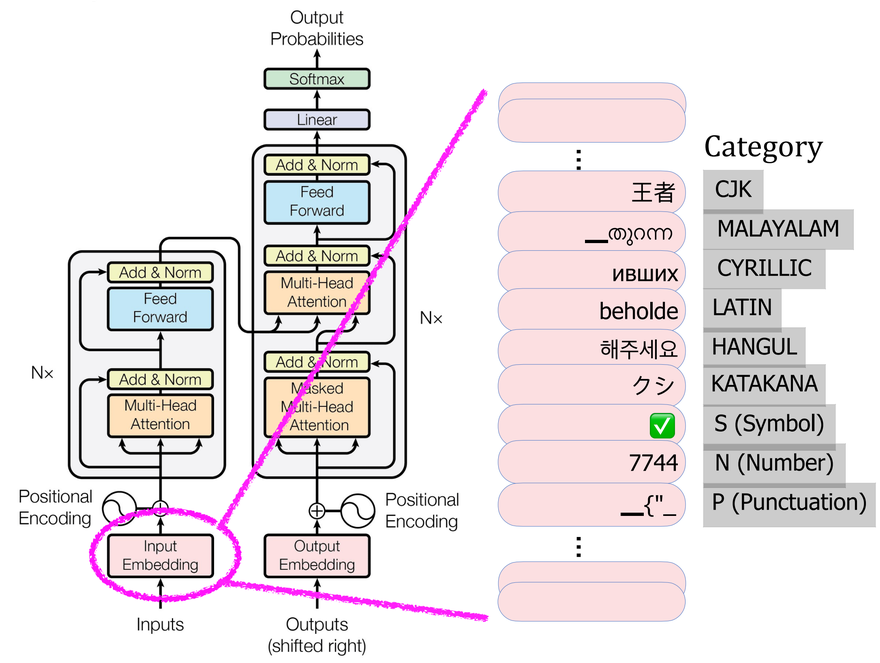
Hyperpolyglot LLMs: Cross-Lingual Interpretability in Token Embeddings

Cornell Bowers CIS Information Science

Andrea W Wen-Yi, David Mimno
(aww66@cornell.edu, mimno@cornell.edu)
Cornell University

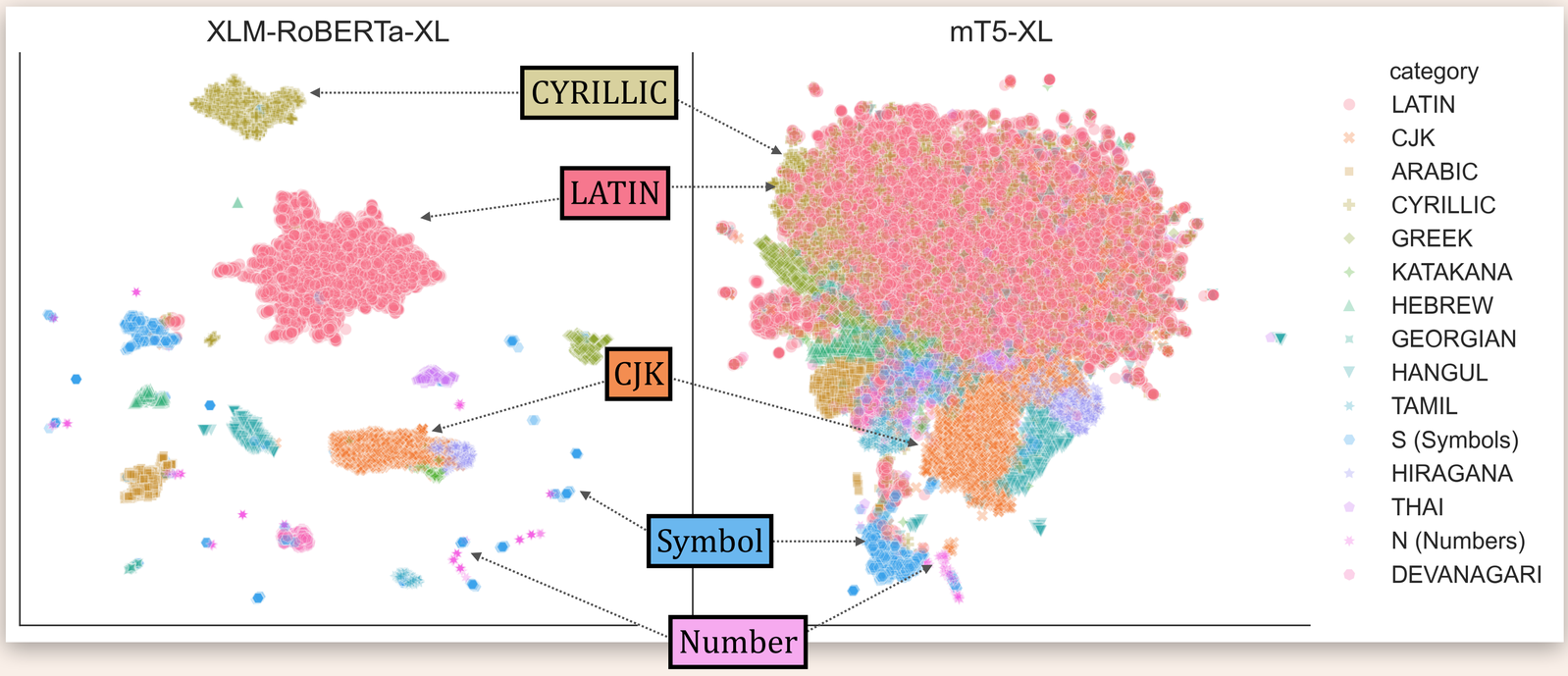
How do multilingual large language models represent meanings across languages?

