Using Learners Language Models for estimating learner language profiles

BERNARDO STEARNS

17 Apr 2023

- Research Associate at University of Galway working mostly in NLP projects
- PhD candidate in machine learning, applying NLP to Language Learning

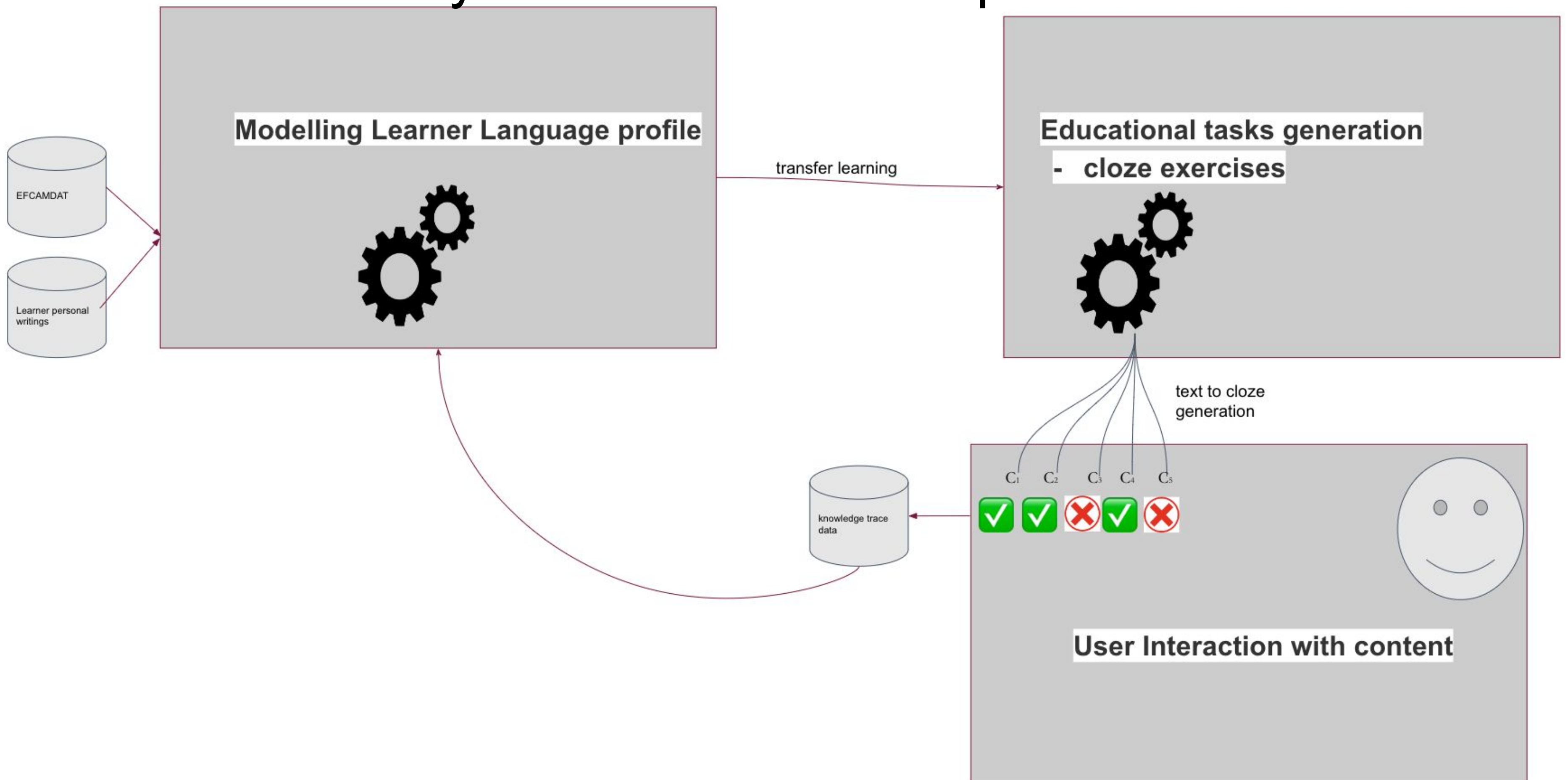
Supervisors:
- Dr. John McCrae
- Dr. Thomas Gaillat
- Dr. Bharathi Raja

