Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model

Tsung-Yuan Hsu[*] **Chi-liang Liu**[*] **Hung-yi Lee**
Graduate Institute of Communication Engineering
National Taiwan University
{sivia89024, liangtaiwan1230, tlkagkb93901106}@gmail.com

Abstract

Because it is not feasible to collect training data for every language, there is a growing interest in cross-lingual transfer learning. In this paper, we systematically explore zero-shot cross-lingual transfer learning on reading comprehension tasks with a language representation model pre-trained on multi-lingual corpus. The experimental results show that with pre-trained language representation zero-shot learning is feasible, and translating the source data into the target language is not necessary and even degrades the performance. We further explore what does the model learn in zero-shot setting[0].

1 Introduction

Reading Comprehension (RC) has become a central task in natural language processing, with great practical value in various industries. In recent years, many large-scale RC datasets in English (Hermann et al., 2015; Hewlett et al., 2016; Rajpurkar et al., 2016; Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017; Rajpurkar et al., 2018) have nourished the development of numerous powerful and diverse RC models (Seo et al., 2016; Hu et al., 2018; Wang et al., 2017; Clark and Gardner, 2018; Huang et al., 2017). The state-of-the-art model (Devlin et al., 2018) on SQuAD, one of the most widely used RC benchmarks, even surpasses human-level performance. Nonetheless, RC on languages other than English has been limited due to the absence of sufficient training data. Although some efforts have been made to create RC datasets for Chinese (He et al., 2018; Shao et al., 2018) and Korean (Seungyoung Lim, 2018), it is not feasible to collect RC datasets for every language since annotation efforts to collect a new RC dataset are often far from trivial. Therefore, the setup of transfer learning, especially zero-shot learning, is of extraordinary importance.

Existing methods (Asai et al., 2018) of cross-lingual transfer learning on RC datasets often count on machine translation (MT) to translate data from source language into target language, or vice versa. These methods may not require a well-annotated RC dataset for the target language, whereas a high-quality MT model is needed as a trade-off, which might not be available when it comes to low-resource languages.

In this paper, we leverage pre-trained multilingual language representation, for example, BERT learned from multilingual un-annotated sentences (multi-BERT), in cross-lingual zero-shot RC. We fine-tune multi-BERT on the training set in source language, then test the model in target language, with a number of combinations of source-target language pair to explore the cross-lingual ability of multi-BERT. Surprisingly, we find that the models have the ability to transfer between low lexical similarity language pair, such as English and Chinese. Recent studies (Lample and Conneau, 2019; Devlin et al., 2018; Wu and Dredze, 2019) show that cross-lingual language models have the ability to enable preliminary zero-shot transfer on simple natural language understanding tasks, but zero-shot transfer of RC has not been studied. To our knowledge, this is the first work systematically exploring the cross-lingual transferring ability of multi-BERT on RC tasks.

2 Zero-shot Transfer with Multi-BERT

Multi-BERT has showcased its ability to enable cross-lingual zero-shot learning on the natural language understanding tasks including XNLI (Conneau et al., 2018), NER, POS, Dependency Parsing, and so on. We now seek to know if a pre-trained multi-BERT has ability to solve RC tasks in the zero-shot setting.

[*] Equal contribution

[0] All the modifications of existing corpora used in this paper would be released in https://github.com/ntu-spml-lab/artificial-reading-comprehension-datasets

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5932–5939,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

2.1 Experimental Setup and Data

We have training and testing sets in three different languages: English, Chinese and Korean. The English dataset is SQuAD (Rajpurkar et al., 2016). The Chinese dataset is DRCD (Shao et al., 2018), a Chinese RC dataset with 30,000+ examples in the training set and 10,000+ examples in the development set. The Korean dataset is KorQuAD (Seungyoung Lim, 2018), a Korean RC dataset with 60,000+ examples in the training set and 10,000+ examples in the development set, created in exactly the same procedure as SQuAD. We always use the development sets of SQuAD, DRCD and KorQuAD for testing since the testing sets of the corpora have not been released yet.

Next, to construct a diverse cross-lingual RC dataset with compromised quality, we translated the English and Chinese datasets into more languages, with Google Translate[1]. An obvious issue with this method is that some examples might no longer have a recoverable span. To solve the problem, we use fuzzy matching[2] to find the most possible answer, which calculates minimal edit distance between translated answer and all possible spans. If the minimal edit distance is larger than min(10, lengths of translated answer - 1), we drop the examples during training, and treat them as noise when testing. In this way, we can recover more than 95% of examples. The following generated datasets are recovered with same setting.

