Multilingual representation

Multilingual Embedding, Multilingual Models and Multilinguality

Neural Representation Learning Seminar, CIS, LMU

Lecturer: Prof. Hinrich Schütze

Tutor: Victor Steinborn

Presenter: Ercong Nie

July 1st, 2022

Overview of the Contens

In this presentation, I would like to introduce ...

1. Some **core concepts** in the multilingual representation, such as multilingual embedding, multilingual models and multilinguality.
2. Some **metrics** to measure the multilinguality of a model.
3. That through the experiment results in the paper, it can be concluded that **4 architectural properties** and **2 linguistic properties** are essential for model's multilinguality.
4. The **knn-replace method** which is proposed to improve the model's multilinguality, based on the insights from the experiment.

Outline

1. Overview of Multilingual Representation
   - Multilingual Embedding
   - Multilingual Models
   - Multilinguality
2. Identification of Properties Essential for Multilinguality
3. Improving mBERT's Multilinguality
4. References

Two sources of multilingual embedding

Multilingual embedding from...

- **static** monolingual word embeddings of several languages
- multilingual pretrained language models (**MPLMs**)

From Static Embeddings

Mapping-based Approaches[1]

General steps of **mapping-based approaches**:

1. Train **static monolingual** word representations independently on monolingual corpora
2. Learn a **transformation matrix** mapping representations in one language to the other
   - Transformation can be learned from **word alignments** or **bilingual dictionaries**
   - Learning can be **supervised**, **semi-supervised** or **unsupervised**

[1] Ruder et al. (2019)

From Static Embeddings
VecMap[2]

VecMap:

- An embedding mapping method in **fully unsupervised** settings without the need of a seed dictionary
- **Core Idea**:
  - Utilizing the corresponding **similarity matrix** of each language: $M_X = X^T X$ and $M_Z = Z^T Z$.
  - If the embedding spaces of both languages are **isometric** (which is the assumption for mapping-based method), the two similarity matrices should be equivalent up to a permutation of their rows and columns.
  - **Solution**:
    1. Sort the matrices $sorted(M_X)$ and $sorted(M_Z)$;
    2. Find the corresponding translation for a word in row x_i of $sorted(M_X)$ through **nearest neighbor retrieval** over the rows of $sorted(M_Z)$.

[2] Artetxe et al. (2018)

From MPLMs

- **Higher** performance across tasks than static word embeddings.

1. Overview of Multilingual Representation
   - Multilingual Embedding
   - Multilingual Models
   - Multilinguality
2. Identification of Properties Essential for Multilinguality
3. Improving mBERT's Multilinguality
4. References

Motivation & Usage of Multilingual Models

Multilingual Models: Models capabale of processing more than one language with comparable performance

- **Fewer** models need to be maintained
- **Low- and mid**-resource languages will benefit from crosslingual transfer
- Useful in machine translation, zero-shot task transfer and typological research

Motivation & Usage of Multilingual Models

Multilingual Models: Models capabale of processing more than one language with comparable performance

- **Fewer** models need to be maintained
- **Low- and mid**-resource languages will benefit from crosslingual transfer
- Useful in machine translation, zero-shot task transfer and typological research

Examples of Multilingual Models

- **mBERT**: BERT-based model pretrained on Wikipedias of **104** languages with a shared subword vocabulary
- **XLM**: Transformer-based model with Masked Language Modeling (**MLM**) and Translation Language Modeling (TLM) as pretraining tasks
- **XLM-R**: RoBERTa-based model pretrained on **2.5TB** size of crawling data including 100 languages with a large vocabulary size of 250 thousand.

Why is mBERT (and other MPLMs) multilingual?

- The reasons for mBERT's multilinguality still remain **obscure**.
- Some explanations:
  - **Deep model structure** and **similar language structure** are necessary for multilinguality[3].
  - **Shared parameters** in the top layers of the model are required for achieving multilinguality[4].
  - Neither **shared vocabulary** nor **joint pretraining** is essential for multilinguality[5].

