It's not Greek to mBERT: Inducing Word-Level Translations from Multilingual BERT

Hila Gonen, Shauli Ravfogel, Yanai Elazar, Yoav Goldberg
hilagnn@gmail.com, shauli321@gmail.com, yanaiela@gmail.com, yoav.goldberg@gmail.com

paper

What is this paper about?

We show that the knowledge needed for **word-level translation** is implicitly encoded in multilingual BERT, and is easy to extract with simple methods.

What are the methods you use?

We present two methods for translation with mBERT: a **template-based** one, and an **analogy-based** one, see **orange** panel for details.

How do they perform?

Surprisingly well, see **cyan** panel for results.

Can you say anything about the way this information is stored in the representations?

Sure, we identify an empirical **language-identity subspace** in mBERT, and show that the representations in different languages are easily separable in that subspace, see **purple** panel for details..

How can I learn more about this?

Read our paper using the QR code.

Translation Methods

- **Template-based method:**
  - The word 'SOURCE' in LANGUAGE is: [MASK].
  - e.g. **The word *'nose'* in *French* is [MASK].**
- **Analogy-based method:**
  - We create language representations by averaging vectors in that language: $\overrightarrow{En}, \overrightarrow{Fr}$
  - To translate "nose" from En into Fr:

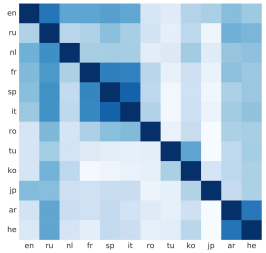
$$\overrightarrow{nose} - \overrightarrow{En} + \overrightarrow{Fr} = ?$$

Results

- **We translate 1016 words** (NorthEuraLex) into 11 languages, following are translation accuracies@1/@10/@100:

	@1	@10	@100
Baseline	0.036	0.244	0.575
Analogies	0.105	0.463	0.737
Template	**0.449**	**0.703**	**0.845**

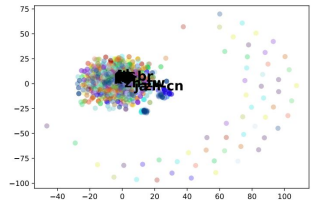
- Translation accuracy between the different languages reflects **typological relations**, as seen in the confusion matrix:



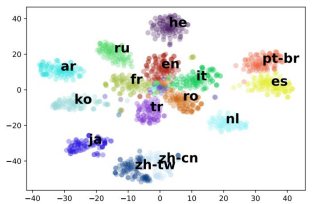
Dissecting mBERT representations

- We linearly decompose the representations into a **language-specific component** and a **language-neutral component** using INLP, a projection-based iterative algorithm (Ravfogel et al. 2020).
- We project representations on both spaces:

Neutral space



Language space



- We predict words from **language-neutral representations**. Instead of related words, we get translations into different languages:

mother		visited	
before	after	before	after
mother	mother	visited	visited
father	moeder	visits	visito
madre	mothers	*attended*	besøkt
mutter	мать	*visit*	visits
native	matki	visiting	besuchte
moeder	μητερα	visito	entered
mary	mutter	entered	visiting
true	madre	*joined*	asked
mothers	جنس	*toured*	vitja
the	مادر	*visite*	посет

Conclusion

- We study the **word-level translation information** embedded in mBERT and show its astonishing translation capabilities.
- We show that mBERT learns representations which contain both a **language-encoding component** and an abstract, **cross-lingual component**.