The Effects of Transfer Learning between Languages in Machine Reading Comprehension

Son Tran, Khoi Le, Uyen Le and **Matthew Kretchmar**
Computer Science Department
Denison University, United States
{tran_s2,le_k2,le_u1,kretchmar}@denison.edu

Abstract

Machine Reading Comprehension (MRC) is a challenging task in Natural Language Processing that requires computers to have rigorous linguistic understanding in order to predict the answer to a question in a related context. Given the complication of the task, researchers have been interested in applying the concept of Transfer Learning to MRC to exploit the powers of pre-trained knowledge and cross-linguistic similarities. In this paper, by utilizing pre-trained multilingual networks, we introduce a simple but effective pipeline for fine-tuning Machine Reading Comprehension models in low-resource languages that circumvents the downsides of a high-quality data shortage.

1 Introduction

Machine Reading Comprehension (MRC) is a fundamental subfield of Natural Language Processing (NLP), in which the computer simulates a human question answering mechanism by reading passages and answering related questions. In this process, the computer tries to infer the syntax and semantic relationships between entities named in the text or explore the implications about the world through the text. Challenging though its task is, MRC is critical in transforming the way humans accumulate their knowledge in an age of information and leveraging text knowledge in developing machine learning models. Nowadays, with the advances of transfer learning powered by pre-trained language models, MRC models have been able to achieve state-of-the-art performances on many benchmarks and surpass human-level parity.

Transfer learning in the context of NLP is the ability to leverage the knowledge from training a deep learning model on a very large corpus using unsupervised tasks and use such knowledge to perform different NLP functions on a different datasets. Accordingly, pre-trained models are saved networks that were previously trained on a large amount of data to obtain universal language representations that can later become beneficial in training specific tasks. This approach helps researchers take advantage of the huge amount of learned knowledge without having to train their models from scratch. All what researchers need to do is fine-tuning the pre-trained models on their specific tasks.

Another aspect of transfer learning in NLP is cross language pre-training, which utilizes the extensive knowledge already learned in a high-resource language like English to leverage the other low- and medium-resource languages such as German, Korean, Vietnamese, etc. A prevalent approach in this type of transfer learning is to develop multilingual models that can handle several languages simultaneously. These pre-trained multilingual models, previously trained on a text corpus of over 100 languages (Conneau et al., 2020); (Devlin et al., 2018), circumvent the problem of having to train a monolingual model for every single language while preserve high performances on different tasks in multilingual settings. Moreover, they have shown surprisingly competitive results when compared to monolingual models in low-resource languages. Given that there is only a limited number of high quality works contributed to the development of MRC in low-resource languages, these pre-trained multilingual models are significant in the growth of future research in MRC in such languages.

It is still open to question whether the growth of MRC in low-resource languages will ever reach the English state-of-the-art considering the lack of high quality data. This motivates us to explore the use of transfer learning in improving the performances of multilingual models on low-resource languages. Particularly, we are interested in whether additionally training a multilingual language model on large dataset of resource-rich lan-

