RMIT
UNIVERSITY

Cross-linguistic comparison of linguistic feature encoding in BERT models for typologically different languages

Yulia Otmakhova, Karin Verspoor, Jey Han Lau

The University of Melbourne, RMIT[1]University

Motivation

Interest in cross-linguistic embeddings is on the rise, but are we studying them correctly?

Pairwise comparisons of languages contrasting in a particular linguistic feature

Languages and tasks we chose

- English vs Russian: fixed vs free word order -> word order corruption
- Russian vs Korean: inflected vs agglutinative with different head directionality -> word order corruption
- English vs Russian: morphologically poor vs rich -> long-distance agreement
- English vs Russian: absence and presence of grammatical gender -> gender bias

General probing approaches

Tasks 1 and 2

Bigram Shift classifier:

Randomly swap two adjacent words in half of the sentences

Train the classifier (frozen BERT + linear layer)

Task 2 has some restrictions in terms of movement within NP V chunk

Tasks 3 and 4

Masked word prediction:

Mask the token which grammatical word we want to predict from context

Check whether the probability of a word in correct grammatical form (number for task 3 and gender for task 4) is higher than that of incorrect one

Task 1: Sensitivity to word order corruption in languages with fixed vs free word order

Hypothesis: harder to detect in Russian

Results: harder to detect at lower layers, but easier at higher

	EN	RU
Layer 1	0.786	0.903
Layer 2	0.902	0.867
Layer 3	0.893	0.824
Layer 4	0.926	0.855
Layer 5	0.931	0.937
Layer 6	0.903	0.944
Layer 7	0.895	0.945
Layer 8	0.893	0.935
Layer 9	0.875	0.935
Layer 10	0.869	0.944
Layer 11	0.873	0.948
Layer 12	0.863	0.911

Task 2: Sensitivity to word order corruption in agglutinative and inflected languages with different head directionality

Hypothesis: models process right- and left context equally well

Results: no difference between SVO and SOV at lower (morphology levels)
Tokenization (and whether we move agglutinative units or whitespace words) influences the results

	KO1	KO2	RU
Layer 1	0.988	0.940	0.948
Layer 2	0.989	0.928	0.928
Layer 3	0.995	0.942	0.886
Layer 4	0.996	0.921	0.899
Layer 5	0.994	0.896	0.981
Layer 6	0.991	0.895	0.983
Layer 7	0.990	0.902	0.982
Layer 8	0.987	0.888	0.977
Layer 9	0.985	0.863	0.976
Layer 10	0.989	0.844	0.979
Layer 11	0.990	0.888	0.979
Layer 12	0.990	0.875	0.921

Table 4: The accuracy of word order corruption for an agglutinative SOV language vs inflected SVO language. KO1 refers to BShift using morphology-based tokenization; KO2 refers to BShift based on whitespace tokenization.

Task 3: Long-distance agreement in morphologically rich and poor languages

Hypothesis: easier for Russian as intervening context can have clues for the correct agreement

Results: true, but the performance suddenly increases only at the last (syntax) layer

	EN uncased		EN cased		RU cased	
	orig.	gen.	orig.	gen.	orig.	gen.
L 1	0.683	0.477	0.683	0.423	0.464	0.471
L 2	0.659	0.477	0.756	0.439	0.489	0.481
L 3	0.707	0.485	0.707	0.472	0.502	0.485
L 4	0.707	0.458	0.659	0.496	0.500	0.501
L 5	0.659	0.466	0.683	0.520	0.523	0.518
L 6	0.732	0.499	0.757	0.537	0.539	0.543
L 7	0.805	0.623	0.780	0.602	0.559	0.538
L 8	0.780	0.612	0.854	0.664	0.520	0.515
L 9	0.878	0.737	0.951	0.734	0.568	0.531
L 10	0.927	0.770	0.976	0.797	0.586	0.558
L 11	0.951	0.816	0.976	0.824	0.618	0.569
L 12	0.951	0.810	0.976	0.821	0.991	0.919

Task 4: Gender bias in languages with and without grammatical gender

Hypothesis: bias is more present for Russian

Results: true, especially for verbs and adjectives

	Winning rate		Avg. prob.	
	M	F	M	F
EN pronouns	50%	50%	0.268	0.296
RU pronouns	93%	7%	0.460	0.03
RU verbs	100%	0%	0.299	0.047
RU adjectives	100%	0%	0.091	0

Main takeaways

- More comparisons are needed to make strong conclusions
- Highlight the necessity of a stricter design on cross-lingual experiments
- Do not chose the languages just because you know them – chose the ones clearly contrasting in the feature you want to explore