Towards automated detection of fluent aphasias: A classification and interpretable natural language processing study

Frank Tsiwah and Ruhi Umesh Mahadeshwar

Aphasia classification

The classical taxonomic approaches classifies **fluent aphasias** into the subtypes Wernicke’s, anomic, transcortical sensory, and conduction aphasia syndromes.

Aphasia classification

The classical taxonomic approaches classifies **fluent aphasias** into the subtypes Wernicke’s, anomic, transcortical sensory, and conduction aphasia syndromes

- ❑ Test batteries (e.g., WAB, BDAE etc)
- ❑ Spontaneous speech analyses

Aphasia classification challenges

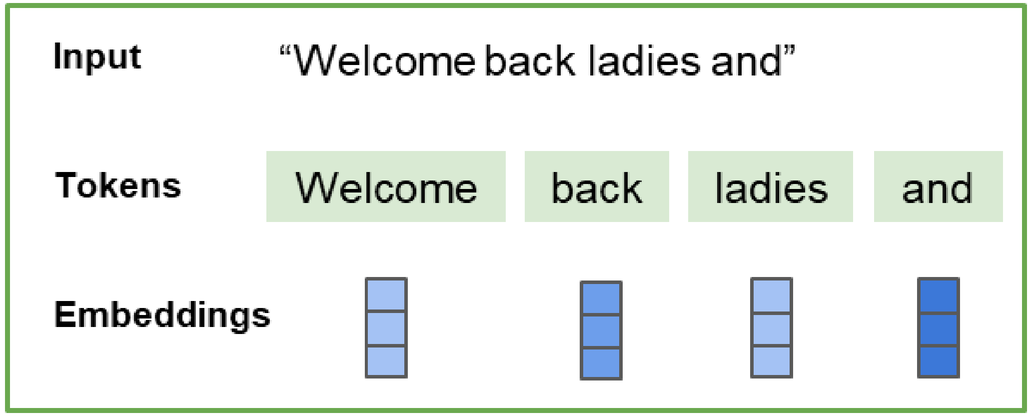
To classify individuals with aphasia based on their linguistic profile can be:

- Time consuming
- Resource intensive
- Language samples collected through test batteries often do not reflect natural spoken language