[3] Wang et al. (2019)
[4] Wu et al. (2019)
[5] Artetxe et al. (2019)

Identifying Essential Elements for Multilinguality

- **Goal**: Analyzing the reasons for the mBERT's multilinguality by identifying the essential properties in experimental setting.
- **Hypotheses**
  - **Architectural properties of model**: Overparameterization, shared special tokens, shared position embeddings, random word replacement
  - **Linguistic properties**: Word order, comparability of corpora

Identifying Essential Elements for Multilinguality

- **Goal**: Analyzing the reasons for the mBERT's multilinguality by identifying the essential properties in experimental setting.
- **Hypotheses**
  - **Architectural properties of model**: Overparameterization, shared special tokens, shared position embeddings, random word replacement
  - **Linguistic properties**: Word order, comparability of corpora
- **Overview of the whole work**:
  1. Design a small and simple version of mBERT for gaining quick insights in multilinguality investigation
  2. Design some metrics for evaluating the model's **degree of multilinguality** and model quality
  3. Design experiments to reduce the properties that are assumed to be essential for model's multilinguality
  4. Analyze the results to see if the model multilinguality is damaged while the model quality remains stable

Setup
Languages

- English and Fake-English
- **Fake-English**: created by shifting token indices after tokenization by a large constant (e.g., the **vocabulary size** of the English)

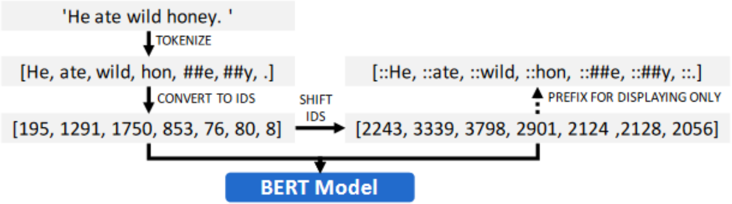

Figure 1: Creating a Fake-English sentence by adding a shift of 2048 to token indices.

- Shifted tokens are prefixed by "::" and added to vocabulary.
- Such created Fake-English has the exact same linguistic properties as English.

Data

- **Training data**: English Easy-to-Read version of the Parallel Bible Corpus
- Get a **sentence-parallel** corpus by creating a Fake-English version
- **Development data**: From the Old Testament of the English King James Bible
- **Vocabulary size**: 2048 * 2

Data

- **Training data**: English Easy-to-Read version of the Parallel Bible Corpus
- Get a **sentence-parallel** corpus by creating a Fake-English version
- **Development data**: From the Old Testament of the English King James Bible
- **Vocabulary size**: 2048 * 2

Model

- A smaller size of BERT-Base model: **BERT-small**
- Less hidden size, a single attention-head
- Pre-training objective: only masked language modeling
- Train a single model in $< 40min$ on a single GPU

Evaluation of Multilinguality I

Basic Idea: Evaluate model's multilinguality by using the representations from layers 0 and 8 for three different tasks.

- **Task 1: Word Alignment**
  - Gold word alignment: identity alignment
  - Alignment extraction method: **Argmax method**
  - Metric: F_1 score

Evaluation of Multilinguality I

Basic Idea: Evaluate model's multilinguality by using the representations from layers 0 and 8 for three different tasks.

- **Task 1: Word Alignment**
  - Gold word alignment: identity alignment
  - Alignment extraction method: **Argmax method**
  - Metric: F_1 score
- **Task 2: Sentence Retrieval**
  - Computing the **sentence similarity matrix** between English and Fake-English
  - Sentence embeddings computed simply by **averaging** token vectors
  - Retrieving sentences by similarity ranking
  - Metric: Mean precision ρ

Evaluation of Multilinguality II

- **Task 3: Word Translation**
  - Obtain word vectors by feeding each word individually to BERT
  - Then evaluate word translation in the similar way with **sentence retrieval**
  - Metric: Precision τ

Evaluation of Multilinguality II

- **Task 3: Word Translation**
  - Obtain word vectors by feeding each word individually to BERT
  - Then evaluate word translation in the similar way with **sentence retrieval**
  - Metric: Precision τ
- **Multilinguality Score**: Computed by averaging retrieval and translation results across both layer 0 and layer 8.

$$\text{Multilingual Score } \mu = 1/4(\tau_0 + \tau_8 + \rho_0 + \rho_8)$$

Evaluation of Model Quality

MLM Perplexity (with base e) is used for evaluating the model quality.

