Multilingual BERT

Bin Liang
Harbin Institute of Technology, Shenzhen

内容提要

- BERT
- Multilingual BERT
- Downstream Tasks

“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ” NAACL 2019

https://github.com/google-research/bert

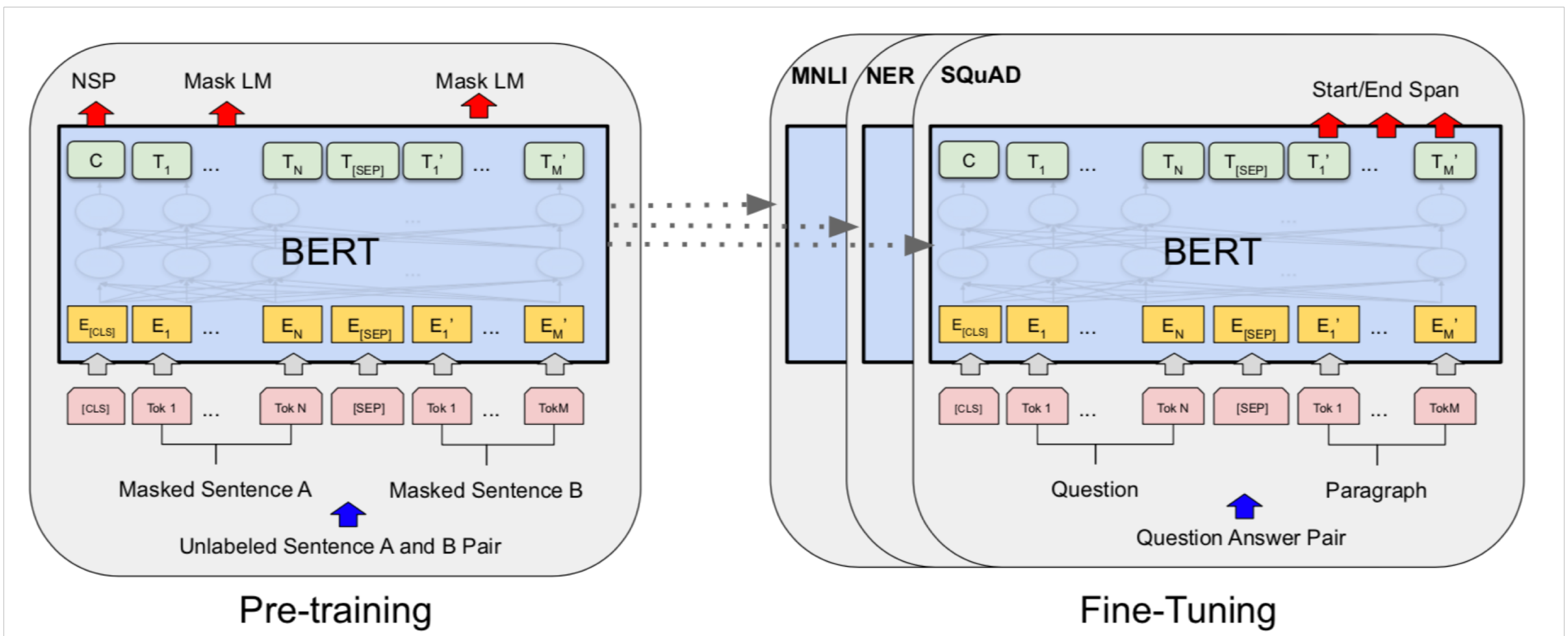
“How multilingual is Multilingual BERT” arXiv 2019

https://github.com/google-research/bert/blob/master/multilingual.md

BERT

- **Pre-training:** the BERT model is trained on unlabeled data over different pre-training tasks.

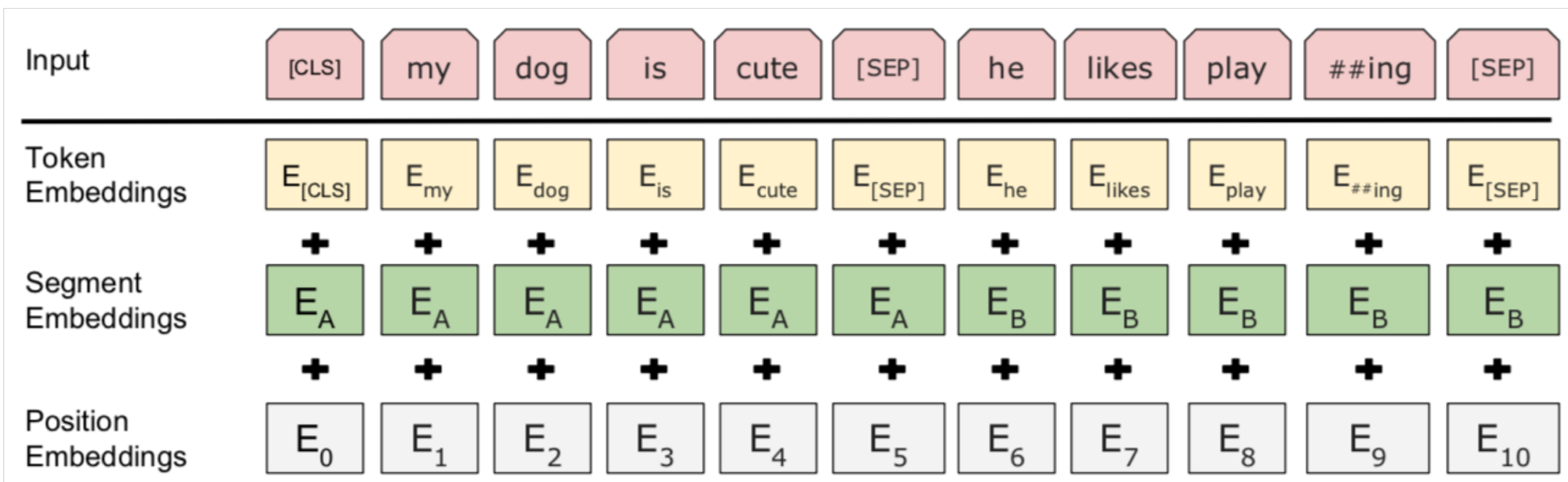
- **Fine-tuning:** the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks.