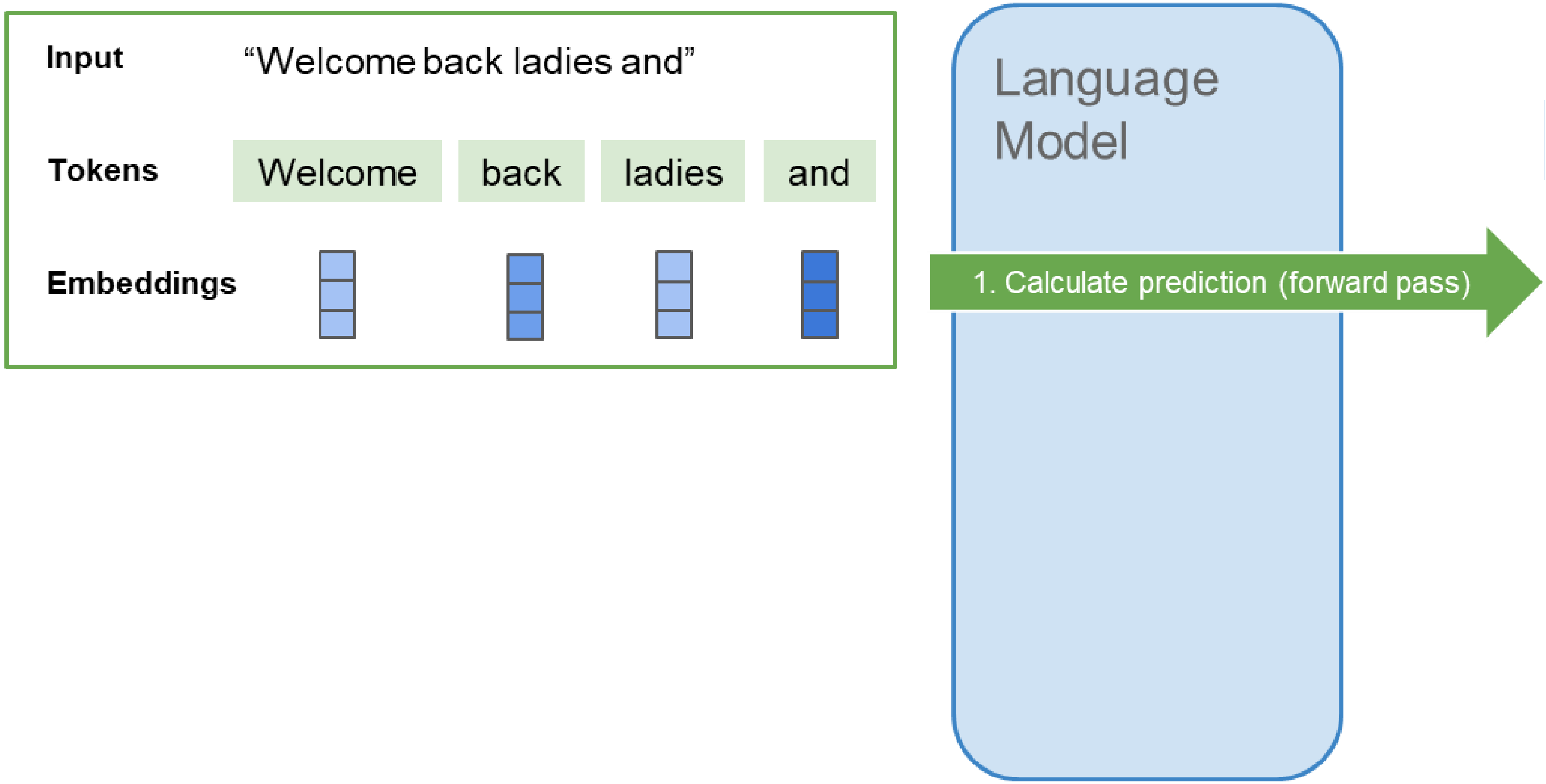
Aim of study

The goal of this project was to examine the use of **large language models** to automatically detect fluent aphasia types using spontaneous speech transcripts

