

Improving Low resource Reading Comprehension via Cross lingual Transposition Rethinking

Gaochen Wu
wgc2019@mail.tsinghua.edu.cn
Computer Science and Technology
Tsinghua University
Beijing, China

Bin Xu
xubin@tsinghua.edu.cn
Computer Science and Technology
Tsinghua University
Beijing, China

Yuxin Qin
tyx16@mails.tsinghua.edu.cn
Computer Science and Technology
Tsinghua University
Beijing, China

Weizhou Wang
1157639917@qq.com
State Grid GanSu Electric Power
Lanzhou, China

Geng Wang
Wanggen@gs.sgcc.com.cn
State Grid GanSu Electric Power
Lanzhou, China

ABSTRACT

Extractive Question Answering (QA) has made tremendous advances enabled by the availability of large scale high quality QA training data. Although such rapid progress and widespread application, extractive QA datasets in languages other than English remain scarce. Moreover, collecting such a sufficient amount of training data for each language is costly and even impossible. To address this issue, we propose a Cross Lingual Transposition ReThinking (XLTT) model by modelling existing high quality extractive reading comprehension datasets in a multilingual environment. To be specific, we present multilingual adaptive attention (MAA) to combine intra attention and inter attention to learn more general generalizable semantic and lexical knowledge from each pair of language families. Furthermore, to make full use of existing datasets, we adopt a new training framework to train our model by calculating task level similarities between each existing dataset and target dataset. The experimental results show that our XLTT model surpasses six baselines on two multilingual ERC benchmarks, especially more effective for low resource languages with 3.9 and 4.1 average improvement in F1 and EM, respectively.

CCS CONCEPTS

• Computing methodologies → Natural language processing.

KEYWORDS

Question Answering; Extractive Reading Comprehension; Low-resource; Multilingual Adaptive Attention; Cross-lingual

ACM Reference Format:

Gaochen Wu, Bin Xu, Yuxin Qin, Weizhou Wang, and Geng Wang. 2021. Improving Low resource Reading Comprehension via Cross lingual Transposition Rethinking. In *IJCKG '21: International Joint Conference on Knowledge*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IJCKG '21, December 06–08, 2021, Online

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-9565-6/21/12...\$15.00

<https://doi.org/10.1145/3502223.3502234>

Graphs, December 06–08, 2021, Online. ACM, New York, NY, USA, 10 pages.
<https://doi.org/10.1145/3502223.3502234>

1 INTRODUCTION

Question Answering (QA) is a comprehensive practically-useful task which also can be used to evaluate how well computer systems understand human languages [5, 18]. Due to a large number of high-quality span-extraction QA datasets released in recent years, such as SQuAD v1.1 [30], HotpotQA [43], CMRC 2018 [10], DRCd [33], and so on, span-extractive QA has become an enormously popular research topic in NLP and a wide variety of neural-based QA models have pushed the boundaries of numerous extractive QA datasets. However, these studies have been essentially monolingual and largely focused on English.

Though tremendous advances in popular and realistic span-extraction QA have been made by the NLP community in recent years, there are still two main challenges making training extractive QA systems more difficult: (1) Large-scale high-quality datasets in other languages remain scarce, even for relatively high-resource languages such as Chinese. (2) Collecting such a sufficient amount of training data for all the world's languages is costly and even impossible. Furthermore, we argue that while large-scale supervised training datasets (e.g., WikiQA [41], ImageNet [11]) are extremely useful in the past decade, but they are too expensive and not a good way in the future. An alternative approach to constructing large-scale high-quality monolingual datasets is to develop multilingual modeling systems which can generalize to a wide range of the world's languages whether training datasets are available or not in target languages.

In this year or so, we have seen a few research works on cross-lingual extractive QA modeling. [2] proposed the first zero-shot extractive reading comprehension system for Japanese and French with no extractive training datasets available, instead by an English extractive QA model (e.g., BiDAF [32] and improved BiDAF [6, 27] and an attention-based neural machine translation (NMT) model. However, (1) their proposed QA system lacks semantic awareness among <Passage, Question, Answer>. Moreover, (2) the attention-based answer alignment approach between two different languages, especially for two different language families with an astonishing breadth of complex and typologically diverse linguistic phenomena and scenarios, is challenging and could introduce additional noise.