Motivation

Cardamom

Comparative deep models for minority and historical languages

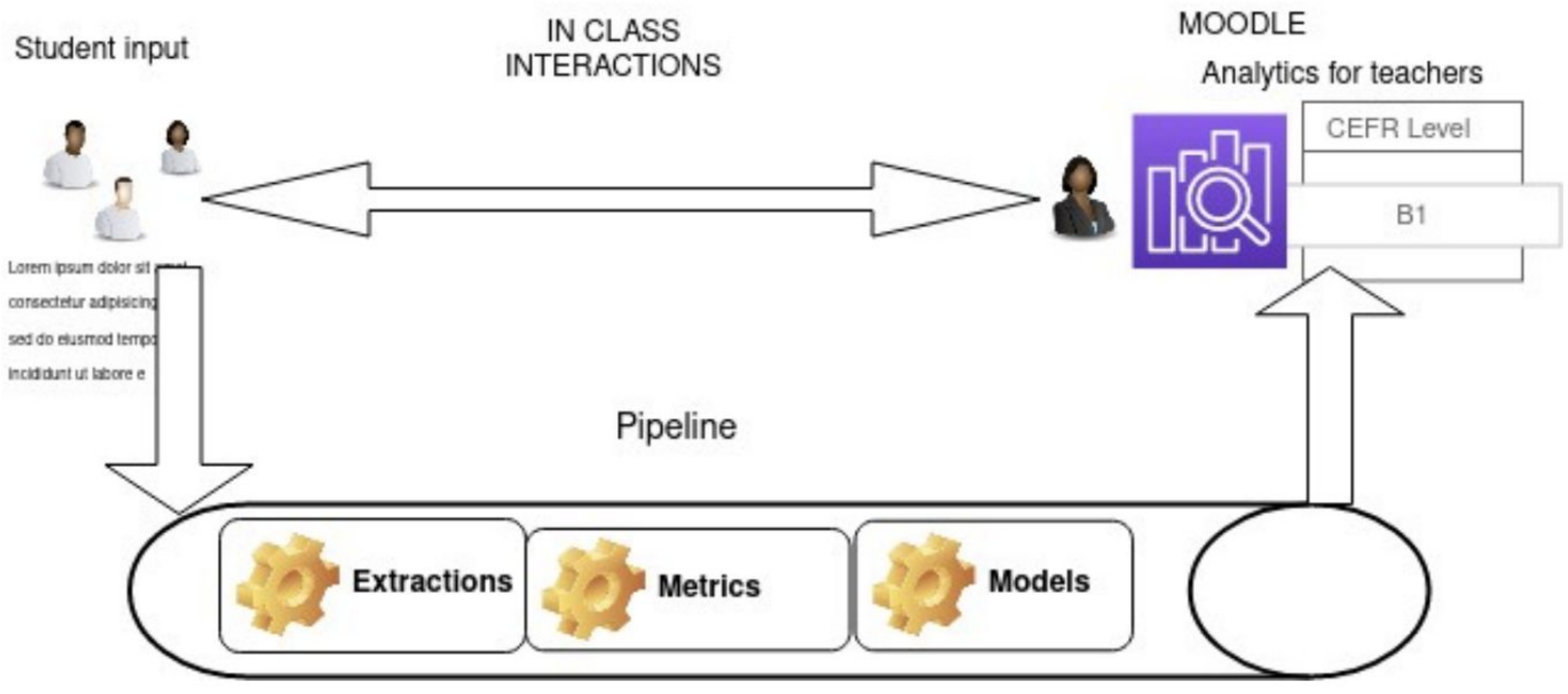
- **Aims to create a language learning app focused on less-resourced languages leveraging NLP tools**
- i) what are the NLP tools that can generate educational tasks from text?
- ii) how can we automatically adapt exercises difficulty based on learners profile?



A4LL

- **Aims to create a language-learning analytics system providing intuitive analysis of student's**
- i) what are the language features related to specific proficiency levels?
- ii) how can these features be measured automatically?

The system



Research Questions

- How can we numerically encode a learner's linguistic knowledge ?

How efficiently do language models adapt to predict tokens in ungrammatical sentences produced by language learners?

- How can those numerical representations be used in CALL downstream tasks ?

Explore how Learner Language Models could be used to build learner language profile

SIMILAR RESEARCH :

- User modeling in language learning with macaronic texts
- Predicting learner knowledge of individual words using machine learning

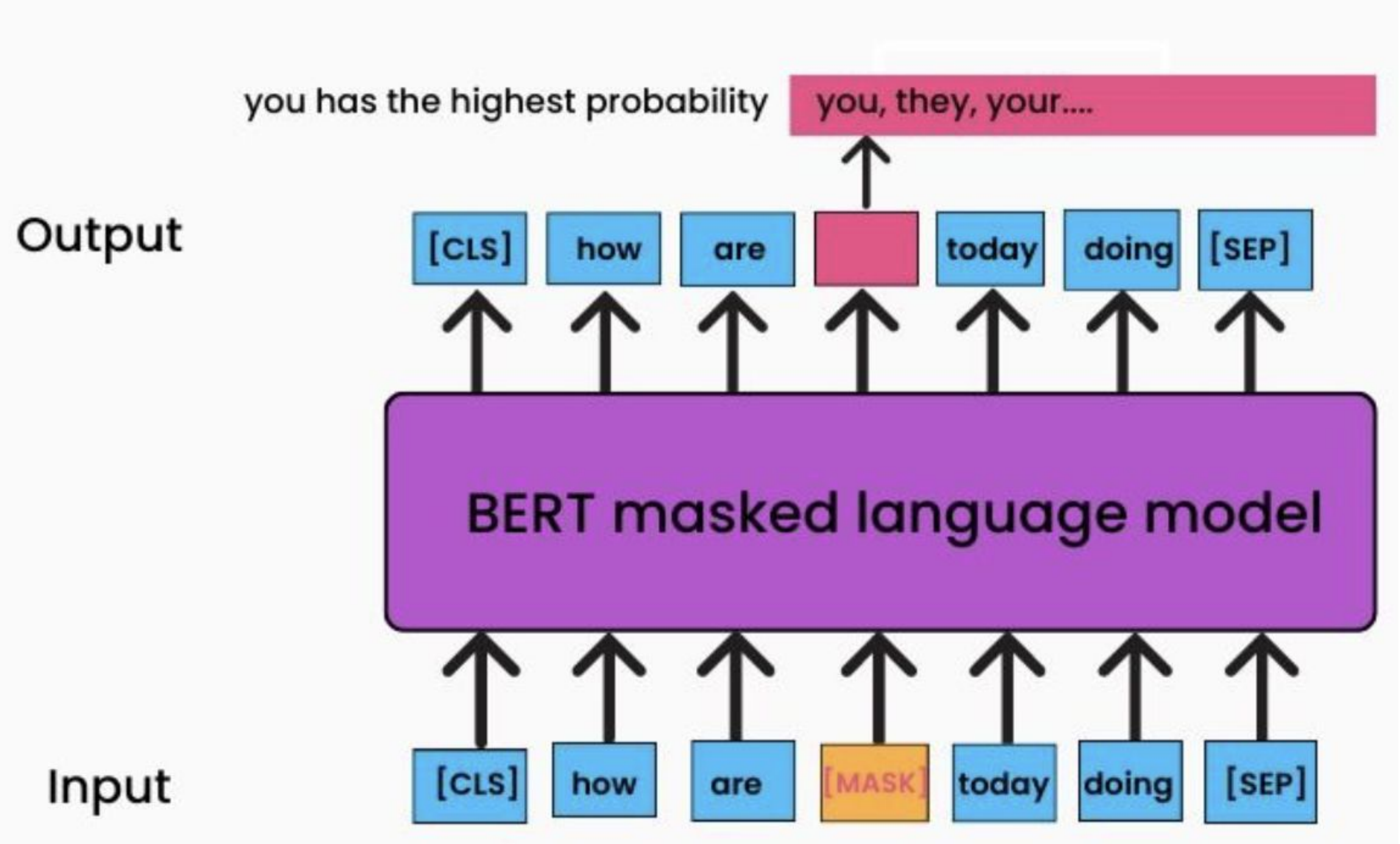
- Comparing Native and Learner Englishes Using a Large Pre-trained Language Model
- Synthetic Data Generation for Grammatical Error Correction with Tagged Corruption Models

