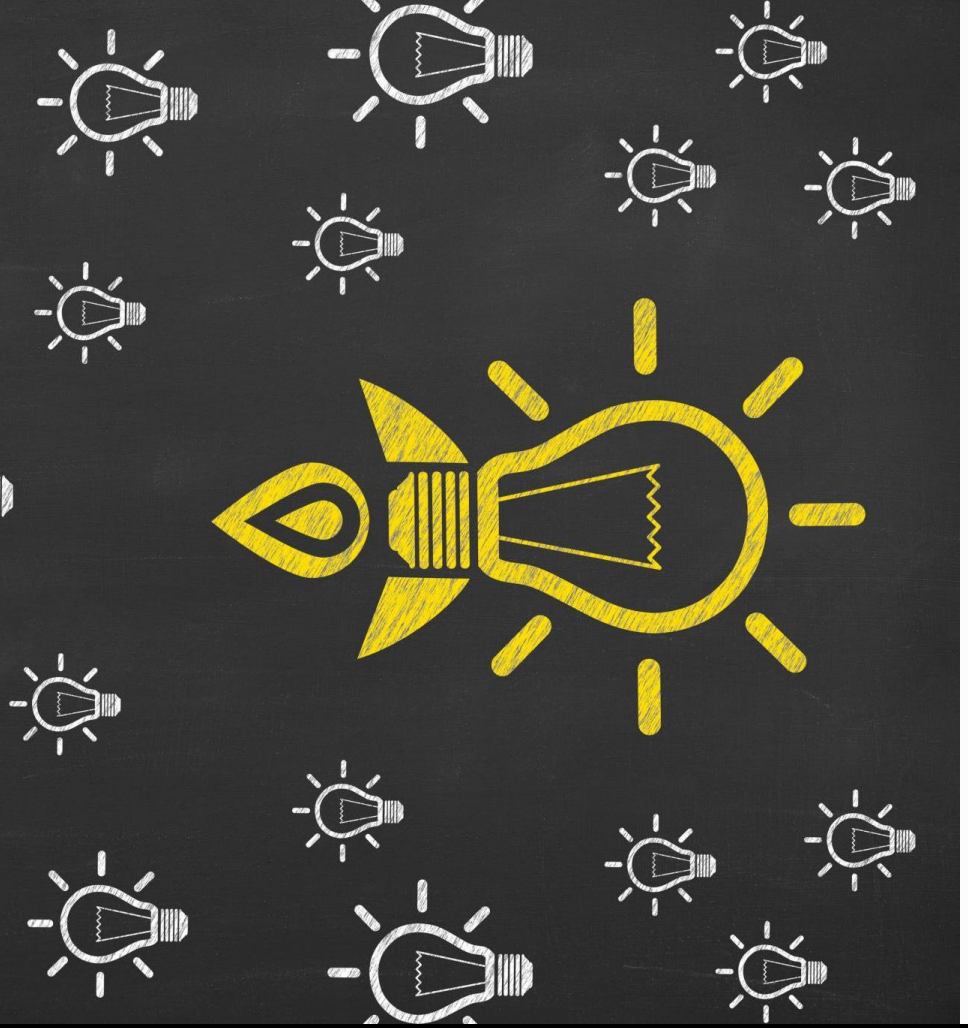
UNIVERSITY OF GOTHENBURG

SPRÅKBANKENTEXT

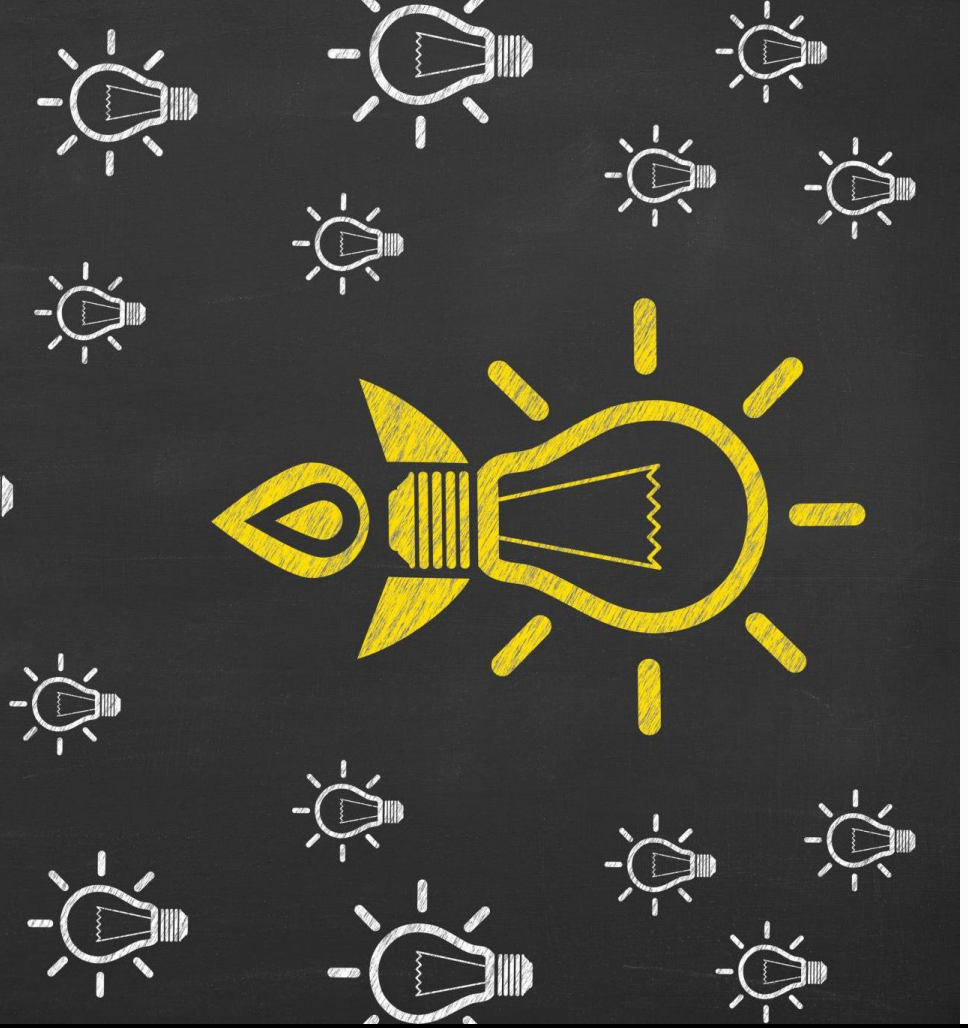
Jingle BERT, Jingle BERT, Frozen All the Way

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Szawerna, Elena Volodina



The Idea

- Transformer-based models are being increasingly used for automated essay scoring (AES)
- We know different layers encode different kinds of linguistic knowledge
- How much of this knowledge should we keep?



The Idea

- Said another way, how much domain adaptation is needed for this task?
- We focus on L2 learner texts in English, French, and Swedish
- We study BERT-like models for the three languages

Why BERT-like Models?

- Much research has aimed to learn which layers encode which aspects of linguistic knowledge
- The architectures of recent decoder-only models tend to vary a lot from each other
- Decoder-only models have had mixed results when dealing with AES of L2 learner texts

Methodology

- We use language-specific versions of BERT
- We truncate the essays to fit the maximum token length of the models
- We freeze the layers of the model bottom-up
  - Lower layers learn basic linguistic features
  - Higher layers learn more task-specific features
- We use the `[CLS]` token for classification

Language – English

- Model – BERT
  - We use the cased model
  - Trained on BookCorpus and Wikipedia dumps

- Dataset – EFCamDat
  - Essays collected from the EF online platform
  - Uses a 16-level scale with equivalence to CEFR levels
  - Grades were assigned based on level reached on a web platform, as opposed to direct assessment
  - Over 400K essays, we sampled 2% of the data

Language – French

- Model – CamemBERT
  - Based on RoBERTa
  - Trained on a French subset of CommonCrawl
- Dataset – TCFLE-8
  - Essays taken from the TFC French language certification exam
  - Each essay is assigned a level by at least two professional graders using the CEFR scale
  - Slightly over 6.5K essays

Language – Swedish

- Model – Swedish BERT
  - We used the cased model
  - Trained on the Nordic Pile
- Dataset – Swell-Pilot
  - Consists of three subcorpora gathered from different time periods
  - The CEFR label for each essay was aggregated from that from two professional graders
  - 502 essays