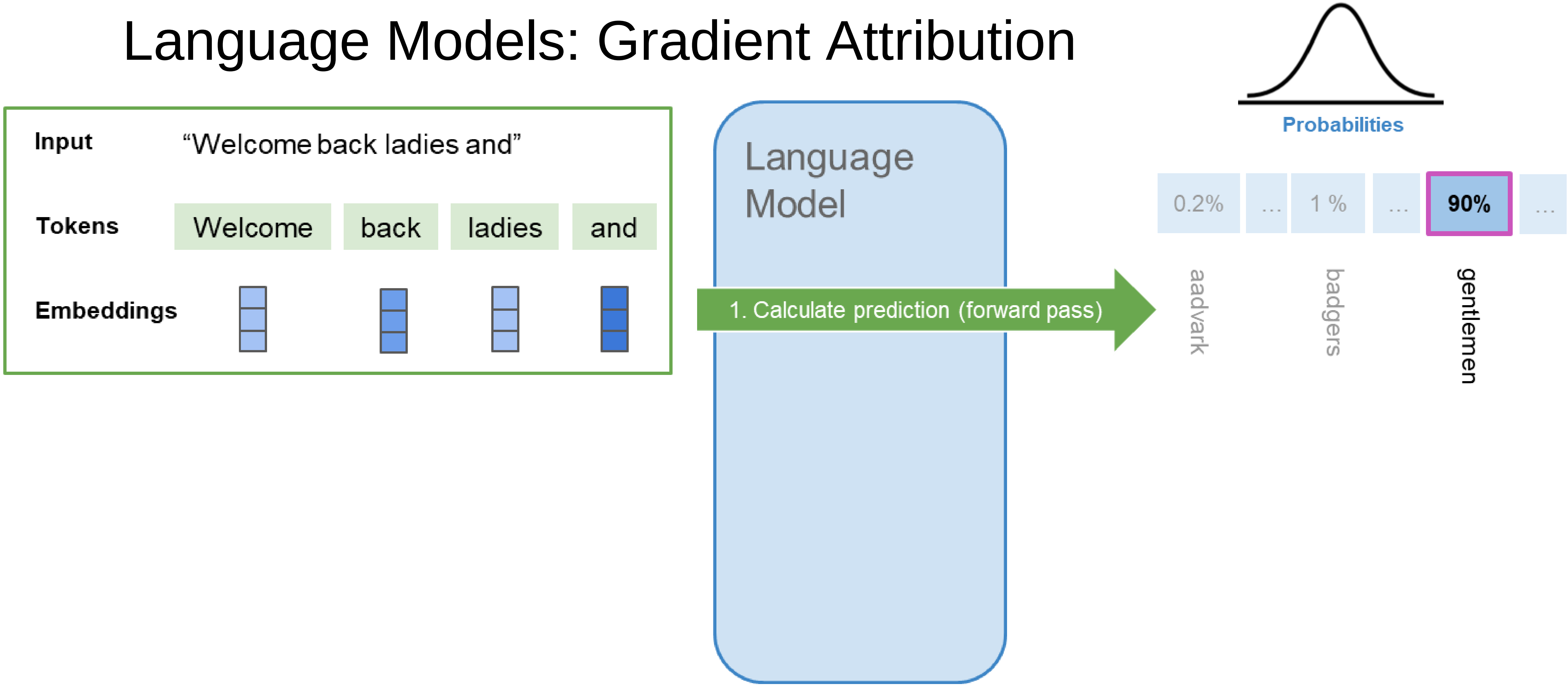
Language Models: Gradient Attribution



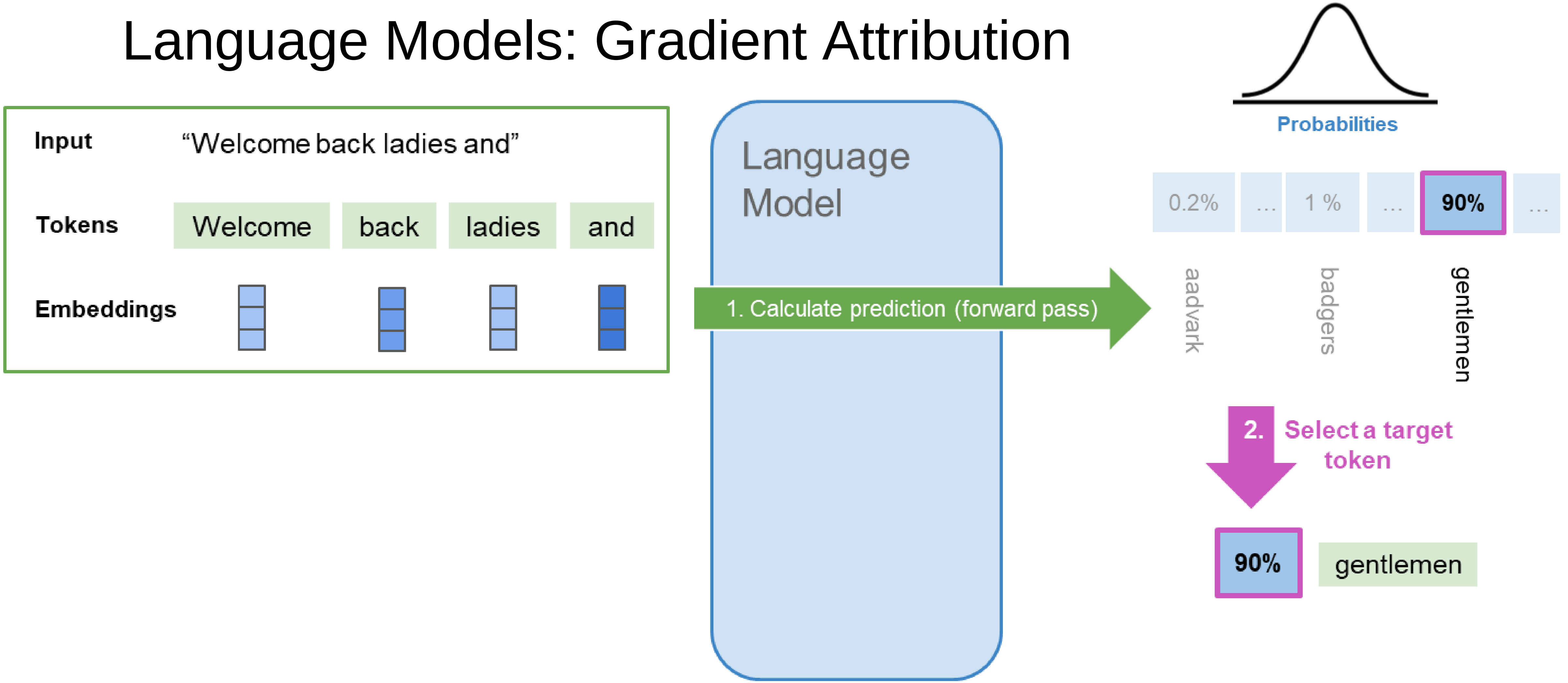
Language Models: Gradient Attribution



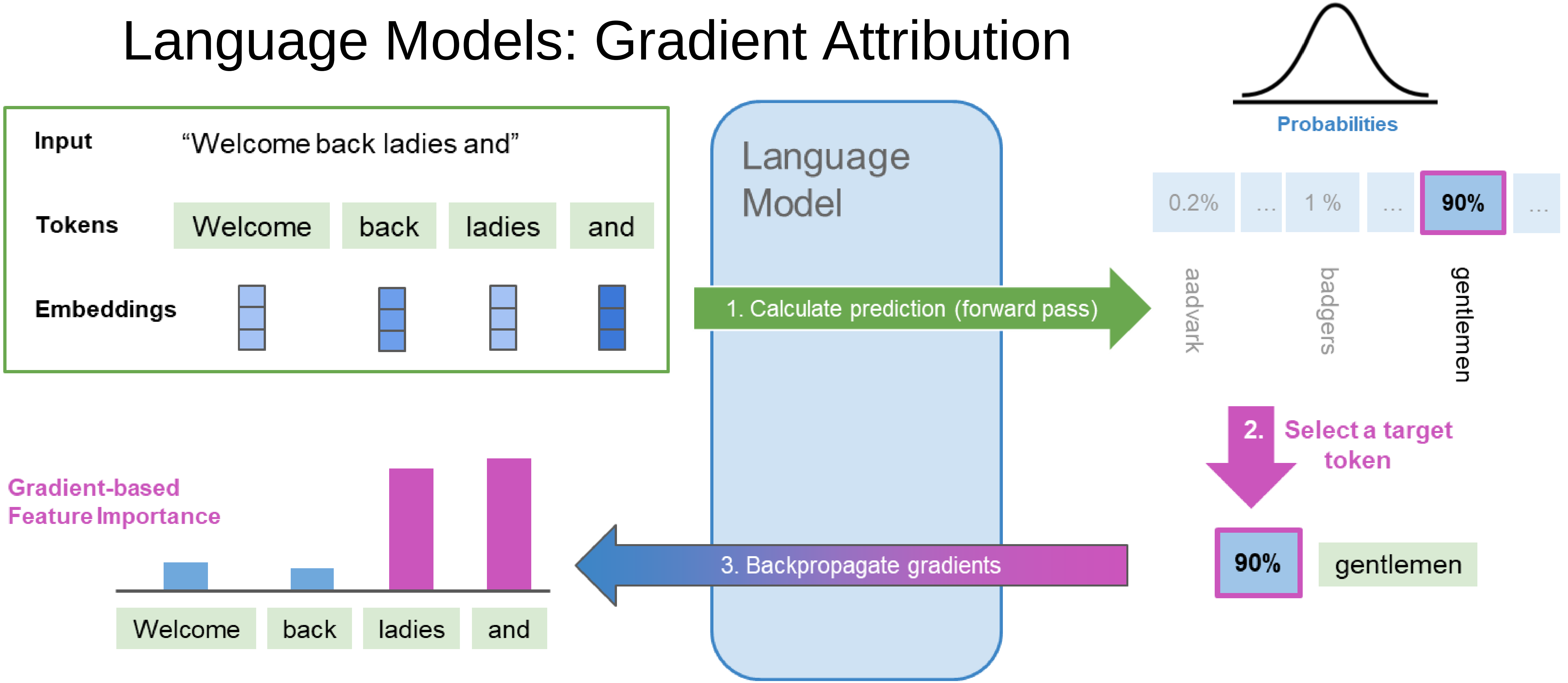
Language Models: Gradient Attribution



Language Models: Gradient Attribution



Language Models: Gradient Attribution



rijksuniversiteit
groningen

Research questions

- ❑ Can we automatically detect fluent aphasias types from healthy controls using spontaneous speech transcripts?

Research questions

- Can we automatically detect fluent aphasias types from healthy controls using spontaneous speech transcripts?
  - How much spontaneous speech sample does a language model require to attain a high accuracy?

Research questions

- Can we automatically detect fluent aphasias types from healthy controls using spontaneous speech transcripts?
  - How much spontaneous speech sample does a language model require to attain a high accuracy?

- Which words or linguistic features are most informative for the model’s predictions in distinguishing fluent aphasia types from healthy controls?

Method