- Probing pretrained language models for lexical semantics.

- Training Learner Language Models
  - Masked language modelling
  - Ungrammatical resources
  - Experiment
- LLMs for learner-language profiling
  - LM outputs & Statistics
  - Possible exploitations
  - Planned experiment

Masked Language Modelling

- How to encode a learners linguistic knowledge and skill ?
  - Transformers showed excellent results in encoding language representations
  - It benefited tremendously from self-supervised tasks, in specific Masked Language Modelling
  - trained over huge texts datasets



Ungrammatical Learner Resources

- How to encode a learners linguistic knowledge and skill ?
    - GEC and Learner Corpora Research created large ungrammatical datasets

more than 1m texts

1. LEARNER 18445817, LEVEL 1, UNIT 1, CHINESE
Hi! Anna,How are you? Thank you to sendmail to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are friends already,I hope we can learn english toghter! Bye! Anfeng.

2. LEARNER 19054879, LEVEL 2, UNIT 1, FRENCH
Hi, my name's Xavier. My favorite days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the afternoon, I play music or go by bicycle. I like sunday. And you ?

3. LEARNER 19054879, LEVEL 8, UNIT 2, BRAZILIAN
Home Improvement is a pleasant protest song sung by Josh Woodward. It's a simple but realistic song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the smooth guitar along with the high pitched resonant drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the negative aspects this new prosperous city brought. I really enjoyed this song.

Figure 1: Three typical scripts, in which learners are asked to introduce themselves (1), describe their favourite day (2), and review a song for a website (3).

200m sentences

Synthetic Data Generation for Grammatical Error Correction with Tagged Corruption Models