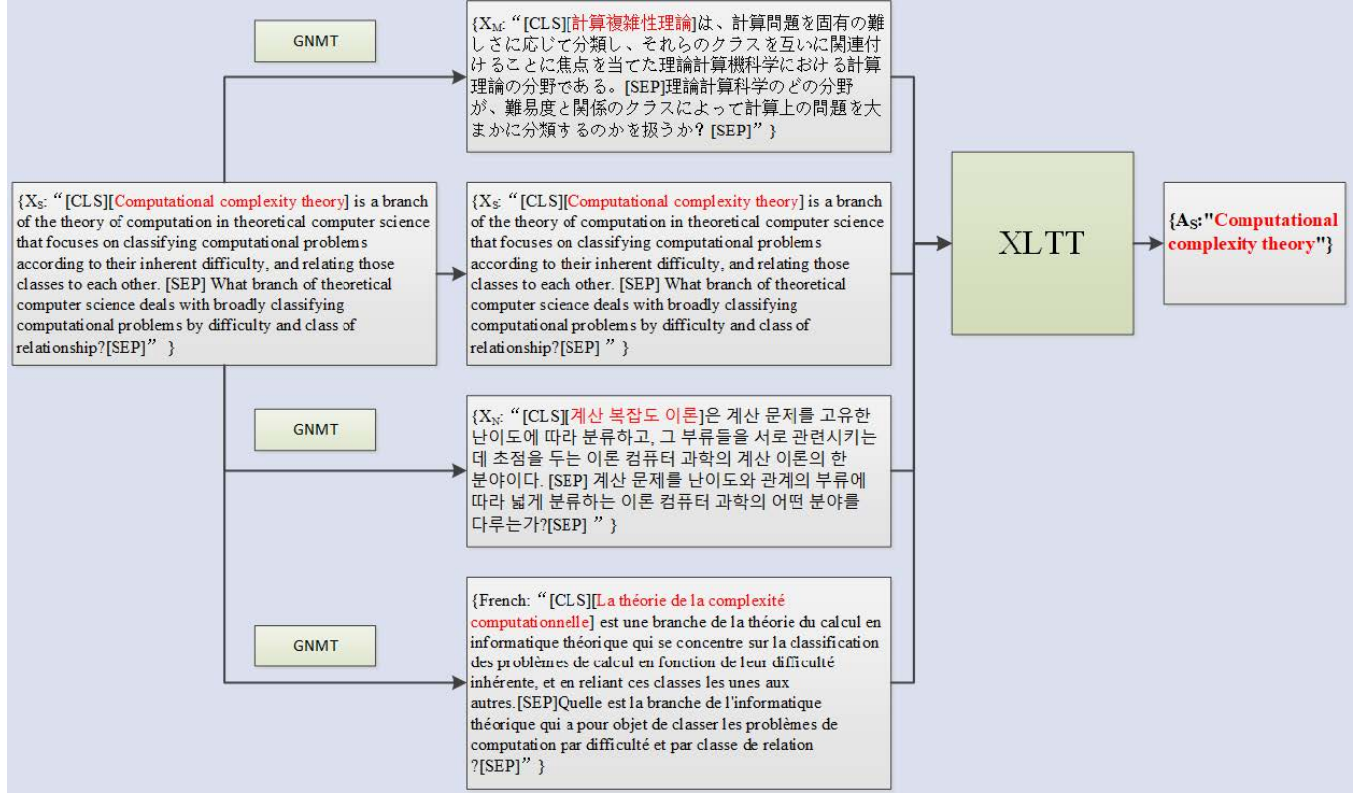


Figure 1: An input example constructed with GNMT from SQuAD v1.1 of XLTT

[9] develop back-translation approaches and Dual BERT [12] for cross-lingual machine reading comprehension task. However, (1) back-translation approaches have to align the translated answer to the original answer span in the target language using Neural Machine Translation system (NMT). (2) Dual BERT considers the training data in a bilingual context to learn the semantic relations. (3) Moreover, Dual BERT is proposed only for the condition when there is training data available in the target languages. Although interesting attempts have been made, multilingual extractive QA modeling still has not been well-studied.

Complex linguistic phenomena in the world’s languages are used to express semantic meanings of sentences and paragraphs. Due to an existing large number of the world’s languages (e.g., about 7000 living languages [8]) with a variety of linguistic features and data scenarios including Latin script, white space tokens, word formation, sentence boundaries, morphosyntax, relativization, and so on [14, 15], these typological diversities make developing general multilingual extractive QA models much more challenging. In order to speedup multilingual extractive QA modeling research, there are several datasets and benchmarks constructed in 2020 or so, including XQuAD [1], MLQA [25], TYDI QA [7], XTREME [19], and so on, their state of the art generalization performance is far behind training language results, showing huge room to improve.