Perplexity

- an evaluation metric for language model quality
- the normalized inverse probability of the test data

The lower the perplexity, the better the language model

Architectural Properties I

- **1. Overparameterization:** *overparam*
  - **Hypothesis**: Models with a smaller number of parameters use parameters more efficiently and are more likely to create a multilingual space.
  - **Experiment**: Train a standard BERT-base model and compare the result with BERT-small
- **2. Shared Special Tokens:** *shift-special*
  - Special tokens: [UNK], [CLS], [MASK]...
  - **Hypothesis**: Shared special tokens may contribute to multilinguality since they could serve as "anchor points[6]".
  - **Experiment**: Shift the special tokens with the same shift applied to token indices.

[6] Wu et al. (2019)

Architectural Properties II

- **3. Shared Position Embedding:** *lang-pos*
  - **Hypothesis**: Position and segment embeddings are usually shared across languages
  - **Experiment**: Investigate their contribution to multilinguality by using language-specific position and segment embeddings by adding a constant to indices

	ENGLISH							FAKE-ENGLISH						
Tok.	195	1291	1750	853	76	80	8	2243	3339	3798	2901	2124	2128	2056
Pos.	1	2	3	4	5	6	7	129	130	131	132	133	134	135
Seg.	0	0	0	0	0	0	0	1	1	1	1	1	1	1

Figure 2: lang-pos

- **4. Random Word Replacement:** *no-random*
  - **Hypothesis**: In MLM task, 10% of the masks are replaced by randomly sampled tokens, which can come from the vocabulary of any language. This random replacement could contribute to multilinguality.
  - **Experiment**: Mask without using random words

Linguistic Properties

Basic Hypothesis: **Structural similarities** across languages contribute to the multilinguality[7].

- **1. Word Order:** *inv-order*
  - **Hypothesis**: Word order has some effect on multilinguality.
  - **Experiment**: Invert each sentence in the Fake-English corpus.
- **2. Comparability of Corpora:** *no-parallel*
  - **Hypothesis**: The similarity of training corpora contributes to structural similarities.
  - **Experiment**: Train on non-parallel corpus created by splitting the Bible into two halves, one half for English and Fake-English each.

[7] Wang et al. (2019)

Main Findings from the Experiment

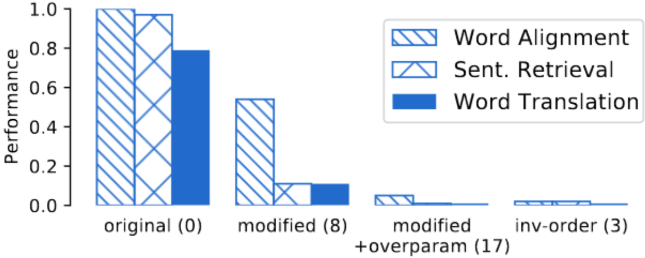

Figure 3: Results are for embeddings from layer 8

- Model 0: original
- Model 8: modified for three architectural properties: shared positional embeddings, shared special tokens, random word replacement
- Model 17: add one modiftication of overparameterization based on Model 8
- Model 3: Pairing a language with its inversion