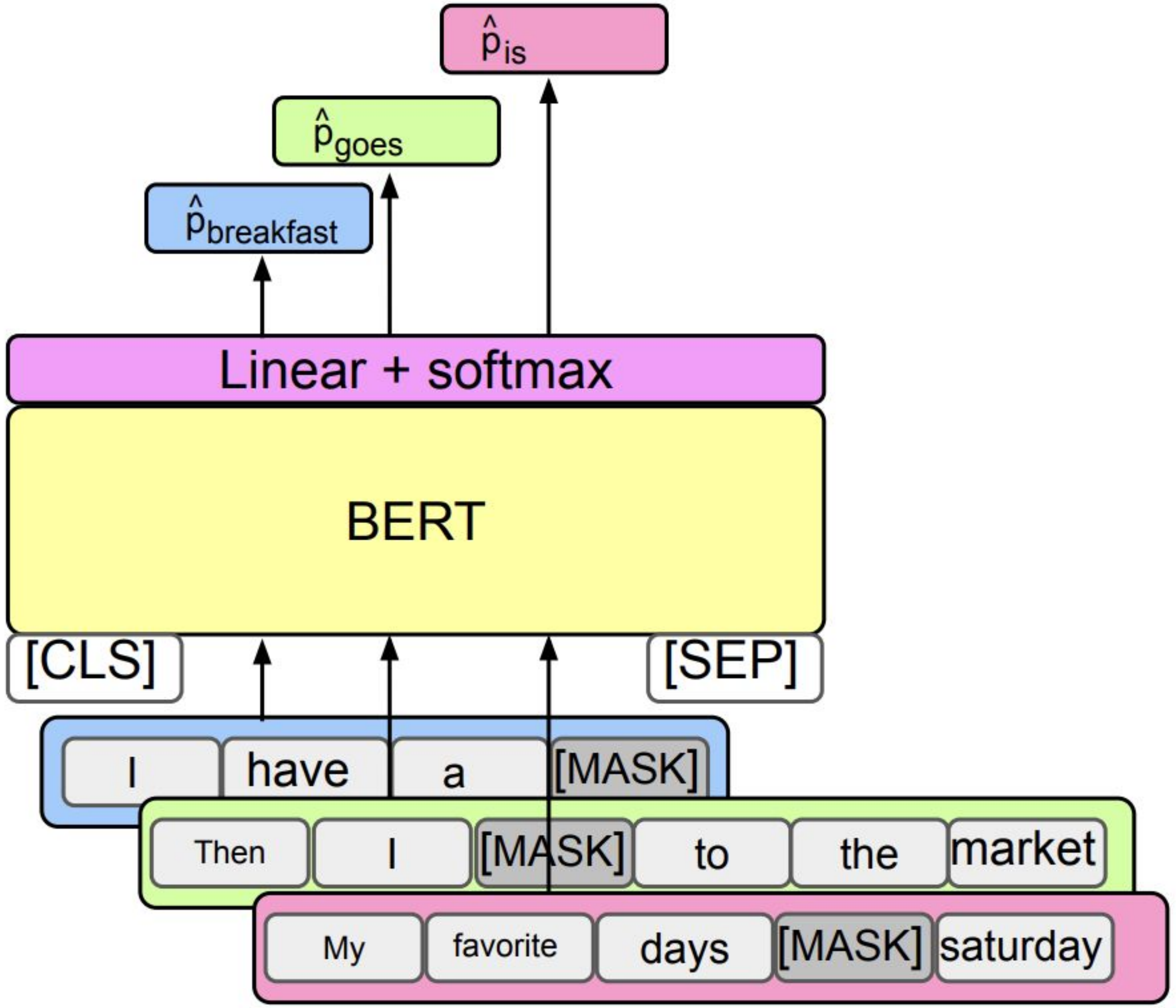
Felix Stahlberg and Shankar Kumar
Google Research
{fstahlberg,shankarkumar}@google.com

- Released 200m synthetic generated ungrammatical sentences
- Pre-training in this data improved SOTA GEC models

Input (clean)	I'm learning a lot and the students are very friendly.
Untagged corruption (1-best)	I'm learning a lot and the students are very friendly.
Untagged corruption (2-best)	I'm learning a lot **and students** are very friendly.
Tagged corruption	
ADJ	I'm learning a lot and the students are very **friendliness**.
ADJ:FORM	I'm learning a lot and the students are very **friendlies**.
ADV	I'm learning a lot and the students are **so** friendly.
CONJ	I'm learning a **lot the** students are very friendly.
CONTR	**I learning** a lot and the students are very friendly.
DET	I'm learning a lot **and students** are very friendly.
K	**I 'm** learning a lot and the students are very friendly.
MORPH	I'm learning a lot and the students are very **friendship**.
NOUN	I'm learning **many things** and the students are very friendly.
NOUN:INFL	I'm learning a lot and the **studentes** are very friendly.
NOUN:NUM	I'm learning a lot and the **student are** very friendly.
NOUN:POSS	I'm learning a lot and the **student's** are very friendly.
ORTH	I'm learning **alot** and the students are very friendly.
OTHER	I'm learning **very much** and the students are very friendly.
PART	I'm learning **up** a lot and the students are very friendly.
PREP	I'm learning **to** a lot and the students are very friendly.
PRON	**Learning** a lot and the students are very friendly.
PUNCT	I'm learning a lot and the students are very **friendly**
SPELL	I'm **lerning** a lot and the students are very friendly.
VERB	I'm learning a lot and the **students very** friendly.
VERB:FORM	I'm **learn** a lot and the students are very friendly.
VERB:INFL	I'm **learnes** a lot and the students are very friendly.
VERB:SVA	I'm learning a lot and the **students is** very friendly.
VERB:TENSE	**I learn** a lot and the students are very friendly.
WO	**I'm a lot learning** and the students are very friendly.

Experiment

- Given a learner in the EFCAMDAT dataset how well can we predict tokens from randomly generated masked sentences of this learner ?
- How effective fine-tuning bert in synthetic data, learner related texts and learner specific texts are for this task?



1. Learner 18445817, Level 1, Unit 1, Chinese

Hi! Anna,How [redacted] you? Thank you to [redacted] to me. My name's Anfeng.I'm 24 years old.Nice to meet you !I think we are friends already,I hope we can learn english toghter! Bye! Anfeng.

