicted answer (A, B, C, or D)
Defective Input Performance

- S. McKnight et al, "Automatic Assessment of Conversational Speaking Tests", UK Speech C

<table>
<thead>
<tr>
<th>Training data</th>
<th>M</th>
<th>H</th>
<th>C</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>25.00</td>
<td>25.00</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td>Q+{O}+C</td>
<td>88.09</td>
<td>84.42</td>
<td>81.64</td>
<td>85.01</td>
</tr>
<tr>
<td>Q+{O}</td>
<td>54.81</td>
<td>57.75</td>
<td>60.31</td>
<td>57.32</td>
</tr>
</tbody>
</table>

- RACE++ data set
- Systems can achieve reasonably high performance without performing comprehension
Effective Number of Options

\[ \mathcal{H}(Y) = 0.01 \]

\[ 2^{\mathcal{H}(Y)} = 1.01 \]

\[ \mathcal{H}(Y) = 1.60 \]

\[ 2^{\mathcal{H}(Y)} = 3.04 \]

\[ 1 \leq 2^{\mathcal{H}(Y|Q,O)} \leq \text{#options} \]
What Are We Assessing?

- Systems can achieve reasonably high performance without performing comprehension.
- ‘Shortcut’ systems can confidently:
  - determine some correct answer options
  - eliminate some unlikely distractors
  - use general knowledge to gain information
- Can exploit this in content creation to flag questions that don’t need comprehension to answer.
Conclusions

• Foundation Models: predictive and generative AI
  • Pre-training on large quantities of semi-supervised data at scale enables
    • Homogeneity: same model useful for many different downstream tasks
    • Emergence: zero-shot learning required to reach good performance on many tasks

• Range of uses in downstream tasks even when in-domain data is limited
  • Examples in Automated Spoken Language Assessment and Learning:
    • Auto-marking, Spoken Grammatical Error Correction, Multiple Choice Reading Comprehension …

• The field of Foundation Models is changing rapidly – definitely worth sticking around for
Thanks to the ALTA Spoken Language Processing Technology Project Team

Prof Mark Gales
Stefano Bannò
Yassir Fathullah
Adian Liusie
Yiting “Edie” Lu
Charlie McGhee
Simon McKnight
Rao Ma
Potsawee Manakul
Mengjie Qian
Vatsal Raina
Vyas Raina
A Acoustic-to-Articulatory Inversion for Pronunciation Feedback
Charles McGhee, Mark Gales, Kate Knill

B N-best T5: Robust ASR Error Correction using Multiple Input Hypotheses and Constrained Decoding Space
Rao Ma, Mark Gales, Kate Knill, Mengjie Qian

C Adapting an Unadaptable ASR System
Mengjie Qian*, Rao Ma*, Mark Gales, Kate Knill

C Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning
Stefano Bannò (FBK), Kate Knill, Marco Matassoni (FBK), Vyas Raina, Mark Gales

C Automatic Assessment of Conversational Speaking Tests
Simon McKnight, Arda Civelekoglu, Mark Gales, Kate Knill
Questions?

Thanks to:

Diane Nicholls and the Humannotator team at ELiT for the Linguaskill Speaking annotations.

This presentation reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

Project website: [http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html](http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html)

Practice your English speaking for free with Speak & Improve

Contact: kmk1001@cam.ac.uk