The pre-trained multi-BERT is the official released one[3]. This multi-lingual version of BERT were pre-trained on corpus in 104 languages. Data in different languages were simply mixed in batches while pre-training, without additional effort to align between languages. When fine-tuning, we simply adopted the official training script of BERT, with default hyperparameters, to fine-tune each model until training loss converged.

Model	Train-set	EM	F1
(a) Shao et al. 2018	Chinese	-	53.78
(b) QANet[3]	Chinese	66.10	78.10
(c) English-BERT	Chinese	65.00	76.96
(d) Chinese-BERT	Chinese	82.00	89.10
(e) multi-BERT	Chinese	81.24	88.68
(f) multi-BERT	English	63.31	78.82
(g) multi-BERT	English +Chinese	82.63	90.10

Table 1: EM/F1 scores over Chinese testing set.

	Test		
Train	English	Chinese	Korean
En	**81.2/88.6**	63.3/78.8	49.2/69.3
Zh	34.1/53.8	**81.2/88.7**	56.4/78.2
Kr	58.5/68.4	73.4/82.7	**69.41/89.3**
En-Fr	67.5/76.4	56.5/72.5	37.2/56.3
En-Zh	59.7/71.4	**61.4/78.8**	49.0/72.7
En-Jp	53.3/64.9	62.4/76.7	50.4/72.0
En-Kr	41.7/50.1	56.7/71.6	**47.1/70.8**
Zh-En	**26.6/44.1**	57.7/71.1	40.5/59.5
Zh-Fr	23.4/39.8	44.9/62.0	39.6/59.9
Zh-Jp	25.5/42.6	60.9/72.4	44.9/65.7
Zh-Kr	26.5/42.2	58.2/69.5	**47.4/67.7**

Table 2: EM/F1 score of multi-BERTs fine-tuned on different training sets and tested on different languages (En: English, Fr: French, Zh: Chinese, Jp: Japanese, Kr: Korean, xx-yy: translated from xx to yy). The text in bold means training data language is the same as testing data language.

2.2 Experimental Results

Table 1 shows the result of different models trained on either Chinese or English and tested on Chinese. In row (f), multi-BERT is fine-tuned on English but tested on Chinese, which achieves competitive performance compared with QANet trained on Chinese. We also find that multi-BERT trained on English has relatively lower EM compared with the model with comparable F1 scores. This shows that the model learned with zero-shot can roughly identify the answer spans in context but less accurate. In row (c), we fine-tuned a BERT model pre-trained on English monolingual corpus (English BERT) on Chinese RC training data directly by appending fastText-initialized Chinese word embeddings to the original word embeddings of English-BERT. Its F1 score is even lower than that of zero-shot transferring multi-BERT (rows (c) v.s. (e)). The result implies multi-BERT does acquire better cross-lingual capability through pre-training on multilingual corpus.

Table 2 shows the results of multi-BERT fine-tuned on different languages and then tested on English , Chinese and Korean. The top half of the table shows the results of training data without translation. It is not surprising that when the training and testing sets are in the same language, the best results are achieved, and multi-BERT shows transfer capability when training and testing sets are in different languages, especially between Chinese and Korean.

In the lower half of Table 2, the results are ob-

[1] https://translate.google.com/

[2] https://github.com/taleinat/fuzzysearch

[3] https://github.com/google-research/bert

tained by the translated training data. First, we found that when testing on English and Chinese, translation always degrades the performance (En v.s. En-XX, Zh v.s. Zh-XX). Even though we translate the training data into the same language as testing data, using the untranslated data still yield better results. For example, when testing on English, the F1 score of the model training on Chinese (Zh) is 53.8, while the F1 score is only 44.1 for the model training on Zh-En. This shows that translation degrades the quality of data. There are some exceptions when testing on Korean. Translating the English training data into Chinese, Japanese and Korean still improve the performance on Korean. We also found that when translated into the same language, the English training data is always better than the Chinese data (En-XX v.s. Zh-XX), with only one exception (En-Fr v.s. Zh-Fr when testing on KorQuAD). This may be because we have less Chinese training data than English. These results show that the quality and the size of dataset are much more important than whether the training and testing are in the same language or not.

2.3 Discussion

2.3.1 The Effect of Machine Translation

Table 2 shows that fine-tuning on un-translated target language data achieves much better performance than data translated into the target language. Because the above statement is true across all the languages, it is a strong evidence that translation degrades the performance.We notice that the translated corpus and untranslated corpus are not the same. This may be another factor that influences the results. Conducting an experiment between un-translated and back-translated data may deal with this problem.