	Context	Question
German	Obwohl die Vereinigten Staaten wie auch viele Staaten des Commonwealth Erben des britischen Common Laws sind, setzt sich das amerikanische Recht bedeutend davon ab. Dies rührt größtenteils von dem langen Zeitraum her, in dem sich das amerikanische Recht unabhängig vom Britischen entwickelt hat. (Although the United States, like many Commonwealth countries, is the heir to British common law, American law differs significantly. This stems in large part from the long period in which American law has developed independently of British law.)	Von welchem Gesetzt stammt das Amerikanische ab? (From which law does the American come from?)
French	Il s ’ agit d ’ une réplique de la grotte de Lourdes , en France , où la Vierge Marie est apparue à Sainte Bernadette Soubirous en 1858 . À la fin de la route principale (et dans une ligne directe qui relie 3 statues et le dôme d ’ or) , simple , statue en pierre moderne de Mary . (It is a replica of the grotto in Lourdes , France , where the Virgin Mary appeared to Saint Bernadette Soubirous in 1858 . At the end of the main road (and in a direct line that connects 3 statues and the Golden Dome), simple, modern stone statue of Mary.)	A qui la Vierge Marie apparaîtrait - elle en 1858 à Lourdes France? (To whom would the Virgin Mary appear in 1858 in Lourdes France?)
Vietnamese	Ông là học trò, cộng sự của Chủ tịch Hồ Chí Minh. Ông có tên gọi thân mật là Tô, đây từng là bí danh của ông. Ông còn có tên gọi là Lâm Bá Kiệt khi làm Phó chủ nhiệm cơ quan Biện sự xứ tại Quế Lâm (Chủ nhiệm là Hồ Học Lãm). (He was a student and associate of President Ho Chi Minh. His nickname is To, this used to be his alias. He was also known as Lam Ba Kiet when he worked as Deputy Director of the Department of Foreign Affairs in Guilin (the director was Ho Hoc Lam).)	Tên gọi nào được Phạm Văn Đồng sử dụng khi làm Phó chủ nhiệm cơ quan Biện sự xứ tại Quế Lâm? (What name is Pham Van Dong using when working as Deputy Head of the Department of Justice in Que Lam?)
Korean	서일본 여객철도 어번 네트워크의 한 축을 구성하고 있다. 오사카 시가지 남부의 철도 교통 터미널인 덴노지에서 남쪽으로 철도를 늘려 오사카 남부의 각 도시를 경유하여 와카야마 에 이르는 노선이다. 한와 선으로부터 서측에 자리한 난카이 전기 철도 본선 · 공항선과 경합하고 있는 것 이외에 신이마미야 역 ~모즈 역 간에서는 난카이 전기 철도 고야 선과, 오사카 시내에서는 오사카 시 교통국 미도스지 · 다니마치 선이나 한카이 전기 궤도와도 경합이 이루어지고 있다. (It forms one axis of the Urban Network of West Japan Railway Company. From Tennoji, a railway transportation terminal in the southern part of Osaka City, the railway is extended to the south and reaches Wakayama via each city in southern Osaka. In addition to competing with the Nankai Electric Railway Main Line and Airport Line located to the west of the Hanwa Line, it also competes with the Nankai Electric Railway Koya Line between Shin-Imamiya Station and Mozu Station, and also with the Osaka City Transportation Bureau Midosuji/ Tanimachi Line and Hankai Electric Railway in Osaka City. Competition is taking place.)	한와 선은 오사카 덴노지에서 어디까지 가는 노선인가? (Where does the Hanwa Line run from Osaka Tennoji?)

Table 1: Examples of Question Answering in four languages

guages like English before training it on the target language can enhance can enhance its performance.

In this paper, we propose a new pipeline for fine-tuning multilingual models in low-resource language that can significantly alleviate the problem of data shortage in low-resource languages and improve the performances of models on all four experimented high-quality MRC datasets.

2 Experiments

2.1 Dataset

	Train	Dev
SQuAD	87,599	10,570
GermanQuAD	11,518	2,204
FQuAD	50,741	5,668
ViQuAD	18,579	2,285
KorQuAD	60,407	5,774

Table 2: Number of questions in our chosen five Question Answering datasets

SQuAD 1.1 (Rajpurkar et al., 2016) is one of the most influential Question Answering datasets as it is a large-scale dataset with samples of very high-quality. With the introduction of this large-scale dataset, Machine Reading Comprehension research community has witnessed great progress on many aspects:

- New tasks of Machine Reading Comprehension with real applications are introduced, such as Open Domain Question Answering (Chen et al., 2017) and Conversational Question Answering (Reddy et al., 2019).
- Many works (Sugawara et al., 2018; Jia and Liang, 2017) of high quality in analyzing MRC benchmarks are carried out.
- Many state-of-the-art MRC systems (Zhang et al., 2021) are developed.

Following the success of SQuAD, Question Answering datasets are introduced in other languages such as German (Möller et al., 2021), French (d'Hoffschmidt et al., 2020), Vietnamese (Nguyen et al., 2020), Korean (Lim et al., 2019), Russian (Efimov et al., 2020), Japanese (So et al., 2022),

etc.
We use five Question Answering datasets in Table 2 for our experiments.