Results – Across Languages

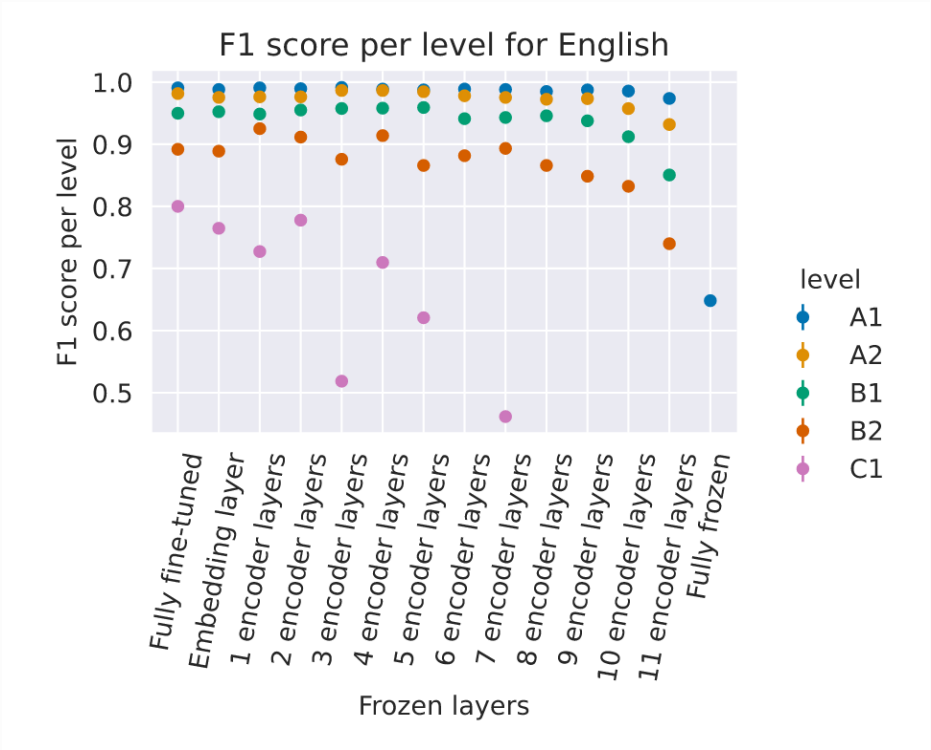
Layers Frozen	English	French	Swedish
State-of-the-art	0.974	0.56	0.23
None	0.975 ± 0.000	0.555 ± 0.003	0.722 ± 0.018
All layers	0.319 ± 0.000	0.443 ± 0.005	0.188 ± 0.001
Embedding Layer	0.971 ± 0.000	0.526 ± 0.005	0.727 ± 0.008
1 Encoder Layer	0.974 ± 0.000	0.517 ± 0.011	0.731 ± 0.019
1 and 2	0.974 ± 0.000	0.524 ± 0.010	**0.744 ± 0.011**
1 to 3	0.974 ± 0.000	0.538 ± 0.002	0.718 ± 0.006
1 to 4	**0.977 ± 0.000**	0.529 ± 0.011	0.720 ± 0.003
1 to 5	0.972 ± 0.000	0.537 ± 0.008	0.725 ± 0.010
1 to 6	0.966 ± 0.000	0.532 ± 0.017	0.705 ± 0.006
1 to 7	0.967 ± 0.000	0.542 ± 0.018	0.671 ± 0.009
1 to 8	0.962 ± 0.000	0.548 ± 0.006	0.664 ± 0.020
1 to 9	0.957 ± 0.000	0.552 ± 0.004	0.612 ± 0.011
1 to 10	0.946 ± 0.000	0.564 ± 0.004	0.596 ± 0.013
1 to 11	0.919 ± 0.000	**0.572 ± 0.001**	0.541 ± 0.004

Results – Across Languages

- The English and Swedish models performed best when freezing just some of the encoder layers
  - This points to the importance of surface-level features for identifying the CEFR levels of the essays

- The French model performs best when freezing most of the decoder layers
  - This indicates that a broader range of linguistic features might be necessary to accurately classify the essays

Results – Across CEFR Levels (English)

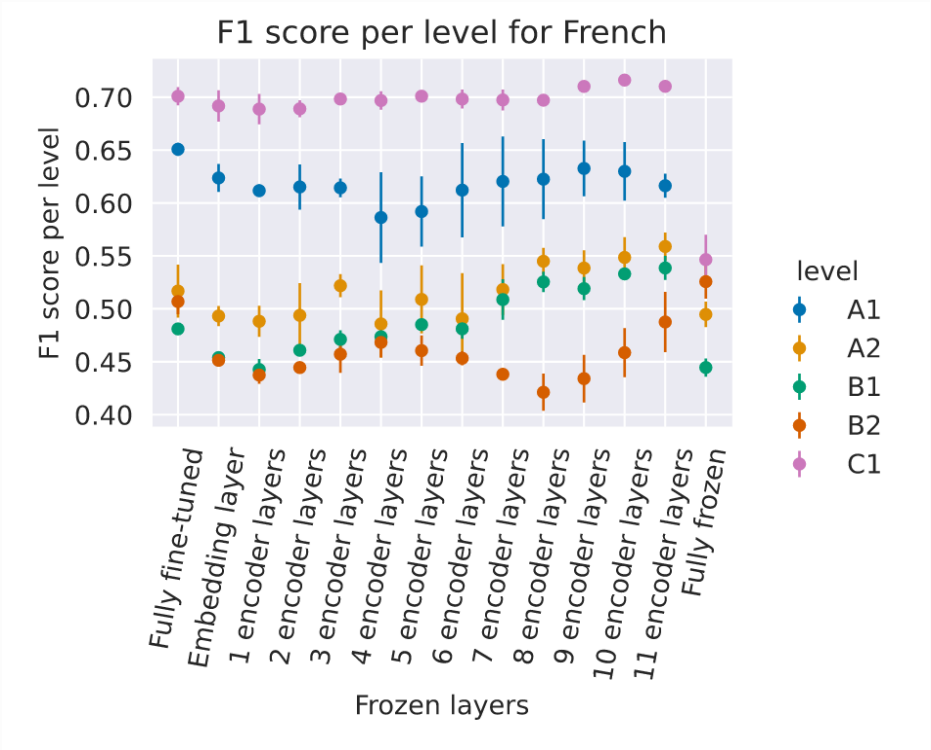


Results – Across CEFR Levels (English)