2.3.2 The Effect of Other Factors

Here we discuss the case that the training data are translated. We consider each result is affected by at least three factors: (1) training corpus, (2) data size, (3) whether the source corpus is translated into the target language. To study the effect of data-size, we conducted an extra experiment where we down-sampled the size of English data to be the same as Chinese corpus, and used the down-sampled corpus to train. Then We carried out one-way ANOVA test and found out the significance of the three factors are ranked as below: (1) > (2) >> (3). The analysis supports that the characteristics of training data is more important than translated into target language or not. Therefore, although translation degrades the performance, whether translating the corpus into the target language is not critical.

Train	Test	EM	F1
English	English-permuted	1.25	11.54
English	Chinese-permuted	5.02	17.49
Chinese	Chinese-permuted	8.91	25.67
Chinese	Chinese	81.24	88.68

Table 3: EM/F1 scores over artificially created unseen languages (English-permuted and Chinese-permuted).

3 What Does Zero-shot Transfer Model Learn?

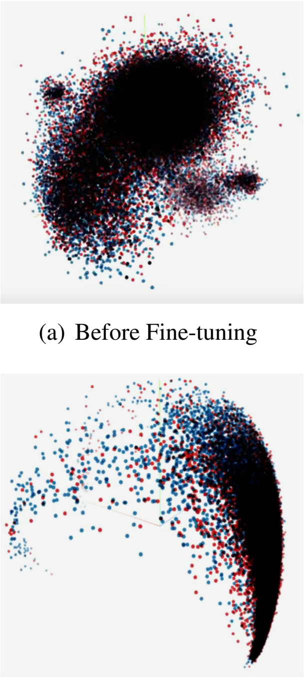
3.1 Unseen Language Dataset

It has been shown that extractive QA tasks like SQuAD may be tackled by some language independent strategies, for example, matching words in questions and context (Weissenborn et al., 2017). Is zero-shot learning feasible because the model simply learns this kind of language independent strategies on one language and apply to the other?

To verify whether multi-BERT largely counts on a language independent strategy, we test the model on the languages unseen during pre-training. To make sure the languages have never been seen before, we artificially make unseen languages by permuting the whole vocabulary of existing languages. That is, all the words in the sentences of a specific language are replaced by other words in the same language to form the sentences in the created unseen language. It is assumed that if multi-BERT used to find answers by language independent strategy, then multi-BERT should also do well on unseen languages. Table 4 shows that the performance of multi-BERT drops drastically on the dataset. It implies that multi-BERT might not totally rely on pattern matching when finding answers.

3.2 Embedding in Multi-BERT

PCA projection of hidden representations of the last layer of multi-BERT before and after fine-tuning are shown in Fig. 1. The red points represent Chinese tokens, and the blue points are for English. The results show that tokens from different languages might be embedded into the same



(a) Before Fine-tuning

(b) After Fine-tuning

Figure 1: PCA visualization of hidden representations from the 12-th transformer layer of multi-BERT before and after fine-tuning on English. The red points represent Chinese tokens, and the blue points are for English.

Train	Mix Lang.	EM	F1	Sub.
English	None	81.17	88.63	0%
English	Chinese	68.79	79.18	31%
English	French	65.7	77.43	61%
English	Japanese	63.32	74.06	30%
English	Korean	39.93	63.46	32%

Table 4: EM/F1 scores on artificial code-switching datasets generated by replacing some of the words in English dataset with synonyms in another languages. (Sub. is the substitution ratio of the dataset)

Source	Example
pred:	second 法律 of 熱力學 (Zh)
gt:	second law of thermodynamics
pred:	エレクトリック motors (Jp)
gt:	electric motors
pred:	fermionic nature des lectrons (Fr)
gt:	fermionic nature of electrons
pred:	the 차이점 in 잠재력 에너지 (Kr)
gt:	the difference in potential energy

Table 5: Answers inferenced on code-switching dataset. The predicted answers would be the same as the ground truths (gt) if we translate every word into English.

space with close spatial distribution. Even though during the fine-tuning only the English data is used, the embedding of the Chinese token changed accordingly. We also quantitatively evaluate the similarities between the embedding of the languages. The results can be found in the Appendix.

3.3 Code-switching Dataset

We observe linguistic-agnostic representations in the last subsection. If tokens are represented in a language-agnostic way, the model may be able to handle code-switching data. Because there is no code-switching data for RC, we create artificial code-switching datasets by replacing some of the words in contexts or questions with their synonyms in another language. The synonyms are found by word-by-word translation with given dictionaries. We use the bilingual dictionaries collected and released in facebookresearch/MUSE GitHub repository. We substitute the words if and only if the words are in the bilingual dictionaries.