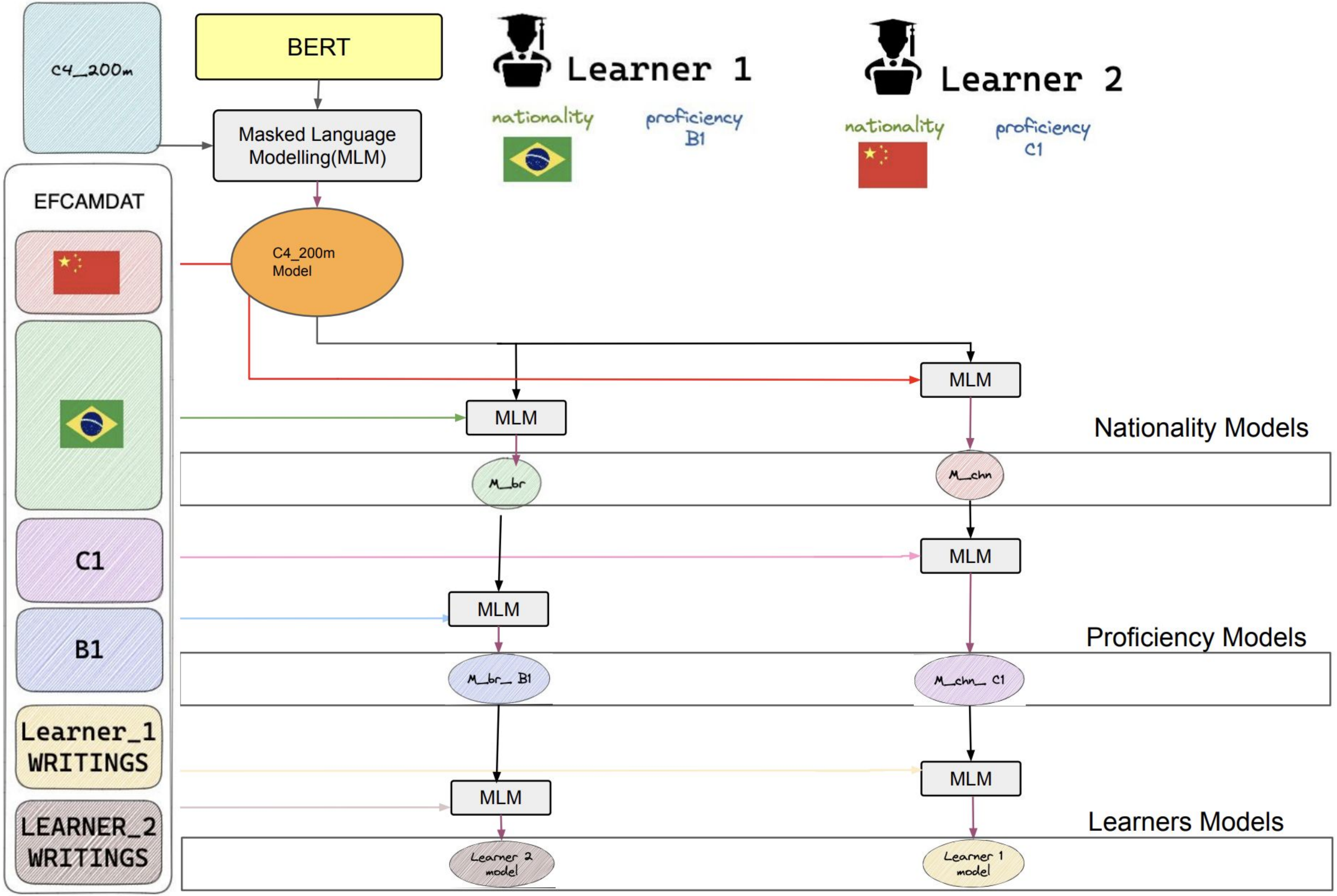
2. Learner 19054879, Level 2, Unit 1, French

Hi, my name's Xavier. My [redacted] days is saturday. I get up at 9 o'clock. I have a breakfast, I have a shower... Then, I goes to the market. In the [redacted], I play music or go by bicycle. I like sunday. And you ?

3. Learner 19054879, Level 8, Unit 2, Brazilian

Home Improvement is a pleasant protest song sung by Josh Woodward. It's a simple but [redacted] song that analyzes how rapid changes in a town affects the lives of many people in the name of progress. The high bitter-sweet voice of the singer, the [redacted] guitar along with the high pitched [redacted] drum sound like a moan recalling the past or an ode to the previous town lifestyle and a protest to the [redacted] aspects this new prosperous city brought. I really enjoyed this song.

Figure 1: Three typical scripts, in which learners are asked to introduce themselves (1), describe their favourite day (2), and review a song for a website (3).



C4_200m
BERT
Masked Language Modelling(MLM)
Learner 1
nationality
proficiency B1
Learner 2
nationality
proficiency C1
EFCAMDAT
C4_200m Model
C1
B1
Learner_1 WRITINGS
LEARNER_2 WRITINGS
MLM
M_br
M_chn
Nationality Models
MLM
MLM
M_br_ B1
M_chn_ C1
Proficiency Models
MLM
MLM
Learner 2 model
Learner 1 model
Learners Models

Model	MRR	average recall at k					
		1	5	10	25	50	100
unmodified bert(baseline)	0.564	0.466	0.677	0.743	0.814	0.851	0.881
+ c4200m	0.552	0.460	0.666	0.712	0.777	0.803	0.830
+ nationality	**0.667**	**0.575**	**0.780**	**0.822**	**0.871**	**0.893**	**0.908**
+ proficiency	0.582	0.480	0.681	0.749	0.831	0.873	0.884
+ learner	0.587	0.483	0.689	0.752	0.835	0.879	0.888

Table 3: Results of each group of pre-trained models on the EFCAMDAT test set

Thoughts

- **The are a significant number of possibilities of ordering and combinations of different sources of resources related to a learner.**
  - making training more expensive
  - a specific combination of resources can lead to better results other than using all resources

- **The masking strategy during training can dictate what linguistic aspects the model would focus**
  - random masking vs masking only verbs

- **masking tokens marked as errors where we investigate the error-annotated corpora to try to predict a given type of error.**

OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

LLMs for learner-language profiling

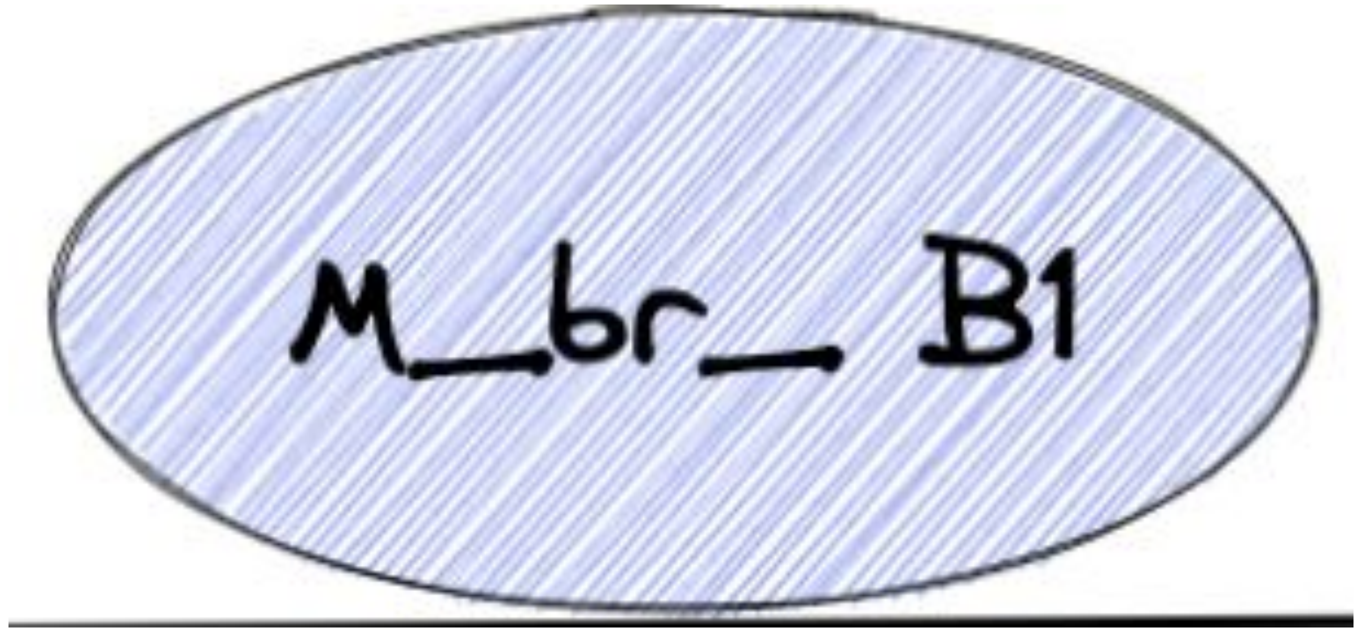
LM outputs

- The outputs are probabilities over cwe
- similar to work where CWE of specific words in an learner sentence is different from CWE of words in a native sentence. We investigate that the CWE created by a learner language model is different from a native language model