In order to (1) address the scarce availability of extractive QA training data in target languages and (2) overcome multilingual

modeling challenges arising from a large amount of complex linguistic phenomena and data scenarios. In this paper, we propose a multilingual span extractive RC modeling approach named XLTT, denoting cross lingual transposition rethinking. The overall architecture and training framework of XLTT is shown as in Figure 3. Specifically, we firstly construct multilingual parallel corpora by translating large existing high quality extractive RC datasets from rich resource languages like English into different language families such as Japanese [21] and Korean [17] with SOTA neural machine translation (e.g., GNMT [39]). It should be noted that although GNMT is much more powerful, we introduce auxiliary loss for considering the quality of translated data. Secondly, we adopt multilingual adaptive attention (MAA) mechanism to combine intra attention and inter attention which can enable our extractive RC model to learn more general abstract semantic and lexical knowledge from parallel instances in each pair of representative language (e.g., Korean) and pivot language (e.g., English), which is beneficial to generalize to low resource languages.

To investigate the effectiveness of our proposed approach, we carry out extensive experiments on MLQA and XQuAD, two multilingual extractive reading comprehension datasets without providing training data. The experimental results show that our XLTT model achieves better performance than six baseline models, especially effective for low resource languages with 5.1 and 4.5 average improvement in F1 and EM on MLQA, 2.7 and 3.6 average improvement in F1 and EM on XQuAD, respectively.

To sum up, we make the following contributions: (1) We propose a cross lingual transposition rethinking (XLTT) approach by leveraging the existing large scale high quality ERC datasets in a multilingual context for solving low resource ERC. (2) We propose multilingual adaptive attention (MAA) mechanism by combining intra attention and inter attention which can enable our extractive RC model to capture more general generalizable semantic and lexical information from each pair of language families, which is beneficial to generalize to low resource languages. We also propose a new training framework to train our ERC model by computing the task level similarities between each existing training dataset and target dataset. (3) Experimental results show that our proposed model outperforms performance of 6 baselines on 2 multilingual extractive reading comprehension benchmarks, which demonstrate the effectiveness of our multilingual modeling approach.

2 RELATED WORK

Question Answering (QA) has been a popular and genuinely-useful NLP research domain in recent years. A large number of datasets are available to study QA from various perspectives, for example, including span-extraction reading comprehension [5, 32, 36–38, 44], multi-hop reading comprehension [3, 13, 46], multi-hop complex KBQA [45, 47], open-domain QA [24, 40], dialogue QA [16, 31], and many more. Due to the availability of large-scale high-quality QA datasets, the community has enjoyed huge progress. Benefiting from created large-scale potentially practically-useful extractive QA datasets such as SQuAD v1.1 [30], TriviaQA [20], NewsQA [34], SQuAD 2.0 [29] and NQ [23], along with powerful pretraining language models including BERT [12], XLNet [42], GPT [28], RoBERTa [26], and best NLP technologies like attention mechanism [4, 38], Transformer [35], GCN [22], extractive QA has become a dominant paradigm and massive progress has been achieved by the community in the past several years. Despite such popularity and widespread application, (1) extractive QA datasets in languages other than English are comparatively rare, even for relatively high-resource languages. Moreover, (2) collecting such large-scale high-quality datasets is difficult and costly. Furthermore, (3) these studies have been monolingual and largely focused on English. Extractive-QA in other languages has not been well-addressed.

2.1 Multilingual Extractive QA Datasets

In order to accelerate research progress in the field of multilingual extractive QA, several potentially practically-useful extractive QA datasets have been released in this year or so. Lewis et al. (2019) [25] present MLQA, a multi-way parallel extractive question answering evaluation benchmark in seven languages. Clark et al. (2020) [7] propose TYDI QA, a QA dataset covering eleven typologically diverse languages with 204K QA pairs. Artetxe et al. (2019) [1] create XQuAD, a dataset of 1190 SQuAD v1.1 QA pairs from 240 paragraphs by translating the instances into ten languages. Hu, J. et al (2020) [19] release XTREME, a massively multilingual multi-task benchmark to verify and validate cross-lingual generalization capability of multilingual models. The goals of these proposed multilingual datasets and benchmarks are to encourage and speedup research on developing high-quality multilingual question answering systems to generalize across a large number of the world's

languages with a wide variety of complex linguistic phenomena and to further accelerate research progress on models that generalize well across the linguistic phenomena and data scenarios of the world's languages. Due to huge challenges arising from typological diversity, generalization capabilities of multilingual models are far behind training-language performance and still have significant room for improvement.