LLMs have the potential to support transfer learning between languages with little to no additional training data. This work describes a mechanism for cross-lingual transfer learning by measuring the properties of token embedding layer, a ubiquitous yet often overlooked layer of LLMs. In the 2010s, many researchers sought ways to align word embeddings across languages [Ruder et al., 2019; Mikolov et al., 2013, Lazaridou et al., 2015, Artetxe et al., 2017], we find that some LLMs achieve the alignment **as an emergent property**.

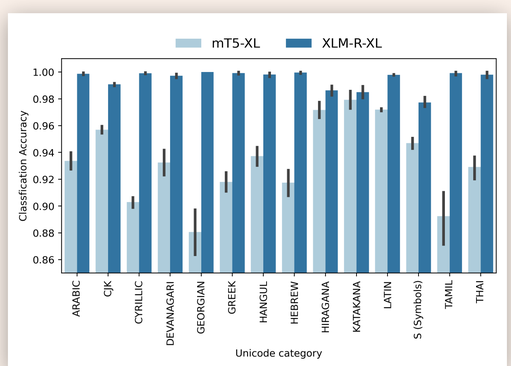


Model families encode languages differently.

(Both XLM-R and mT5 have 250,000 tokens.)



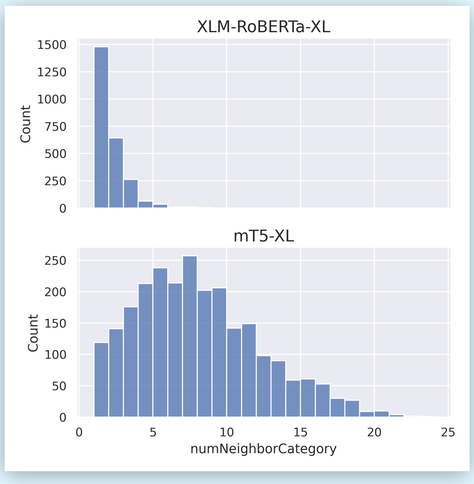
Even in high-dimensional space, mT5 cannot linearly separate languages as well as XLM.



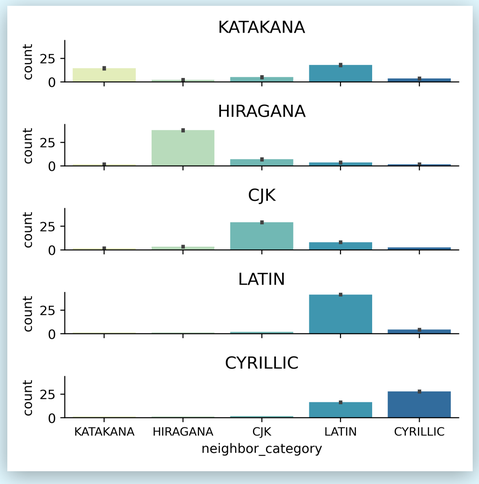
Embeddings encode semantics — mT5 discovers a shared semantic space that spans languages.

Token	när ("when", sv)		アメリカ ("America", ja)		Comment (en)	
Model	XLM-R-XL	mT5-XL	XLM-R-XL	mT5-XL	XLM-R-XL	mT5-XL
Nearest neighbors	when (en) når ("when", da) då ("then", sv) eftersom ("since", sv) två ("two", sv) nær ("closer", is) där ("where", sv) för ("for", sv) första ("first", sv) från ("from", sv) här ("here", sv) också ("also", sv) nästan ("almost", sv) även ("even", sv) är ("are", sv) innan ("before", sv) stora ("large", sv) människor ("people", sv)	når ("when", da) nær ("closer", is) when (en) lähe ("go", et) cuando ("when", es) quando ("when", pt) ש כ ("when", he) för ("for", sv) innan ("before", sv) near (en) att ("to", sv) Wenn ("if", de) hur ("how", sv) quand ("when", fr) quan	米国 ("USA", ja) イギリス ("England", ja) 韓国 ("Korea", ja) 미국 ("USA", ko) フランス ("France", ja) ドイツ ("Germany", ja) イタリア ("Italy", ja) アジア ("Asia", ja) 日本の ("Japanese", ja) ロシア ("Russia", ja) ヨーロッパ ("Europe", ja) 東京 ("Tokyo", ja) 海外 ("abroad", ja) 美国 ("USA", zh) 英国 ("England", zh) آمریکا ("America", fa) امریکا ("America", fa) 일본 ("Japan", ko) 美國 ("USA", zh) 現代 ("modern", zh)	アメリカの ("American", ja) 미국 ("USA", ko) Amerika ("America", ms) 米国 ("USA", ja) 美国 ("USA", zh) イギリス ("England", ja) Америка ("America", mk) ამერიკ ("America", ka) 美國 ("USA", zh) フランス ("France", ja) அமெரிக்க ("American", ta) America (en) അമേരിക്ക ("America", ml) amerikansk ("American", da) amerikanische ("American", de) Meerika อเมริกา ("America", th)	Comments (en) Review (en) Blog (en) Update (en) Text (en) Group (en) Info (en) Support (en) Photo (en) Share (en) Work (en) Information (en) Article (en) Link (en) Video (en) Chat (en) Search (en) News (en)	Comments (en) komment ("comment", is) Kommentar ("comment", sv) omentário Commentaire ("Comment", fr) Коментар ("Comment", bg) Kommentare ("Comments", de) Komentar ("Comment", ms) Koment ("Commentary", sq) kommentarer comentário ("comment", es) coment

Embedding neighborhoods in mT5 are more linguistically diverse.



The diversity of neighborhoods in mT5 vary by token category.



Embedding geometries are similar between models and scales.