Participants and data

- Data from AphasiaBank (MacWhinney et al., 2011)
- Spontaneous speech (SS) transcripts of:
  - 202 Anomic aphasia
  - 47 Wernicke
  - 99 Conduction
  - 267 Healthy Controls (HCs)

rijksuniversiteit groningen

Method

Participants and data

- Data from AphasiaBank (MacWhinney et al., 2011)
- Spontaneous speech (SS) transcripts of:
  - 202 Anomic aphasia
  - 47 Wernicke
  - 99 Conduction
  - 267 Healthy Controls (HCs)

Spontaneous Speech data:

- Picture description: *e.g. cookie theft, cat rescue, broken window*
- Narration: *e.g. Cinderella story*
- Open-ended questions: e.g. *how do you think your speech is these days?*

Method

Example sentences containing pauses and data preprocessing

Raw*: and and &-um it didn't fit of course. so &-um cinderella was able to finally put the shoe on.*

Preprocessed*: and and [FP] it didn't fit of course. so [FP] cinderella was able to finally put the shoe on.*

Method: Models

Pretrained **DistilBERT** (Distilled Bidirectional Encoder Representations from Transformers: Sanh et al., 2019) language model (LM)

Method: Models

Pretrained **DistilBERT** (Distilled Bidirectional Encoder Representations from Transformers: Sanh et al., 2019) language model (LM)

We finetuned using part (70%) of healthy control (HC) and the clinical data, and then tested on the remaining (30%) of the data

Method: Models

Pretrained **DistilBERT** (Distilled Bidirectional Encoder Representations from Transformers: Sanh et al., 2019) language model (LM)

We finetuned using part (70%) of healthy control (HC) and the clinical data, and then tested on the remaining (30%) of the data

Transformer Interpret (Pierse 2021) was used for interpretability

Results: using all data

	Acc.	Precision	Recall	
Anomic vs HC	91%	89%	92%	
Wernicke vs HC	96%	95%	92%	
Conduction vs HC	95%	94%	92%	

Results: using all data

	Acc.	Precision	Recall	
Anomic vs HC	91%	89%	92%	
Wernicke vs HC	96%	95%	92%	
Conduction vs HC	95%	94%	92%	
Anomic, Wernicke, Conduction	55%	60%	47%	

Results: only open-ended questions

	Acc.	Precision	Recall	
Anomic vs HC	93%	90%	94%	
Wernicke vs HC	84%	86%	79%	
Conduction vs HC	92%	93%	91%	
Anomic, Wernicke, Conduction	46%	46%	56%	