2.2 Multilingual Extractive QA Modeling

The increasing interest in cross-lingual extractive QA modeling has been seen in this year or so. We observe a few recently-published cross-lingual extractive QA research works briefly introduced below. Asai et al. (2018) [2] developed the first zero-shot extractive QA system for non-English languages (e.g., Japanese and French), instead by an English extractive QA model and an attention-based neural machine translation (NMT) model. They evaluated their proposed zero-shot extractive QA system on small Japanese and French datasets translated from the SQuAD v1.1 development set. However, the generalization results of their proposed QA system are quite low. In addition, their proposed attention-based answer alignment method between various language families with an astonishing breadth of complex diverse linguistic phenomena are different and could result in additional noise. Furthermore, attention-based system lacks semantic awareness among <Passage, Question, Answer>.

Cui et al. (2019) [9] proposed several zero-shot back-translation models and Dual BERT for cross-lingual machine reading comprehension. However, back-translation approaches have to align the translated answer with the source answer span in the original paragraph to meet the requirement of extractive reading comprehension. Dual BERT is proposed for improving Chinese extractive reading comprehension task with a certain amount of training data. Moreover, Dual BERT only models the <Passage, Question, Answer> in a bilingual environment to learn the hidden semantic relations and is not a general multilingual extractive QA model. To sum up, multilingual extractive question answering modeling still has not been well-studied.

In summary, to address the scarce availability of ERC training data in low-resource languages and overcome challenges from a large variety of complex linguistic phenomena and scenarios, we propose XLTT shown in Figure 2, a cross-lingual transposition rethinking approach. The experimental results show that our approach obtains better performance than six baselines on two multilingual ERC benchmarks.

3 PROBLEM DEFINITION

In this paper, we focus on addressing low-resource multilingual extractive reading comprehension like MLQA without training data in target languages, in this work low-resource languages. We propose a cross-lingual generalization (XLG) task, which requires training a model with existing training data (P_S, Q_S, A_S) in high-resource languages other than target languages like Hindi. Development data in high-resource language S is used for picking hyper-parameters. At test time, the trained model is required to extract answer AT given context PT and question QT in low-resource language T . It should be noted that we only use the supervised signals in high-resource