Detailed Experiment Results

ID	Description	Mult.-score μ	Layer 0 Align. F_1	Layer 0 Retr. ρ	Layer 0 Trans. τ	Layer 8 Align. F_1	Layer 8 Retr. ρ	Layer 8 Trans. τ	MLM-Perpl. train	MLM-Perpl. dev
0	original	.70	1.00 $_{.00}$	.16 $_{.02}$	.88 $_{.02}$	1.00 $_{.00}$	.97 $_{.01}$	.79 $_{.03}$	9 $_{0.2}$	217 $_{7.8}$
1	lang-pos	.30	.87 $_{.05}$	.33 $_{.13}$	.40 $_{.09}$	.89 $_{.05}$	.39 $_{.15}$	.09 $_{.05}$	9 $_{0.1}$	216 $_{9.0}$
2	shift-special	.66	1.00 $_{.00}$	.15 $_{.02}$	.88 $_{.01}$	1.00 $_{.00}$	.97 $_{.02}$	.63 $_{.13}$	9 $_{0.1}$	227 $_{17.9}$
4	no-random	.68	1.00 $_{.00}$	.19 $_{.03}$	.87 $_{.02}$	1.00 $_{.00}$	.85 $_{.07}$	.82 $_{.04}$	9 $_{0.6}$	273 $_{7.7}$
5	lang-pos;shift-special	.20	.62 $_{.19}$	.22 $_{.19}$	.27 $_{.20}$	.72 $_{.22}$	.27 $_{.21}$	.05 $_{.04}$	10 $_{0.5}$	205 $_{7.6}$
6	lang-pos;no-random	.30	.91 $_{.04}$	.29 $_{.10}$	.36 $_{.12}$	.89 $_{.05}$	.32 $_{.15}$	.25 $_{.12}$	10 $_{0.4}$	271 $_{8.6}$
7	shift-special;no-random	.68	1.00 $_{.00}$	.21 $_{.03}$	.85 $_{.01}$	1.00 $_{.00}$	.89 $_{.06}$	.79 $_{.04}$	8 $_{0.3}$	259 $_{15.6}$
8	lang-pos;shift-special;no-random	.12	.46 $_{.26}$	.09 $_{.09}$	.18 $_{.22}$	.54 $_{.31}$	.11 $_{.11}$	.11 $_{.13}$	10 $_{0.6}$	254 $_{15.9}$
15	overparam	.58	1.00 $_{.00}$	.27 $_{.03}$	.63 $_{.05}$	1.00 $_{.00}$	.97 $_{.01}$	.47 $_{.06}$	2 $_{0.1}$	261 $_{4.5}$
16	lang-pos;overparam	.01	.25 $_{.10}$	.01 $_{.00}$	.01 $_{.00}$	.37 $_{.13}$	.01 $_{.00}$	.00 $_{.00}$	3 $_{0.0}$	254 $_{4.9}$
17	lang-pos;shift-special;no-random;overparam	.00	.05 $_{.02}$	.00 $_{.00}$	.00 $_{.00}$	.05 $_{.04}$	.00 $_{.00}$	.00 $_{.00}$	1 $_{0.0}$	307 $_{7.7}$
3	inv-order	.01	.02 $_{.00}$	.00 $_{.00}$	.01 $_{.00}$	.02 $_{.00}$	.01 $_{.01}$	.00 $_{.00}$	11 $_{0.3}$	209 $_{14.4}$
9	lang-pos;inv-order;shift-special;no-random	.00	.04 $_{.01}$	.00 $_{.00}$	.00 $_{.00}$	.03 $_{.01}$	.00 $_{.00}$	.00 $_{.00}$	10 $_{0.4}$	270 $_{20.1}$
18	untrained	.00	.97 $_{.01}$	.00 $_{.00}$	.00 $_{.00}$	.96 $_{.01}$	.00 $_{.00}$	.00 $_{.00}$	3484 $_{44.1}$	4128 $_{42.7}$
19	untrained;lang-pos	.00	.02 $_{.00}$	.00 $_{.00}$	.00 $_{.00}$	.02 $_{.00}$	.00 $_{.00}$	.00 $_{.00}$	3488 $_{41.4}$	4133 $_{50.3}$
30	knn-replace	.74	1.00 $_{.00}$	.31 $_{.08}$	.88 $_{.00}$	1.00 $_{.00}$	.97 $_{.01}$	.81 $_{.01}$	11 $_{0.3}$	225 $_{12.4}$

Figure 4: Results of multilinguality and model fit for different models

Analysis of the Results

- **Lang-pos** has the largest negative impact.
- Adding more than one modification makes multilinguality go down more.
- Language model quality stays stable on train and dev across models (with an exception of overparameterization).
- **Overparameterization** brings a better-performed language model with low perplexity but less multilingual.

Discussion Questions:

1. Why does layer 0 works better than layer 8 on the word translation task?
2. Why does model 7 (shift-special + no random) perform even better than single modification (model 2, 4)?

Results for Corpora Comparability Property

ID	Description	μ	Layer 0 F_1	Layer 0 ρ	Layer 0 τ	Layer 8 F_1	Layer 8 ρ	Layer 8 τ	Perpl. train	Perpl. dev
0	original	.70	1.00	.16	.88	1.00	.97	.79	9	217
21	no-parallel	.25	.98	.06	.28	.98	.50	.15	14	383
21b	lang-pos;no-parallel	.07	.60	.10	.07	.73	.11	.02	16	456

Figure 5: Results on comparable corpora

- Multilinguality decreases as the training corpus becomes non-parallel.
- Notable that the model quality also decreases when using non-parallel training corpus.