Table 4 shows that on all the code-switching datasets, the EM/F1 score drops, indicating that the semantics of representations are not totally disentangled from language. However, the examples of the answers of the model (Table 5) show that multi-BERT could find the correct answer spans although some keywords in the spans have been translated into another language.

3.4 Typology-manipulated Dataset

There are various types of typology in languages. For example, in English the typology order is subject-verb-object (SVO) order, but in Japanese and Korean the order is subject-object-verb (SOV). We construct a typology-manipulated dataset to examine if the typology order of the training data influences the transfer learning results. If the model only learns the semantic mapping between different languages, changing English typology order from SVO to SOV should improve the transfer ability from English to Japanese. The method used to generate datasets is the same as Ravfogel et al. 2019.

The source code is from a GitHub repository named Shaul1321/rnn_typology, which labels given sentences to CoNLL format with StanfordCoreNLP and then re-arranges them greedily.

Table 6 shows that when we change the English typology order to SOV or OSV order, the perfor-

Train/Test	English	Chinese	Korean
En	81.2/88.6	63.3/78.8	49.2/69.3
En-SOV	78.4/86.5	62.8/78.3	47.6/70.4
En-VOS	79.4/87.1	59.1/74.6	46.2/67.0
En-VSO	79.4/87.1	60.9/76.8	44.2/65.4
En-OSV	78.9/86.9	63.5/78.0	49.0/70.7
En-OVS	73.6/82.5	57.6/72.1	45.8/67.0

Table 6: EM/F1 scores over artificially created typology-manipulated dataset.

mance on Korean is improved and worsen on English and Chinese, but very slightly. The results show that the typology manipulation on the training set has little influence. It is possible that multi-BERT normalizes the typology order of different languages to some extent.

4 Conclusion

In this paper, we systematically explore zero-shot cross-lingual transfer learning on RC with multi-BERT. The experimental results on English, Chinese and Korean corpora show that even when the languages for training and testing are different, reasonable performance can be obtained. Furthermore, we created several artificial data to study the cross-lingual ability of multi-BERT in the presence of typology variation and code-switching. We showed that only token-level pattern matching is not sufficient for multi-BERT to answer questions and typology variation and code-switching only caused minor effects on testing performance.

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.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual Extractive Reading Comprehension by Runtime Machine Translation. *arXiv e-prints*, page arXiv:1809.03275.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. *CoRR*, abs/1608.03542.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4099–4106. International Joint Conferences on Artificial Intelligence Organization.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *CoRR*, abs/1711.07341.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

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

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, and Rangan and Majumder. 2016. Ms marco: A human generated machine reading comprehension dataset.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of rnns with synthetic variations of natural languages. *CoRR*, abs/1903.06400.

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

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Jooyoul Lee Seungyoung Lim, Myungji Kim. 2018. KorQuAD: Korean QA Dataset for Machine Comprehension.

Chih-Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: a chinese machine reading comprehension dataset. *CoRR*, abs/1806.00920.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

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.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada. Association for Computational Linguistics.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural qa as simple as possible but not simpler.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *CoRR*, abs/1904.09077.

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

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. 2.

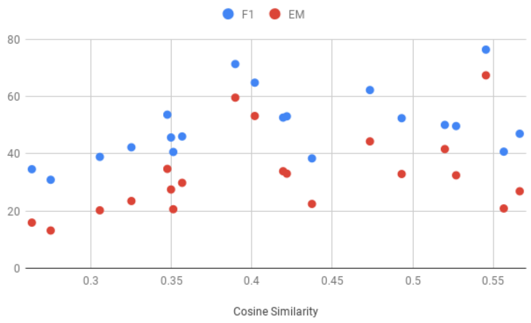


Figure 2: The relation of cosine similarity of answer words with EM/F1 scores in red and blue respectively. Each point represents a source-target language pair of datasets.

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 3.

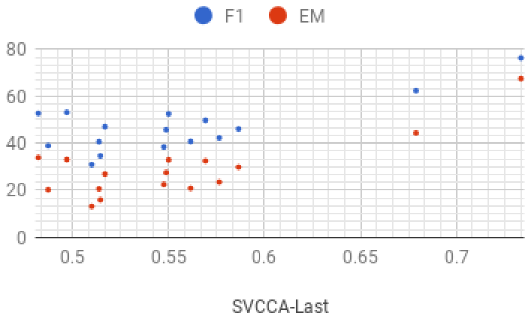


Figure 3: The relation of SVCCA similarity with EM/F1 scores in red and blue respectively. Each point represents a source-target language pair of datasets.

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 7 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 7: EM/F1 scores on DRCD dev-set.

methods even causes devastating effect on EM/F1 scores.