BigNLI

Native Language Identification with Big Bird Embeddings

Sergey Kramp, Giovanni Cassani, Chris Emmery | LREC-Coling 2024

github.com/SergeyKramp/mthesis-bigbird-embeddings

Native Language Identification

What’s your L1?

- Given a text written in English:

“Thanks for borrowing me your phone.”

- Predict the native language (L1) of an author.
- Here: languages that do not differentiate between “lending” and “borrowing”.

Use Cases

Can be used for:

- Inferring country / region of origin (demographics).
- eLearning software.
- (Personalized) spelling and grammar checkers.
- Fraud detection.

Motivation

Replication and SOTA

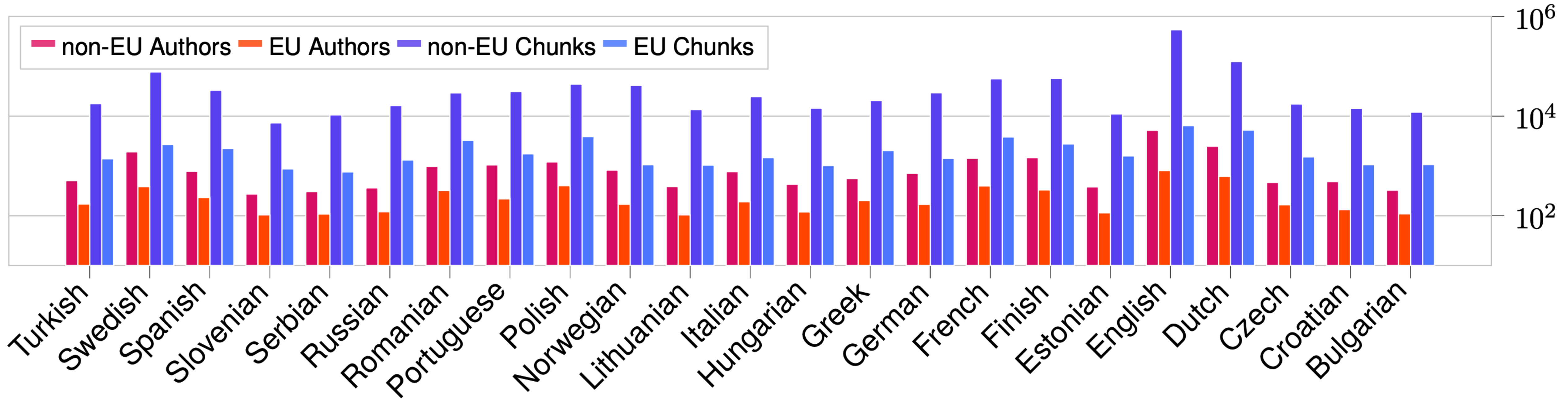
- Replicated Goldin et al. (2018); a feature-engineering approach.
- Shown to compete with transformer-based models at the time.
- We assumed input length might be a limit factor.
- Proposed to use Big Bird (Zaheer et al., 2020) to extract features.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Brussels, Belgium. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Pro- cessing Systems 2020, NeurIPS 2020, Decem- ber 6-12, 2020, virtual*.

Data

(Altered) Reddit L1



- 200M sentences (~3B tokens).
- Annotated using flairs in Europe-centered subreddits.
- One subset with ∧ posts (europe), part with a variety (non_europe).
- Chunked per 100 sentences per author.
- Downsampled and filtered for multi-lingual countries.
- Split non_europe in D_{tune} and D_{exp}.
- Evaluate on europe as D_{oos} and TOEFL as D_{ood}.