Then | I | [MASK] | to | the | market

NATIVE BERT

Went	go	walked	walk	headed
0.668	0.196	0.040	0.026	0.09

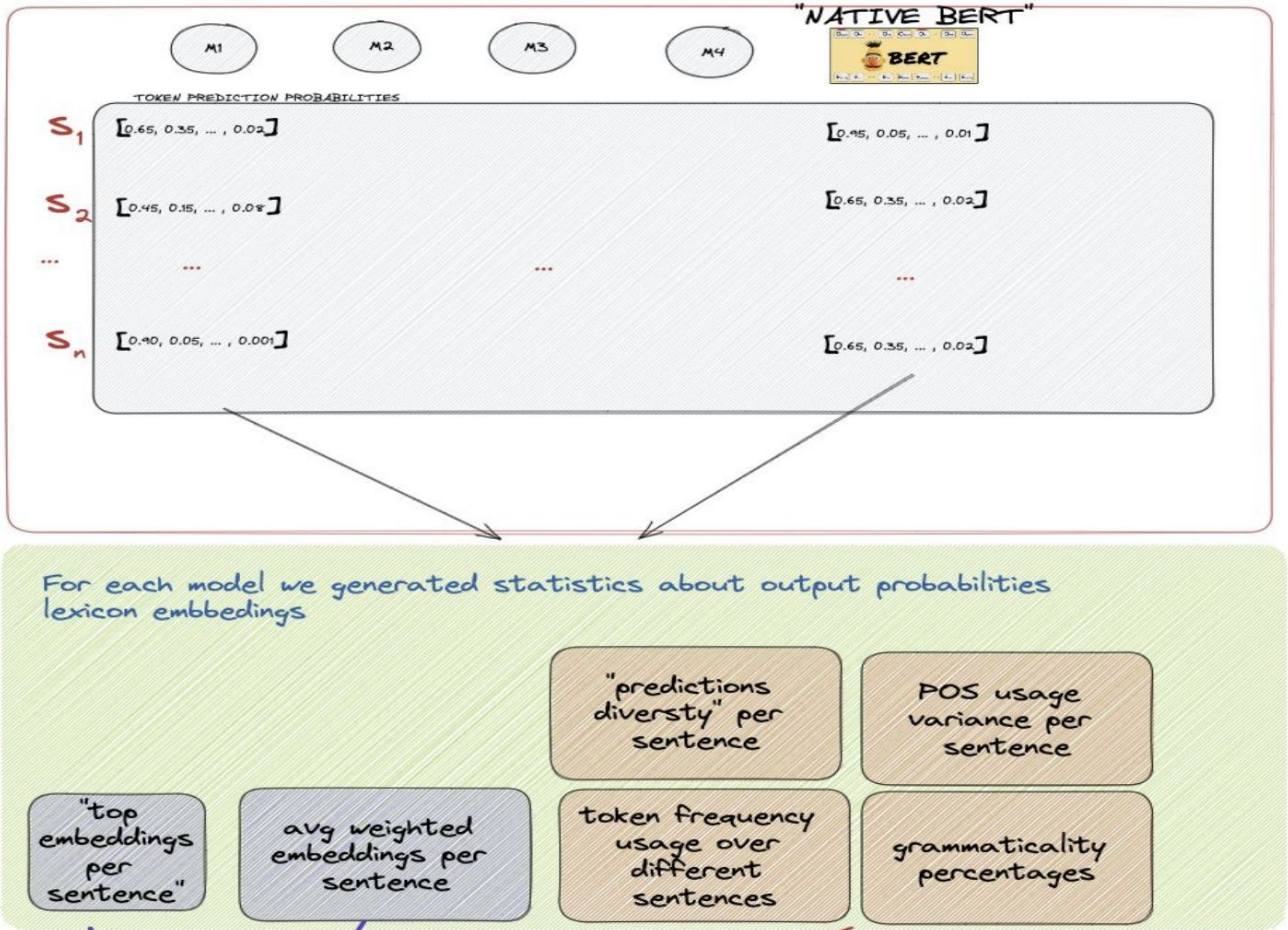


Goes	Went	walked	walk	headed
0.448	0.362	0.08	0.014	0.07

Statistics

- Conciliate linguistic statistics with probabilities of CWE

- Ideally we want to find a set of sentences that find difference in prediction statistics for different learner models



Exploitations

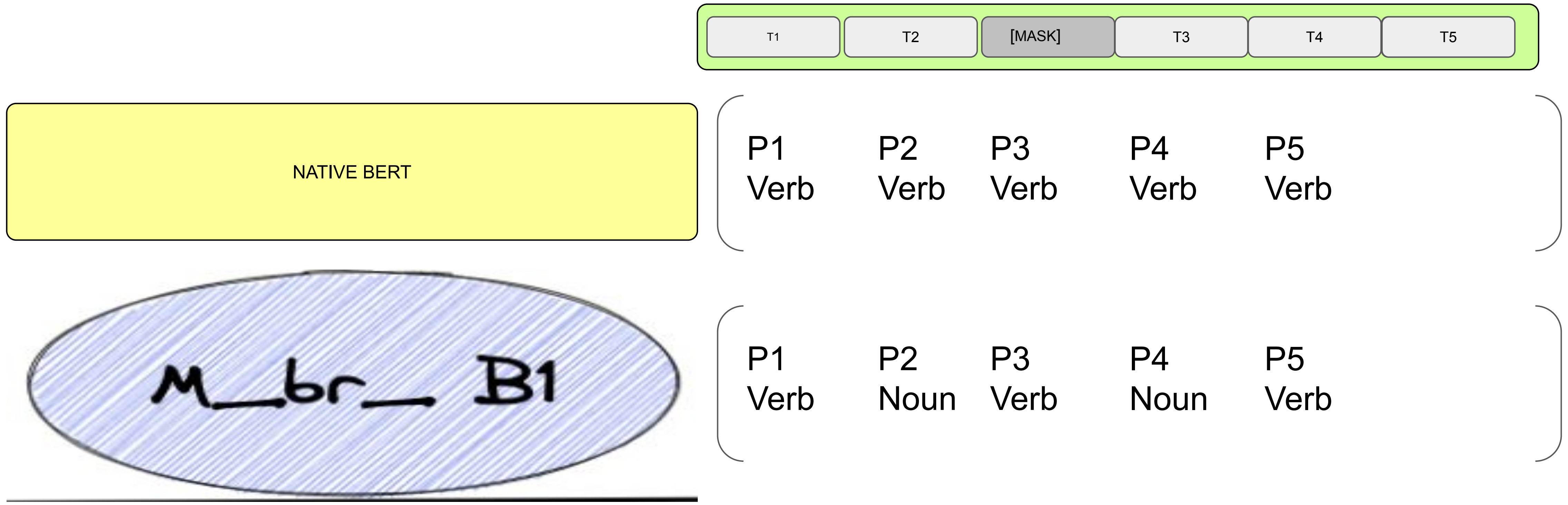
- **Efficient Learner Language Models can enable the simulation of learner behavior in an innumerable number of sentences/scenarios where it would be costly/infeasible to evaluate students.**

- **we have ungrammatical models and native models that can make inference in grammatical or ungrammatical sentences**
  - predictions of ungrammatical models over ungrammatical sentences gives us evidence of predicting token usage behavior
    - Investigate how well LM can replicate token usage behavior
    - Investigate which tokens in which contexts are hard to predict

	Learner Language Models	Native Language Models
Ungrammatical sentences	Evaluate how Language Models replicate a learner token production in a masked sentence	
Grammatical sentences	Evaluate sentences that Learners would be likely to make mistakes	Canonical behavior of Language Models

Exploitations

- **analyse the grammatical variance of the lexicon**.
  - Strong variance could indicate instability.
  - Find computationally sentences that tend to cause instability for specific CEFR levels or/and nationalities
  - How this is distributed across CEFR levels.