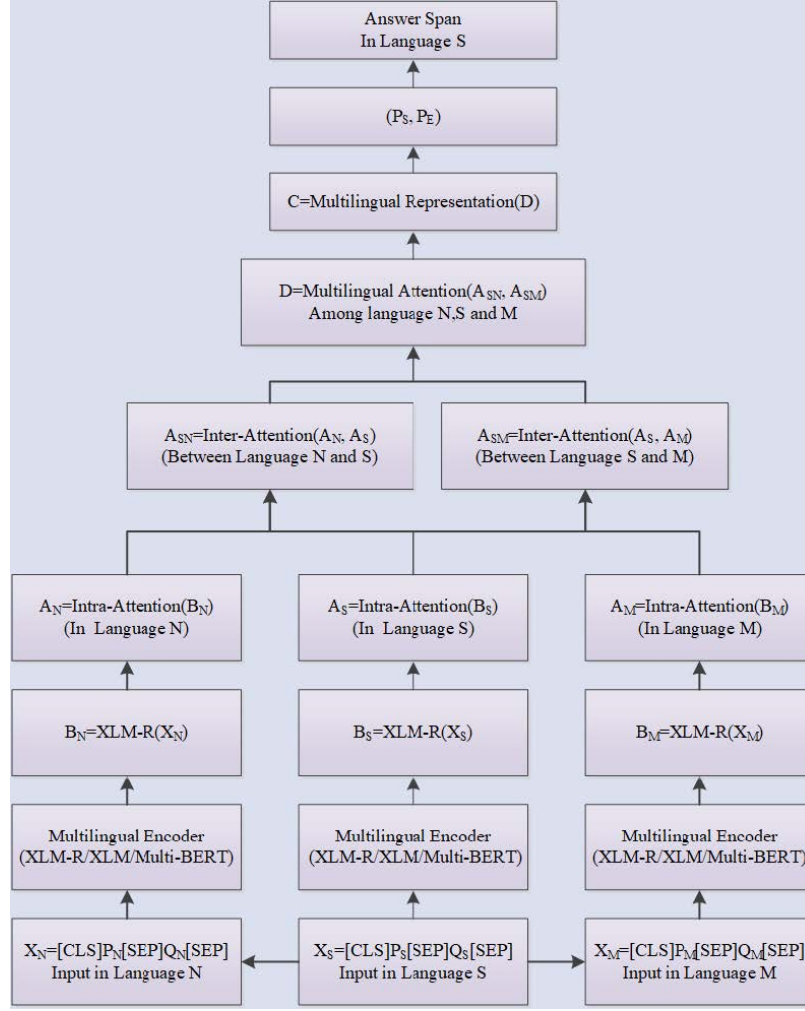


Figure 2: The overall procedure and architecture of XLTT using multilingual adaptive attention mechanism with intra-attention, inter-attention and multilingual attention mechanism.

language to train the model. We provide an example shown in Figure 1.

4 MODEL

In this section, we turn to detail our proposed XLTT illustrated in Figure 2, mainly consisting of the following 3 steps: (1) For existing extractive reading comprehension datasets, we firstly construct multilingual parallel corpora in multiple language families such as (English, Japanese, Korean, ...) using GNMT. (2) Secondly, we utilize multilingual adaptive attention to learn linguistics-related semantic and lexical knowledge in a multilingual environment for generalizing to low-resource languages. (3) Then, we calculate task-level cosine similarities between each training dataset and target dataset for using multiple training datasets, and propose the loss function for training our model.

4.1 Building Multilingual Parallel Corpora

In this study, we choose different language families from the 11 typologically diverse languages covered by the TYDI QA benchmark to evaluate our proposed XLTT modeling approach. At the same time, to guarantee the zero-shot evaluation, in which representative languages we choose for training multilingual models are also different from target test languages contained by MLQA and XQuAD. Therefore, during the training stage, we choose (English, Japanese, Korean) as language combination. In the test stage, we select (target language such as Hindi, Japanese, Korean).

For the SQuAD v1.1 dataset, we translate to Japanese and Korean. For the CMRC2018 and DRCD, we translate to (English, Japanese, Korean) and discard the Chinese instances due to MLQA and XQuAD including Chinese.

Formally, given an extractive QA instance (e.g., <Passage, Question, Answer>) in high-resource language S (e.g., English) represented as $\langle P_S, Q_S, A_S \rangle$, we organize each source language

question-passage pair as an input token sequence X_S for multilingual encoder (e.g., multilingual BERT and XLM-R), as follows:

$$X_S = [CLS]Q_S[SEP]P_S[SEP] \quad (1)$$

Similarly, we can obtain constructed multilingual parallel question-passage pairs of X_S in different language families represented as input token sequences, such as X_M in the language M and X_N in the language N as follows:

$$X_S = [CLS]Q_M[SEP]P_M[SEP] \quad (2)$$

$$X_S = [CLS]Q_N[SEP]P_N[SEP] \quad (3)$$

4.2 Multilingual Adaptive Attention Reader

In this part, we firstly use a shared multilingual encoder such as Multi-BERT and XLM-R, to encode the constructed multilingual parallel corpus to obtain contextualized vector representations. Then we simultaneously model training datasets in a multilingual environment by using multilingual adaptive attention (MAA) mechanism.

4.2.1 Multilingual Encoding. To be specific, we use the input token sequence X (e.g., X_S , X_M and X_N) for multilingual encoder to obtain contextualized representations B (e.g., B_S , B_M and B_N , respectively) as follows:

$$B = \text{XLMR}(X) \quad (4)$$

$$B_S \in R^{L_S * h}, B_M \in R^{L_M * h}, B_N \in R^{L_N * h}$$

Where L indicates the length of the input token sequence and h represents the dimensionality of the hidden state (e.g., 768 for multilingual BERT, 1152 for XLM-R).