2.2 Language model

XLM-RoBERTa$_{\text{BASE}}$ has 12 transformer-based layers with 270M parameters. It was previously trained on 2394.3 GiB of text in 100 languages. Recent works in multilingual Natural Language Processing show that XLM-RoBERTa significantly outperforms its multilingual counterparts, such as mBERT, on both multilingual and monolingual benchmarks.
Researchers from all over the world are using XLM-RoBERTa as the core of many state-of-the-art MRC systems in many languages such as Vietnamese (Nguyen et al., 2021; Hai et al., 2021; Nguyen and Do, 2021). These works also show that in many languages other than English, XLM-RoBERTa significantly outperforms monolingual language models on the task of Question Answering. One important reason for this phenomenon is the limited budget for researching Natural Language Processing in many low-resource languages. Pre-training an enormous language model with several hundreds of parameters requires a lot of computational resources.

2.3 Experimental Settings

The base model in our experiments is XLM-RoBERTa. The model is either fine-tuned with 2000 questions from each language or with questions from the SQuAD 1.1 and the same 2000 questions from each language. All models are trained in 2 epochs using the Adam optimizer with learning rate of $2{\cdot}10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ on a single NVIDIA Tesla K80 provided by Google Colaboratory. In addition, each model is fine-tuned with batch size of 4.

2.4 Model Evaluation

Following previous works (Rajpurkar et al., 2016) in Question Answering, we use two metrics, Exact Match (EM) and F1-score, to evaluate the performances of different models on Reading Comprehension task.

- **EM**: (Exact Match) The percentage of answers predicted by the MRC system match exactly any one of the gold answer(s) annotated by the human reader.
- **F1**: F1-score measured the average overlap between predicted answers with those in the gold answers. For each question, we calculate the F1 score of predicted answer with each gold answer, and take the maximum F1 as the F1 of the corresponding question.

3 Experimental Results

In the training phase of our experiments, while we use a full version of SQuAD 1.1 (Rajpurkar et al., 2016), we only train on 2,000 questions randomly sampled from each of the following datasets: GermanQuAD, FQuAD, ViQuAD, and KorQuAD.
On the other hand, we use the full development set of GermanQuAD, FQuAD, ViQuAD, and KorQuAD when evaluating the performances of the language models.

3.1 Fine-tuning on SQuAD

In this section, we first start evaluating the comprehension skills of XLM-RoBERTa in multilingual settings by fine-tuning XLM-RoBERTa on SQuAD, which is a question answering dataset in English. We then evaluate the fine-tuned XLM-RoBERTa on question answering datasets of four different low-resource languages including German, French, Vietnamese, and Korean. Our goal of this part is to investigate how successfully the multilingual model can transfer its comprehension skills among different languages.
Table 3 shows us the performances of XLM-RoBERTa on the Question Answering task in five different languages after being fine-tuned on 87,599 question-answer pairs in SQuAD 1.1 (Rajpurkar et al., 2016).
However, due to the linguistic differences among different languages, there is a variety of questions in the development sets of the five experimenting datasets that require simple reasoning skills to answer correctly. Thus, XLM-RoBERTa fine-tuned only on SQuAD 1.1 fails to predict correct answers to such questions.

3.2 Machine Reading Comprehension in Low-resource Settings

In the previous section, we see that XLM-RoBERTa failed to answer many questions with simple reasoning skills. In this section, we experiment another training pipeline with limited numbers of training samples in the target languages. Our proposed training pipeline include

		SQuAD	GermanQuAD	FQuAD	ViQuAD	KorQuAD
XLM-R finetuned SQuAD	EM	80.21	45.01	49.36	41.31	47.02
	F1	87.84	61.54	72.17	66.69	55.75

Table 3: Performances on Question Answering task in 5 different languages of XLM-RoBERTa$_{\text{BASE}}$ fine-tuned only on training set of SQuAD 1.1 (Rajpurkar et al., 2016)

two phases:

1. Fine-tuning multilingual models on a Question Answering dataset with a large number of training samples. We use SQuAD 1.1 (Rajpurkar et al., 2016) in our experiment. We refer this step of the the pipeline as pre-tuning.