TILBURG UNIVERSITY

Baseline

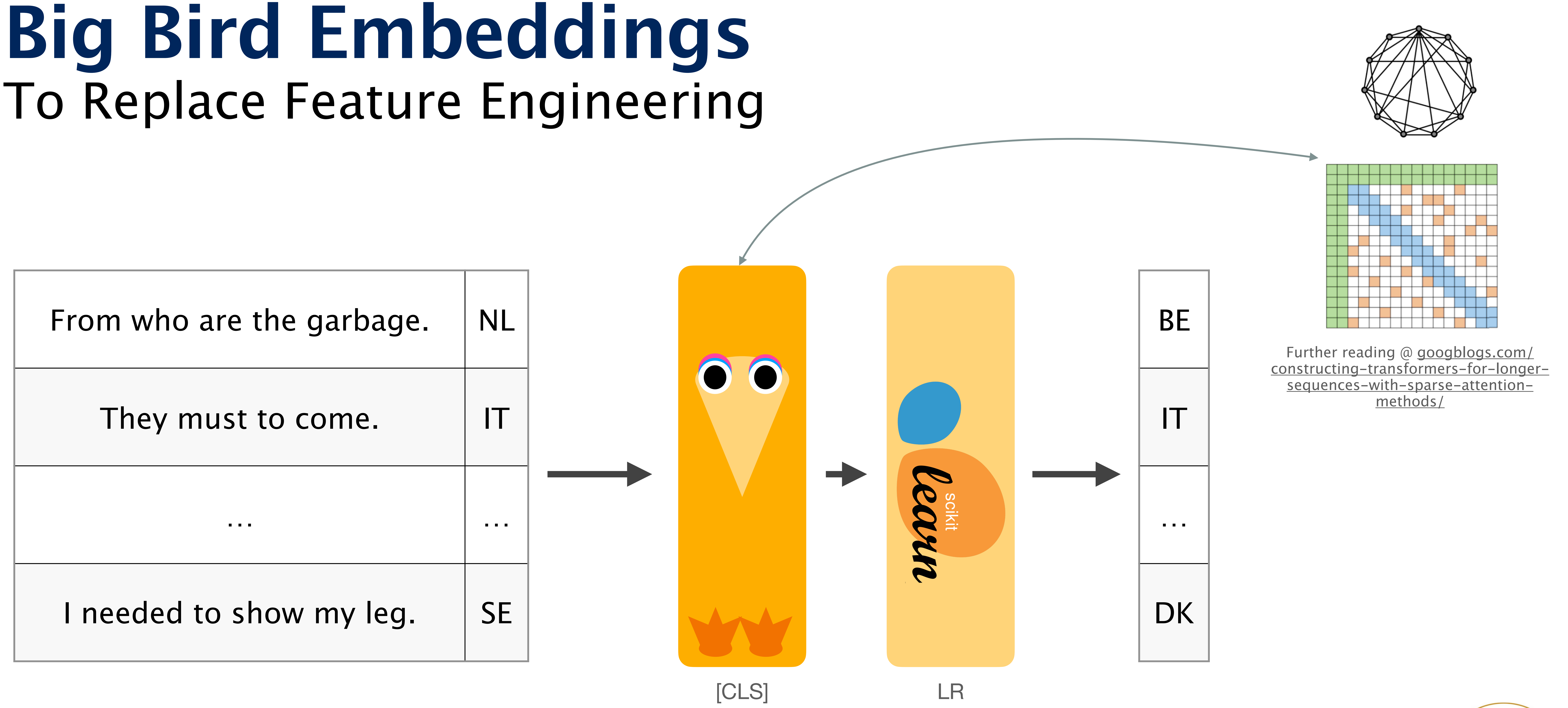
Goldin et al.’s Model; Generally Expensive

n–Gram + Logistic Regression model, using:

- Top 1000 word uni–gram and character tri–grams.
- Top 300 POS tri–grams
- Top 400 most frequent misspellings (average Levenshtein distance).
- Grammar errors (binary).
- Function word frequency.
- Average sentence length.

Big Bird Embeddings

To Replace Feature Engineering



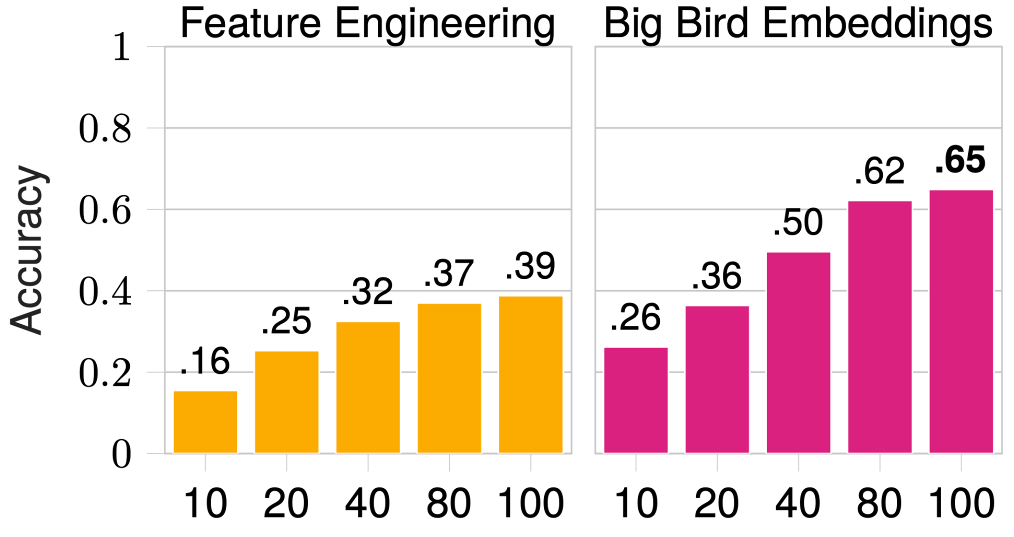
From who are the garbage.	NL
They must to come.	IT
…	…
I needed to show my leg.	SE

BE
IT
…
DK

Further reading @ googblogs.com/constructing-transformers-for-longer-sequences-with-sparse-attention-methods/