Results: only open-ended questions

	Acc. (Prev)	Precision (Prev)	Recall (Prev)	
Anomic vs HC	93% (**91**%)	90% (**89**%)	94% (**92**%)	
Wernicke vs HC	84% (**96**%)	86% (**95**%)	79% (**92**%)	
Conduction vs HC	92% (**95**%)	93% (**94**%)	91% (**92**%)	
Anomic, Wernicke, Conduction	46% (**55**%)	46% (**60**%)	56% (**47**%)	

Previous Model's performance (using all data) highlighted in bold

Results: interpretability – Anomic vs HC

Results: interpretability – Anomic vs HC

```
True:   ANOMIC Pred:   ANOMIC
Text:   okay. [FP] went to school at jamaica. [FP] head boy. so [FP] head boy and wear epaulets and a tie. no i can't. i don't remember. yeah okay.
```

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
n/a	(0.07)	CONTROL	-4.67	[CLS] okay . [FP] went to school at jamaica . [FP] head boy . so [FP] head boy and wear epa ##ule ##ts and a tie . no i can ' t . i don ' t remember . yeah okay . [SEP]
n/a	(0.91)	ANOMIC	4.79	[CLS] okay . [FP] went to school at jamaica . [FP] head boy . so [FP] head boy and wear epa ##ule ##ts and a tie . no i can ' t . i don ' t remember . yeah okay . [SEP]

Results: interpretability – Wernicke vs HC

```
True:  WERNICKE Pred:  WERNICKE
Text:  very good. yes. it's always different. [FP] the talking they don't work. yeah.
```

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
n/a	(0.23)	CONTROL	-3.22	[CLS] very good . yes . it ' s always different . [FP] the talking they don ' t work . yeah . [SEP]
n/a	(0.78)	WERNICKE	3.21	[CLS] very good . yes . it ' s always different . [FP] the talking they don ' t work . yeah . [SEP]

Results: interpretability – Conduction vs HC

```
True:   CONDUCTION Pred:   CONDUCTION
Text:   how do they how do they what? okay. okay. okay. [FP] bad. bad. [FP] my [FP] me that i've been there then before that the stroke. [FP] it's it
```

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
n/a	(0.11)	CONTROL	-8.79	[CLS] how do they how do they what ? okay . okay . okay . [FP] bad . bad . [FP] my [FP] me that i ' ve been there then before that the stroke . [FP] it ' s it ' s [FP] deep . [FP] it doesn ' t it doesn ' t me at all . and it looks like [FP] echo hello echo . it doesn ' t like i ' m saying it . but it ' s going like [FP] it doesn ' t somebody else . it really does . [SEP]
n/a	(0.92)	CONDUCTION	8.65	[CLS] how do they how do they what ? okay . okay . okay . [FP] bad . bad . [FP] my [FP] me that i ' ve been there then before that the stroke . [FP] it ' s it ' s [FP] deep . [FP] it doesn ' t it doesn ' t me at all . and it looks like [FP] echo hello echo . it doesn ' t like i ' m saying it . but it ' s going like [FP] it doesn ' t somebody else . it really does . [SEP]

Conclusions I

- Language models can automatically detect fluent aphasia(s) with a very high accuracy

Conclusions I

- Language models can automatically detect fluent aphasia(s) with a very high accuracy

- This level of accuracy is largely maintained when using spontaneous speech sample based on only open-ended questions are used

Conclusions I

- Language models can automatically detect fluent aphasia(s) with a very high accuracy

- This level of accuracy is largely maintained when using spontaneous speech sample based on only open-ended questions are used

- We have demonstrated language model’s capability to learn linguistic patterns and contextual cues from fluent aphasic speech
    - Pauses, interjections, or discourse markers seem to be the features that consistently predicts fluent aphasia(s)