Multilinguality during Training

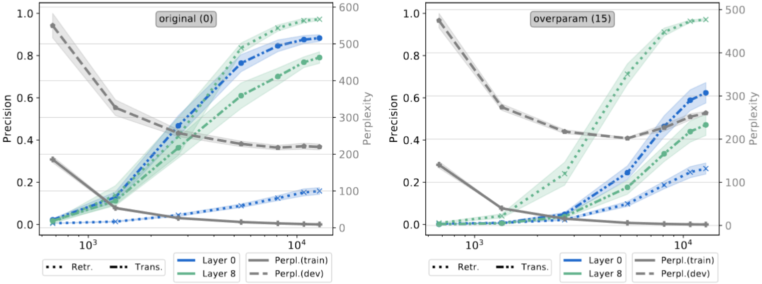

Figure 6: Multilinguality and Model Quality during the Training

- The longer a model is trained, the more multilingual it gets.
- Multilinguality rises later during the training in larger model.
- Multilingual does not start to rise sharply until model fit improvements become flat.
  $\Longrightarrow$ **Trade-off** between good generalization and high degree of mutilinguality

Improvement of Multilinguality

Motivation

- One of the conclusions from the previous experiment is that: replacing some masked tokens with random words during the MLM pretraining can boost multilinguality
- **Further Induction:** Replacing masked tokens with semantically similar words from other languages could further improve the multilinguality
- **Idea:** Introduce a fourth masking option to the MLM pretraining

knn-replace method

- **Retrieve similar words** from another language by **mapping-based** approach for bilingual embedding:
  1. Train **static fastText**[8] monolingual embeddings for both languages on their training set.
  2. Project them into a common space using **VecMap**[9]
- **Replace 30% of the masked tokens** with nearest neighbors from the other language

[8] Bojanowski et al. (2017)
[9] Artetxe et al. (2018)

Results from the experimental setup

ID	Description	Mult.-score μ	Layer 0 Align. F_1	Layer 0 Retr. ρ	Layer 0 Trans. τ	Layer 8 Align. F_1	Layer 8 Retr. ρ	Layer 8 Trans. τ	MLM-Perpl. train	MLM-Perpl. dev
0	original	.70	1.00 $_{.00}$	.16 $_{.02}$	.88 $_{.02}$	1.00 $_{.00}$	.97 $_{.01}$	.79 $_{.03}$	9 $_{0.2}$	217 $_{7.8}$
30	knn-replace	.74	1.00 $_{.00}$	.31 $_{.08}$	.88 $_{.00}$	1.00 $_{.00}$	.97 $_{.01}$	.81 $_{.01}$	11 $_{0.3}$	225 $_{12.4}$

Figure 7: Results of knn-replace method

- The model with knn-replace method **outperforming** the original model in multilinguality score.
- During the training, the multilingual score of the model with knn-replace achieves higher score **earlier**.

Results from Real Data

ID	Description	ENG	DEU	HIN
0-base	original	**.75** $_{.00}$	.57 $_{.02}$	.45 $_{.01}$
3-base	inv-order[DEU]	**.75** $_{.00}$	.41 $_{.01}$	.46 $_{.04}$
8-base	lang-pos;shift-special;no-random	.74 $_{.00}$	.37 $_{.02}$	.38 $_{.02}$
30-base	knn-replace	.74 $_{.01}$	**.61** $_{.01}$	**.54** $_{.00}$
mBERT	Results by (Hu et al., 2020)	.81	.70	.59

Figure 8: Accuracy on XNLI test

- **Setup**: Train a multilingual BERT of **three** languages (English, German and Hindi) on about 3GB of training corpora sampled from Wikipedia.
- **Evaluation**: Finetune the pretrained mBERT on English XNLI then zero-shot evaluate on German and Hindi.
- **Results**: knn-replace model exhibits strong ability to boost the degree of multilinguality.
- ***Discussion Question**: Why does the accuracy of English decrease with knn-replace?

Conclusions

In this presentation,

- We take an overview of some **core concepts** in the multilingual representation, such as multilingual embedding, multilingual models and multilinguality.
- We know about some **metrics** to measure the multilinguality of a model.
- Through the experiment results in the paper, it can be concluded that **4 architectural properties** and **2 linguistic properties** are essential for model's multilinguality.
- Based on the insights from the experiment, the **knn-replace** method is proposed to improve the model's multilinguality.

References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Wang, Z., Mayhew, S., Roth, D., et al. (2019). Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Wu, S., Conneau, A., Li, H., Zettlemoyer, L., and Stoyanov, V. (2019). Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Thanks for your attention!