How Multilingual is Multilingual BERT?

Telmo Pires, Eva Schlinger, Dan Garrette

Google Research

Background

BERT (Devlin et al, 2019) is a pretrained language model providing contextualized embeddings.

Multilingual BERT is BERT trained on concatenated Wikipedias of 104 languages.

- Language-agnostic: language not given as an input.
- One vocabulary for representing all languages.
- Therefore: can be used for cross-lingual transfer learning (train on one language, test on another)

Multilingual BERT

Multilingual BERT does **NOT**:

- Take a language identifier as input.
- Train with any explicit notion of translation.
- Explicitly project different languages into a “shared space”.

BUT, Multilingual BERT facilitates transfer across languages **REALLY WELL.**

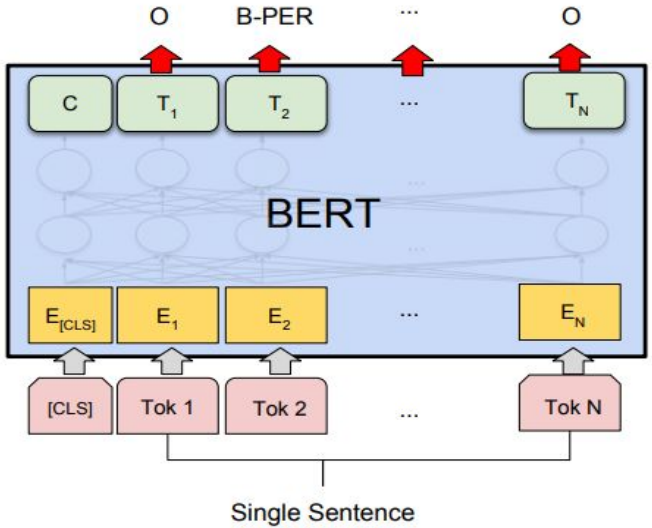
WHY??

Research Questions

1. Does transfer depend on **vocabulary overlap**? (no)
2. Does transfer depend on **typological similarity**? (yes)
3. Can it transfer to **mixed-language** or **transliterated** targets? (sort of)
4. Do **translations** have similar representations? (yes)

Experimental Setup

- Cross-lingual transfer:
  - Fine-tune on one language, test on another
- Sequence prediction tasks:
  - POS: 41 languages (Univ. Dependencies)
  - NER: 16 languages (CoNLL + Internal)



Does transfer depend on vocabulary overlap?

Pretrained on multiple languages: Transfer mostly independent of overlap.

Cross-lingual NER transfer F1 score

Pretrained on English only: Performance depends on overlap

Language pair's vocabulary overlap $= \frac{|E_{train} \cap E_{eval}|}{|E_{train} \cup E_{eval}|}$

Does transfer depend on vocabulary overlap?

Urdu: یہ ایک مثال کی سزا ہے

Hindi: यह एक उदाहरण वाक्य है

Different scripts ⇒ no vocabulary overlap

Urdu → **Hindi** transfer: **91%** POS accuracy

- Model has *never seen* an annotated Hindi word.
- Knows how to map Urdu annotations to Hindi words.

Conclusion: BERT is learning a multilingual representation.

Can it transfer to mixed-language or transliterated targets?

Code-mixing: I thought मौसम different होगा बस fog है

Code-mixing + transliteration: I thought mosam different hoga bas fog hy

BERT can handle code-mixing: small loss (90.56% ⇒ 86.59%) when fine-tuning on monolingual instead of code-mixed corpus.

But can't handle transliteration: huge loss (85.64% ⇒ 50.41%) when fine-tuning on non-transliterated corpus (instead of transliterated corpus).

Does transfer depend on typological similarity?

We compare language similarity using a set of used WALS typological features. **It's easier to generalize between similar languages.**

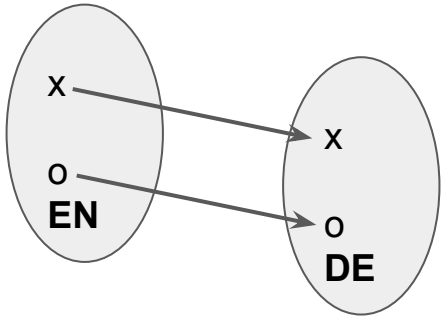
Zero-shot transfer works better when languages **share word order features** ⇒ BERT doesn't learn the systematic transformations necessary to accommodate different orders.

	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

Example: **English** → **Japanese** transfer: **49.4%** POS accuracy.

Do translations have similar representations?

We compute representations for each sentence in 5000 translation pairs, and find the displacement between the centroids for each language.



A **sentence's translation** is likely (for some layers, p > 70%) the **nearest neighbor** of that sentence plus the displacement vector.

Summary

1. Does transfer depend on **vocabulary overlap**? No.
2. Does transfer depend on **typological similarity**? Yes, there is a performance drop when changing word orders.
3. Can it transfer to **mixed-language** or **transliterated** targets? It is able to handle mixed-language, but not transliterated targets.
4. Do **translations** have similar representations? Yes.