A Supplemental Material

A.1 Internal Representation of multi-BERT

The architecture of multi-BERT is a Transformer encoder (Vaswani et al., 2017). While fine-tuning on SQuAD-like dataset, the bottom layers of multi-BERT are initialized from Google-pretrained parameters, with an added output layer initialized from random parameters. Tokens representations from the last layer of bottom-part of multi-BERT are inputs to the output layer and then the output layer outputs a distribution over all tokens that indicates the probability of a token being the START/END of an answer span.

A.1.1 Cosine Similarity

As all translated versions of SQuAD/DRCD are parallel to each other. Given a source-target language pair, we calculate cosine similarity of the mean pooling of tokens representation within corresponding answer-span as a measure of how much they look like in terms of the internal representation of multi-BERT. The results are shown in Fig. 1.

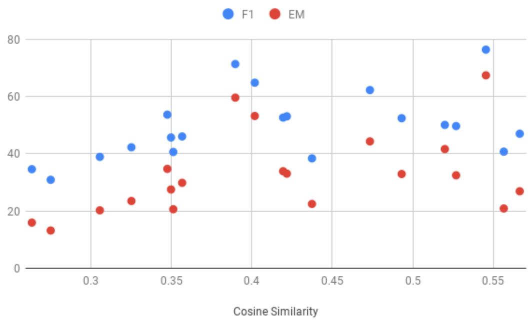

Figure 1: The relation of cosine similarity of answer words with EM/F1 scores in red and blue respectively. Each point represents a source-target language pair of datasets.

A.1.2 SVCCA

Singular Vector Canonical Correlation Analysis (SVCCA) is a general method to compare the correlation of two sets of vector representations. SVCCA has been proposed to compare learned representations across language models (Saphra and Lopez, 2018). Here we adopt SVCCA to measure the linear similarity of two sets of representations in the same multi-BERT from different translated datasets, which are parallel to each other. The results are shown in Fig 2.

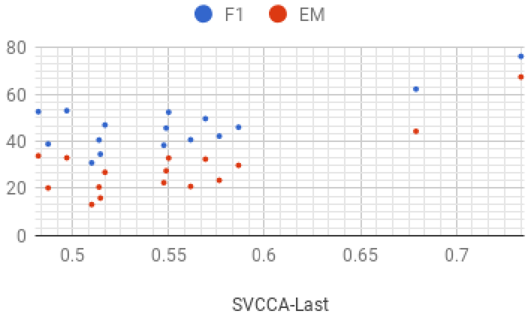

Figure 2: The relation of SVCCA similarity with EM/F1 scores in red and blue respectively. Each point represents a source-target language pair of datasets.

A.2 Improve Transfering

In the paper, we show that internal representations of multi-BERT are linear-mappable to some extent between different languages. This implies that multi-BERT model might encode semantic and syntactic information in language-agnostic ways and explains how zero-shot transfer learning could be done.

To take a step further, while transfering model from source dataset to target dataset, we align representations in two proposed way, to improve performance on target dataset.

A.2.1 Linear Mapping Method

Algorithms proposed in (Lample et al., 2018; Artetxe et al., 2018; Zhou et al., 2019) to unsupervisedly learn linear mapping between two sets of embeddings are used here to align representations of source (training data) to those of target. We obtain the mapping generated by embeddings from one specific layer of pre-trained multi-BERT then we apply this mapping to transform the internal representations of multi-BERT while fine-tuning on training data.

A.2.2 Adversarial Method

In Adversarial Method, we add an additional transform layer to transform representations and a discrimination layer to discriminate between transformed representations from source language (training set) and target language (development set). And the GAN loss is applied in the total loss of fine-tuning.

A.2.3 Discussion

As table 1 shows, there are no improvements among above methods. Some linear mapping

Approach	EM	F1
MUSE(Lample et al., 2018)	33.03	49.48
DeMa(Zhou et al., 2019)	55.64	72.59
Vecmap(Artetxe et al., 2018)	14.05	24.83
GAN-layer 8	54.26	71.04
GAN-layer 11	60.47	76.14

Table 1: EM/F1 scores on DRCD dev-set.

methods even causes devastating effect on EM/F1 scores.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Chunting Zhou, Xuezhe Ma, Di Wang, and Graham Neubig. 2019. Density matching for bilingual word embedding. *CoRR*, abs/1904.02343.