BERT

- **Masked LM:** the man **[MASK]** up , put his basket on [MASK] [MASK] [MASK] ' s head. (80%+10%+10%)

- **NSP:** [CLS] the man went to the store [SEP] he bought a gallon of milk [SEP]. (50%+50%)



Multilingual BERT

- `BERT-Base, Multilingual Cased (New, recommended)` : 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Multilingual Uncased (Orig, not recommended)` : 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
- `BERT-Base, Chinese` : Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

Results (English: 84.2)对于语料资源大的语言，多语言模型的表现不如单语言模型

System	English	Chinese	Spanish	German	Arabic	Urdu
XNLI Baseline - Translate Train	73.7	67.0	68.8	66.5	65.8	56.6
XNLI Baseline - Translate Test	73.7	68.3	70.7	68.7	66.8	59.3
BERT - Translate Train Cased	**81.9**	**76.6**	**77.8**	**75.9**	**70.7**	61.6
BERT - Translate Train Uncased	81.4	74.2	77.3	75.2	70.5	61.7
BERT - Translate Test Uncased	81.4	70.1	74.9	74.4	70.4	**62.1**
BERT - Zero Shot Uncased	81.4	63.8	74.3	70.5	62.1	58.3

Multilingual BERT

System	Chinese
XNLI Baseline	67.0
BERT Multilingual Model	74.2
BERT Chinese-only Model	77.2

训练和维护数十种单语言模型是不可行的。如果你的目标是使用英语和中文以外的语言最大限度地提高性能，那么从我们的多语言模型开始，对你感兴趣的语言数据进行额外的预训练是有益的。

Multilingual BERT

The multilingual model does **not** require any special consideration or API changes. We did update the implementation of `BasicTokenizer` in `tokenization.py` to support Chinese character tokenization, so please update if you forked it. However, we did not change the tokenization API.

To test the new models, we did modify `run_classifier.py` to add support for the XNLI dataset. This is a 15-language version of MultiNLI where the dev/test sets have been human-translated, and the training set has been machine-translated.

To run the fine-tuning code, please download the XNLI dev/test set and the XNLI machine-translated training set and then unpack both .zip files into some directory `$XNLI_DIR`.

To run fine-tuning on XNLI. The language is hard-coded into `run_classifier.py` (Chinese by default), so please modify `XnliProcessor` if you want to run on another language.

This is a large dataset, so this will training will take a few hours on a GPU (or about 30 minutes on a Cloud TPU). To run an experiment quickly for debugging, just set `num_train_epochs` to a small value like `0.1`.