2. Fine-tuning the multilingual models on Question Answering dataset of the low-resource target languages.

To determine the effects of our proposed training pipeline, we then compare the models trained in our pipeline with the models trained in traditional monolingual settings, which includes only the Question Answering dataset of the low-resource target language.

The results show that XLM-RoBERTa fine-tuned using our proposed pipeline shows significantly higher results compared to XLM-RoBERTa fine-tuned only on 2,000 questions per each of the five other Question Answering datasets. Table 4 shows that the differences in performances of these two kinds of models are 12.05% F1 and 13.72% EM, on average.

Besides, we also discover that fine-tuning XLM-RoBERTa$_{\text{BASE}}$ using our proposed pipeline is considerably more effective for target languages of French, German, and Vietnamese than for Korean. We hypothesize that this phenomenon is due to the linguistic differences between English, the language of SQuAD (Rajpurkar et al., 2016), and Korean.

4 Conclusion and Future Work

4.1 Conclusion

As shown in Table 3, we can see that SQuAD 1.1 has certain impacts on MRC not only in English, but also in other languages. Compared to the results in Table 4, we notice that the results when fine-tuning on SQuAD 1.1 alone are better than those when fine-tuning on a specific language for some low-resource languages like Germany or French as the former approach gives better exact match (EM) and F1 scores. The power the SQuAD 1.1 is further proven by the data in Table 4 as additional fine-tuning on the SQuAD 1.1 substantially improves the performances of the models on MRC task in the four low-resource languages. We can then verify the positive effects that fine-tuning SQuAD 1.1 has on the MRC task of the other languages. This is a demonstration of the effects of Transfer Learning among Languages in MRC.

4.2 Future Work

As we have pointed out in Section 3, due to the linguistic differences between English and Korean, SQuAD might not be the best dataset for the pre-tuning step of our pipeline when our target language of the QA system is Korean. Therefore, we believe that it is necessary to investigate the effects of using different Question Answering datasets in pre-tuning step of our pipeline. QA datasets that should be carefully investigated are CMRC (Cui et al., 2019) in Chinese and KorQuAD (Lim et al., 2019) in Korean.

These proposed experiments are important because the Question Answering datasets following the dataset SQuAD, inspired by many analysis (Sugawara et al., 2018; Jia and Liang, 2017) into weaknesses of SQuAD in evaluating the real comprehension skills of language models, had many efforts in improving the quality of annotated questions such as modifying annotating process or hiring annotators process. With a higher quality in training samples, the multilingual models might show a more robust performance on the same task in different languages.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal,

		GermanQuAD	FQuAD	ViQuAD	KorQuAD
XLM-R finetuned traditionally	EM	42.56	42.20	47.85	61.00
	F1	61.22	62.77	69.06	69.61
XLM-R_{our}	EM	58.39	53.85	65.01	71.25
	F1	74.56	75.02	82.40	78.88

Table 4: We treat German, French, Vietnamese and Korean as low-resource languages and experiment our proposed training pipeline on QA datasets of these languages. To highlight the effects of our proposed pipeline, we compare XLM-RoBERTa fine-tuned using our pipeline with the XLM-RoBERTa fine-tuned traditionally.

Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. 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.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – russian reading comprehension dataset: Description and analysis. In *Lecture Notes in Computer Science*, pages 3–15. Springer International Publishing.

Nam Le Hai, Duc Nguyen Sy, Quan Chu Quoc, and Vi Ngo Van. 2021. Vc-tus at vlsp 2021 - vimrc challenge: Improving retrospective reader for vietnamese machine reading comprehension. In *Proceedings of the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021).*

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean qa dataset for machine reading comprehension.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kiet Nguyen, Nhat Nguyen, Phong Do, Anh Nguyen, and Ngan Nguyen. 2021. Vireader: A wikipedia-based vietnamese reading comprehension system using transfer learning. *Journal of Intelligent and Fuzzy Systems*, 41:1–19.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nhat Duy Nguyen and Phong Nguyen-Thuan Do. 2021. Uitsunwind at vlsp 2021 - vimrc challenge: A simply self-ensemble model for vietnamese machine reading comprehension. In *Proceedings of the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021).*

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.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. Jaquad: Japanese question answering dataset for machine reading comprehension.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14506–14514.