4.2.2 Multilingual Intra/Inter Attention. Typically, intraattention mechanism is used to relate different positions of a single sequence in order to calculate more accurate representation of the sequence. Intra-attention has already been used successfully in a variety of NLP tasks including extractive RC. In our multilingual RC modeling approach, to enrich the language representation for final prediction, we firstly take advantage of intra-adaptive attention to filter the irrelevant part of each representation (e.g., B_S , B_M and B_N) in order to make the attention more precise and then apply the softmax function to obtain selfattended representations as in Eqn. 5 6 and 7:

$$A_S = \text{softmax}(B_S \cdot B_S^T), A_S \in R^{L_S * h} \quad (5)$$

$$A_M = \text{softmax}(B_M \cdot B_M^T), A_M \in R^{L_M * h} \quad (6)$$

$$A_N = \text{softmax}(B_N \cdot B_N^T), A_N \in R^{L_N * h} \quad (7)$$

Then, we use multi-head inter-attention to further enhance the final representation by modeling the linguistics-related semantic relations between the passage and question in each pair of languages (i.e., the pivot language S and one of various language families, such as language M or N). we calculate inter-attention of each pair of languages (i.e., S and M , S and N) as in Eqn. (8) and (9):

$$A_{SM} = B_S \cdot B_M^T, A_{SM} \in R^{L_S * L_M} \quad (8)$$

$$A_{SN} = B_S \cdot B_N^T, A_{SN} \in R^{L_S * L_N} \quad (9)$$

To utilize the benefit of both intra-attention and inter-attention, we take advantage of multilingual adaptive attention mechanism and calculate the dot product between intra-adaptive attention and all inter-attention representations to obtain attended representations in the following 4 equations:

$$\tilde{A}_{SM} = A_S \cdot A_{SM} \cdot A_M^T, \tilde{A}_{SM} \in R^{L_S * L_M} \quad (10)$$

$$\tilde{A}_{SN} = A_S \cdot A_{SN} \cdot A_N^T, \tilde{A}_{SN} \in R^{L_S * L_N} \quad (11)$$

$$C'_M = \text{softmax}(\tilde{A}_{SM}) \cdot B_M, C'_M \in R^{L_S * h} \quad (12)$$

$$C'_N = \text{softmax}(\tilde{A}_{SN}) \cdot B_N, C'_N \in R^{L_S * h} \quad (13)$$

Furthermore, we propose multilingual attention in order to capture more general semantic knowledge from various languages families by concatenating all attended representations in above Eqn. (12) and (13). Multilingual attention allows the model to jointly attend to useful information from different representation subspaces of different languages.

$$\text{Multilingual}(S, M, N, \dots) = \text{Concat}(C'_M, C'_N, \dots) \quad (14)$$

In this work, we only consider pivot language S and two language families M and N to form a 3-language multilingual context, so the multilingual attention representation can be simplified as in Eqn. (15):

$$\begin{aligned} C' &= \text{Multilingual}(S, M, N) \\ &= \text{Concat}(C'_M, C'_N), \\ C' &\in R^{L_S * 2h} \end{aligned} \quad (15)$$

After getting multilingual attention representation, we use another fully connected layer along with residual connection and normalization layer to get final knowledge-enhanced source representation as in Eqn. (17). Finally, we use GS as the multilingual adaptive attention representation for final answer prediction.

$$C = W_C C' + b_C, W_C \in R^{2h * h}, C \in R^{L_S * h} \quad (16)$$

$$\begin{aligned} G_S &= \text{concat}[B_S, \text{LayerNorm}(B_S + C)], \\ G_S &\in R^{L_S * 2h} \end{aligned} \quad (17)$$

4.3 Task-level Similarity

For each training dataset D_i , we first collect all its questions to form a large question document Q_i , and then use the TF-IDF to represent the D_i dataset and target dataset D_t such as MLQA (using the English test set). Finally, we directly use the task-level similarity St_i between D_i and D_t as the cosine similarity as the weight of the dataset D_i during the training stage based on their TF-IDF in Eqn (18):

$$\text{Weight}_{t_i} = \text{cosine}(\text{TFIDF}(Q_i), \text{TFIDF}(Q_t)) \quad (18)$$