Conclusions I

- Language models can automatically detect fluent aphasia(s) with a very high accuracy
- This level of accuracy is largely maintained when using spontaneous speech sample based on only open-ended questions are used
- We have demonstrated language model’s capability to learn linguistic patterns and contextual cues from fluent aphasic speech
  - Pauses, interjections, or discourse markers seem to be the features that consistently predicts fluent aphasia(s)
  - A systematic profiling of the words or phrases with highest attribution should be done in future work

Conclusions II

- Classification between the clinical groups shows a reduced level of accuracy, although still above chance level

Conclusions II

- Classification between the clinical groups shows a reduced level of accuracy, although still above chance level
- This could be interpreted as:
  - Presence of overlapping linguistic features across the three groups

Conclusions II

- Classification between the clinical groups shows a reduced level of accuracy, although still above chance level
- This could be interpreted as:
  - Presence of overlapping linguistic features across the three groups
  - Over-reliance on the ‘gold standard labels’ from AphasiaBank

Thank you

Contact: f.tsiwah@rug.nl

Ruhi mahadeshwar

TalkBank

AphasiaBank

AphasiaBank is a shared database of multimedia interactions for the study of communication in aphasia. Access to the data in AphasiaBank is password protected and restricted to members of the AphasiaBank consortium group.

Researchers, educators, and clinicians working with aphasia who are interested in joining the consortium should read the ***Ground Rules*** and then send email to macw@cmu.edu with contact information and affiliation. Please include a brief general statement about how you envision using the data. Students interested in using the data should ask their faculty advisors to join as members. AphasiaBank is supported by NIH-NIDCD grant R01-DC008524 for 2022-2027.

Results: using all data

	Acc. (BM)	Precision (BM)	Recall (BM)	
Anomic vs HC	91% (**47**%)	89% (**50**%)	92% (**100**%)	
Wernicke vs HC	96% (**39**%)	95% (**47**%)	92% (**64**%)	
Conduction vs HC	95% (**72**%)	94% (**87**%)	92% (**18**%)	
Anomic, Wernicke, Conduction	55% (**40**%)	60% (**42**%)	47% (**41**%)	

Model’s baseline performance highlighted in bold rijksuniversiteit groningen

Results: interpretability – Wernicke vs HC

```
True:  WERNICKE Pred:  WERNICKE
Text:  and then [FP] [FP] let's see. and then [FP] yeah. okay [FP] and then page. [FP] [FP] and then [FP] laughing and talking mean man girl. and [
```

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
n/a	(0.23)	CONTROL	-5.71	[CLS] and then [FP] [FP] let ' s see . and then [FP] yeah . okay [FP] and then page . [FP] [FP] and then [FP] laughing and talking mean man girl . and [FP] and then [FP] and then horse . and a nice girl and horse . and then [FP] laughing and talking [FP] a [FP] three girls mean man . [FP] and then [FP] and then [FP] cinderella and . oh boy ! and then [FP] [FP] and then [FP] dancing . cinderella and another man and dancing . and then [FP] at twelve o ' clock at midnight and . uh ##oh . and . [FP] [FP] oh boy . oh ah yeah . and then [FP] and then next day and then [FP] mean man tree mean man . and then sit down and [FP] slip [FP] slip [FP] perfect . and . ah . [FP] [FP] two girl and man married . and wow , lord . [SEP]
n/a	(0.77)	WERNICKE	5.22	[CLS] and then [FP] [FP] let ' s see . and then [FP] yeah . okay [FP] and then page . [FP] [FP] and then [FP] laughing and talking mean man girl . and [FP] and then [FP] and then horse . and a nice girl and horse . and then [FP] laughing and talking [FP] a [FP] three girls mean man . [FP] and then [FP] and then [FP] cinderella and . oh boy ! and then [FP] [FP] and then [FP] dancing . cinderella and another man and dancing . and then [FP] at twelve o ' clock at midnight and . uh ##oh . and . [FP] [FP] oh boy . oh ah yeah . and then [FP] and then next day and then [FP] mean man tree mean man . and then sit down and [FP] slip [FP] slip [FP] perfect . and . ah . [FP] [FP] two girl and man married . and wow , lord . [SEP]