- **the tokens we mask can evaluate different lexical, grammatical or semantic skills**
  - hypernyms

References

Drilon Avdiu,Vanessa Bui, and Klára Ptacinová Klimcíková. 2019. Predicting learner knowledge of individual words using machine learning. In Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning
, pages 1–9, Turku, Finland. LiU Electronic Press.

Tatsuya Aoyama. 2022. Comparing native and learner englishes using a large pre-trained language model.In
Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning
, pages 1–9.

Adithya Renduchintala, Rebecca Knowles, Philipp Koehn, and Jason Eisner. 2016. User modeling in language learning with macaronic texts. In
Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)
,pages 1859–1869, Berlin, Germany. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications
, pages 37–47, Online. Association for Computational Linguistics.

Ivan Vulic, Edoardo Maria Ponti, Robert Litschko,Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) pages 7222–7240, Online. Association for Computational Linguistics.

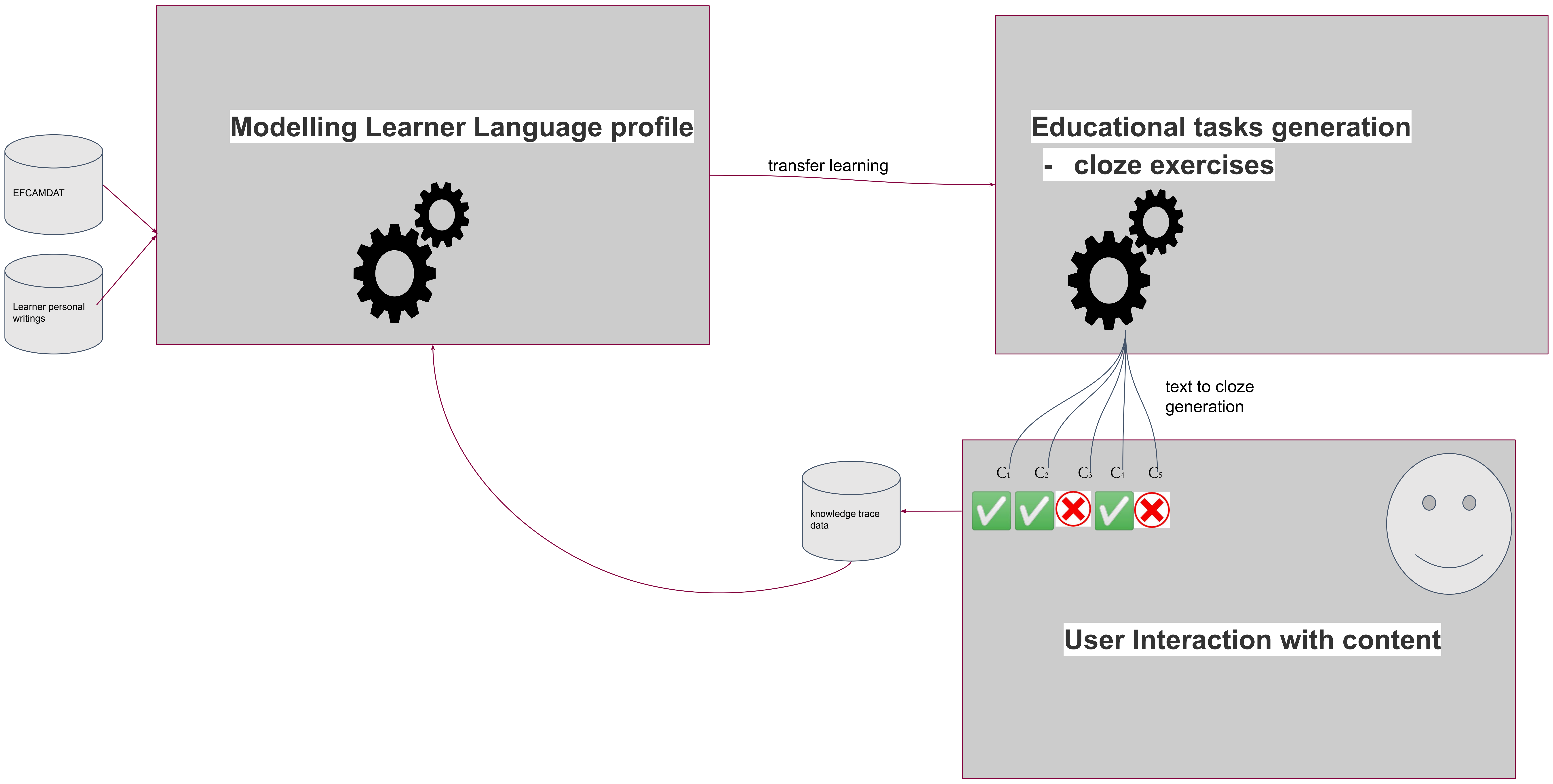
Ollscoil na Gaillimhe
University of Galway

Thank You

UniversityofGalway.ie

OLLSCOIL NA GAILLIMHE

UNIVERSITY OF GALWAY



Modelling Learner Language profile
EFCAMDAT
Learner personal writings
transfer learning
Educational tasks generation
- cloze exercises
text to cloze generation
C1 C2 C3 C4 C5
knowledge trace data
User Interaction with content