https://github.com/google-research/bert/blob/master/multilingual.md

Downstream Tasks

Cross-lingual Task

M-Bert特别适合across language任务的探索研究

优势：某一特定任务（例如NER）使用一种语言对模型进行微调，并使用另一种语言进行评估，体现M-Bert的通用性。

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

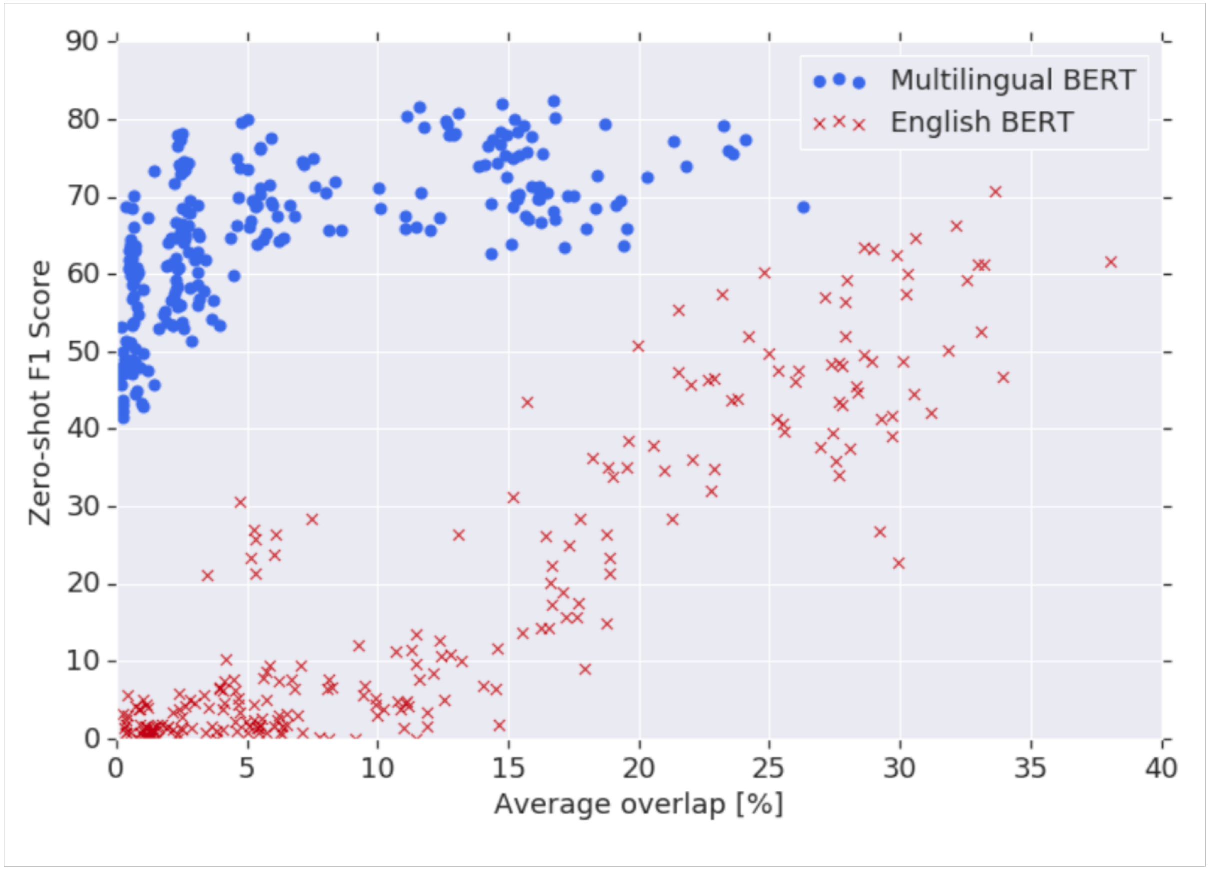
Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

Downstream Tasks

Cross-lingual Task--NER



降低不同语言词典重叠的依赖

Downstream Tasks

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

	Corrected	Transliterated
Train on monolingual HI+EN		
M-BERT	86.59	50.41
Ball and Garrette (2018)	—	77.40
Train on code-switched HI/EN		
M-BERT	90.56	85.64
Bhat et al. (2018)	—	90.53

Table 6: M-BERT's POS accuracy on the code-switched Hindi/English dataset from Bhat et al. (2018), on script-corrected and original (transliterated) tokens, and comparisons to existing work on code-switch POS.

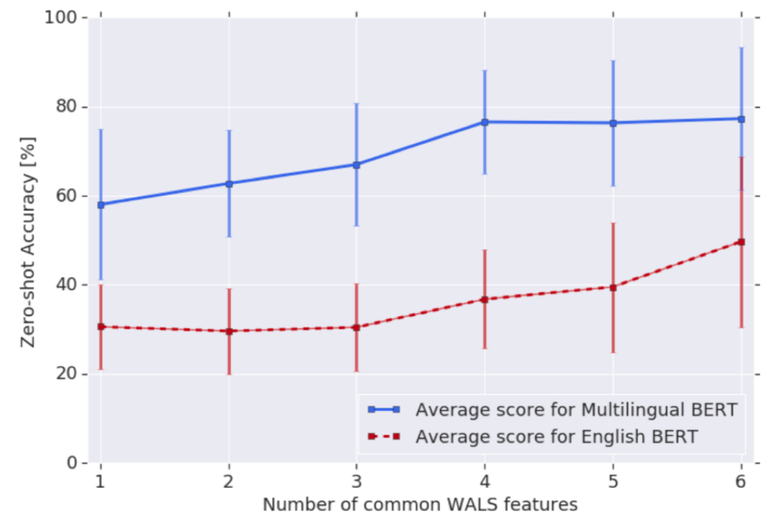


Figure 2: Zero-shot POS accuracy versus number of common WALS features. Due to their scarcity, we exclude pairs with no common features.

Downstream Tasks

- **多语言：**某一种语言缺乏大量预训练数据或严谨的微调数据；
- **跨语言：**一种语言到另外一种语言的迁移和学习；
- **跨任务：**使用一个通用任务进行预训练，使用特定任务微调，最后使用目标任务评估；
- **细粒度**：target、aspect、relation的zero-shot learning等

Thank you

Q&A