Results

Regular, Out-of-Sample, and Out-of-Domain



Name	Hours	ACV	OOS	OOD
Feature Eng.	13.00	.475	.637	.172
BigBird-2048	2.50	.493	.774	.102
BigBird-2048-t	2.50	**.654**	**.855**	**.204**

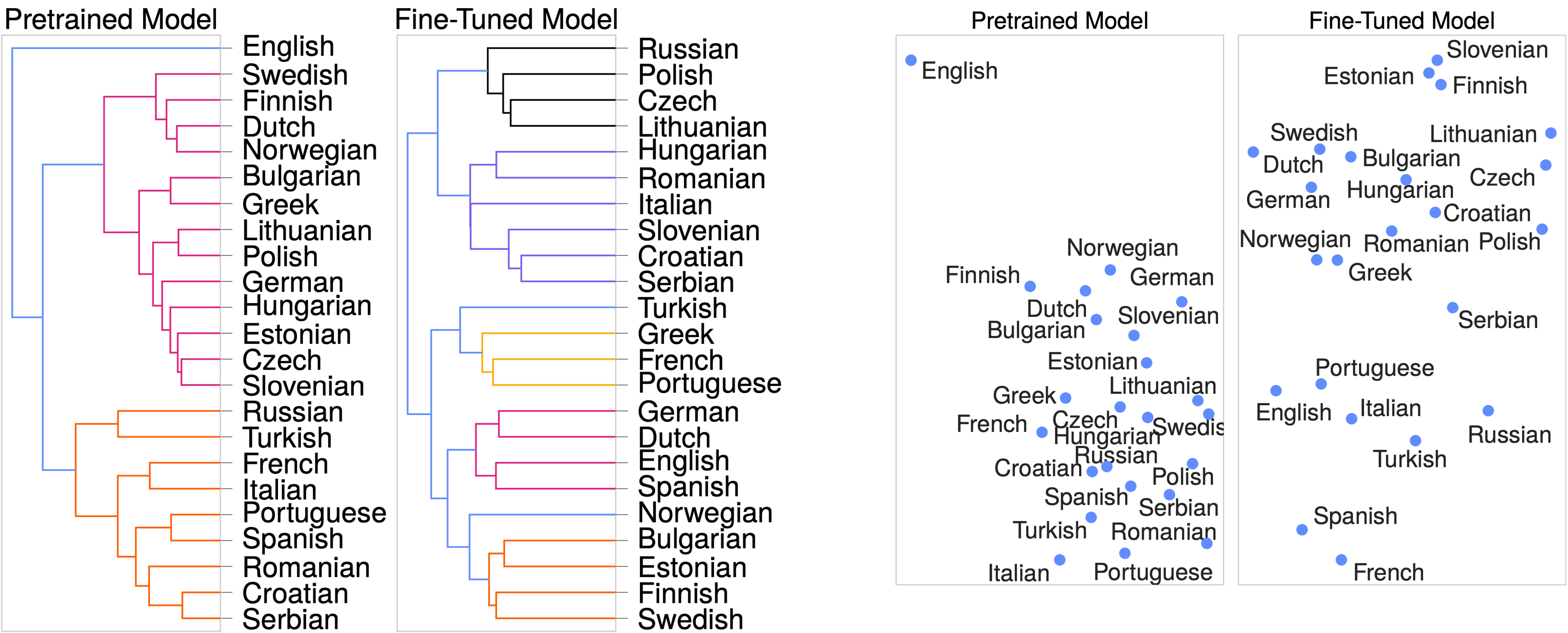
The models (Name) annotated with their input dimensions and if they were fine-[t]uned, how long feature extraction took on D_{exp} (Hours), their average cross-validation accuracy scores on D_{exp} (ACV) and accuracy scores on D_{cos} (OOS, europe) and D_{ood} (OOD, TOEFL-11). Fine-tuning took a day on a Titan X (Pascal).

Train	~~EU~~		TFL	
Test	~~EU~~	TFL	~~EU~~	TFL
Feature Eng.	.729	.262	.406	**.754**
BigBird-2048	.748	.280	.312	.660
BigBird-2048-t	**.821**	**.370**	**.610**	.560

Cross-evaluation accuracy scores between different models trained and tested on the `non_europe` (EU) and TOEFL-11 (TFL) datasets.

TILBURG UNIVERSITY

Qualitative Analyses



Left: Hierarchical clustering dendograms of native language centroids in the Big Bird embedding space before and after fine-tuning. **Right**: Dimensional PCA space showing the language centroids before and after fine-tuning.

Conclusion

Promising Results, Broader Perspective

- Simple embedding feature-extraction performs well (and is fast).
- Reddit is restricted (and biased); worth expanding data, classifiers, and architectures.
- Promising for text classification tasks on larger inputs (and on a budget).
- All models, code, and data for reproduction openly available via:
  - github.com/SergeyKramp/mthesis-bigbird-embeddings