- Performance is inversely correlated to CEFR level
  - This might be due to the prompts given to the students
  - Another reason was that course level was used as a proxy for CEFR level

- When looking at individual levels
  - F1 score tends to decrease as we freeze more layers
  - There does not seem to be a particular pattern regarding variations

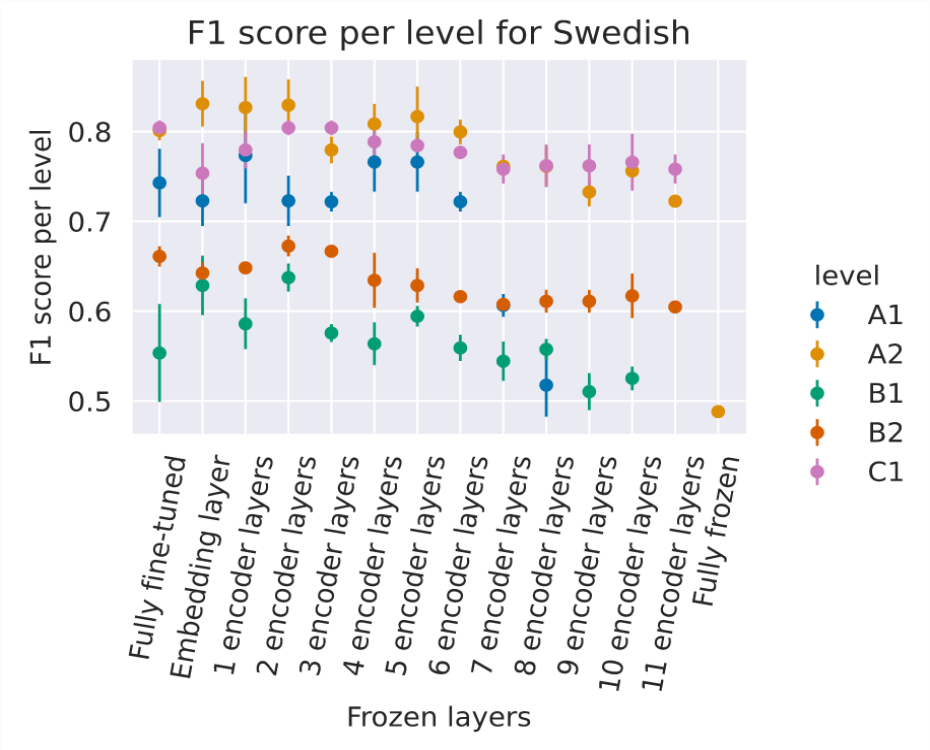
Results – Across CEFR Levels (French)



Results – Across CEFR Levels (French)

- Most levels have a slight increase in performance as we freeze more layers
- Different levels get better performance when freezing different numbers of layers
- This points to low, mid and high level features being important for AES in French

Results – Across CEFR Levels (Swedish)



Results – Across CEFR Levels (Swedish)

- For levels A1 and A2
  - There are two humps: one at the first few layers and one at around the fifth or fourth layers
  - This points to the importance of lexical and syntactic features
- Level B1 follows a similar pattern to A1 and A2 albeit more erratic
- For levels B2 and C1
  - Freezing the first two encoder layers leads to the highest performance
  - This points to the importance of lexical features

Results – General

- All partially fine-tuned models outperformed the fully-frozen ones
- Misclassified essays were usually assigned to one of the adjacent levels
  - CEFR levels are ordinal to humans but not for computers
  - This points to the models relying on linguistic characteristics to identify the level of an essay
- The levels where the model performs best are those at the edges of the CEFR scale for French and for Swedish

Takeaways

- Domain adaptation through partial fine-tuning seems to be the best strategy
- Maintaining basic knowledge of the language within the models is important for AES
- Different layers are important for different languages, but they all follow the model's general pattern

Caveats

- Analyzing prompts and the terms the essays have is important
- Having different models with different languages with different datasets means a lot of moving pieces
  - Having a multilingual model might not make things better, though
- Language learning is complex and using a single label might be overtly simplistic

GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io