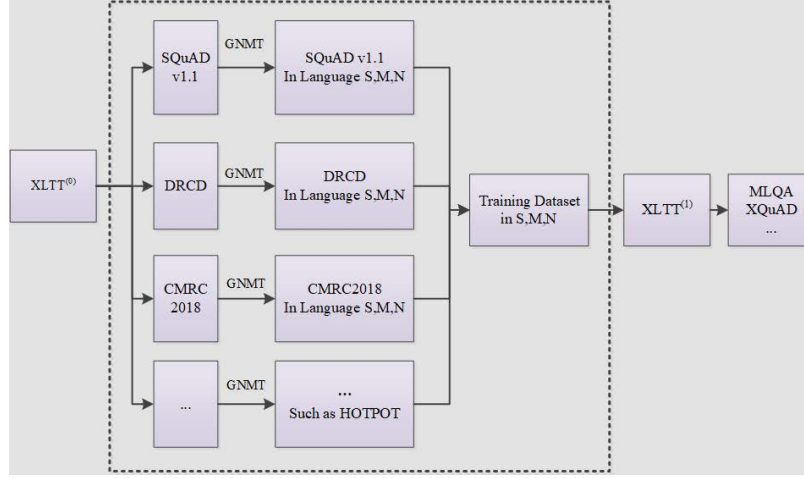


Figure 3: The training framework of XLTT using multiple existing datasets like SQuAD v1.1, DRCD, CMRC2018

Then, we normalize the above calculated similarity using Eqn (19).

$$\text{Weight}_{ti} = \text{softmax}(\text{Weight}_{ti}) \quad (19)$$

Note that we only use the question words to calculate the similarity.

4.3.1 Loss Function. We use Eqn. (17) to GSi and use GSi to calculate standard cross entropy loss for the answer span in the source training language in Eqn. (20), (21) and (22).

$$P_{Si}^s = \text{softmax}(W_S^s G_{Si} + B_S^s) \quad (20)$$

$$P_{Si}^e = \text{softmax}(W_S^e G_{Si} + B_S^e) \quad (21)$$

$$L_S = - \sum_{i \in S} \text{Weight}_{ti} \cdot \left(\frac{1}{|S_i|} \sum_{k=1}^K y_{Si}^s \log(P_{Si}^s) + y_{Si}^e \log(P_{Si}^e) \right) \quad (22)$$

Where S_i denotes the i -th training dataset in language S such as SQuAD v1.1. In order to estimate how the translated instances assemble the original instances, we take advantage of cosine semantic similarity α_R obtained in Eqn. (23) between the golden span representations in source S and representative R languages denoted as H_S and H_R described in [10], we are not going to detail how to calculate H_S and H_R in this paper.

$$\alpha_R = \max\{0, \text{Cos} \langle H_S, H_R \rangle\} \quad (23)$$

$$L_R^{GNMT} = - \frac{1}{K} \sum_{k=1}^K (y_r^s \log(P_R^s) + y_r^e \log(P_R^e)) \quad (24)$$

$$L_{\text{Objective}} = L_S + \sum_R \alpha_R L_R^{GNMT} \quad (25)$$

We also calculate the auxiliary cross-entropy loss in Eqn. (24) for translated samples in R , where a α_R parameter is applied to this loss. The overall objective function is calculated in Eqn. (25).

5 EXPERIMENTAL DETAILS

We now turn toward how to evaluate our proposed multilingual modeling approach using MLQA and XQuAD with the zero-shot evaluation.

5.1 Implementation Details

We use ADAM with weight decay optimizer using an initial learning rate of $4e-5$ and adopt cosine learning rate decay scheme instead of the original linear decay, which we found it beneficial for stabilizing results. All systems are trained on multiple hosts

and multiple graphic units of GPU Cloud platform with Tesla V100 with 32 GB RAM. In this paper, we choose SQuAD v1.1, CMRC2018 and DRCD as our original training data to train our XLTT model, and use the English development data for picking hyper-parameters and evaluate the model on the MLQA and XQuAD test set.

5.2 Datasets

XQuAD [1] is a benchmark dataset for evaluating cross-lingual span-extractive QA performance. The dataset consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1 together with professional translations into 10 languages.

MLQA [25] is a multi-way parallel extractive multilingual QA evaluation benchmark dataset including 7 languages detailed in Table 6 in Appenx A.

SQuAD v1.1 [30] is built based on Wikipedia articles in English. 536 high-quality hand-annotated articles are sampled and annotators formulate questions based on each individual paragraph requiring that the answer must be highlighted from the given paragraph. The data statistics of SQuAD v1.1 is shown in Table 7 in Appendix A

DRCD [33] and CMRC2018 [10] are two Chinese extractive reading comprehension datasets. The statistics are shown in Table 8 in Appendix A.

Table 1: F1 and EM scores on the MLQA test set. XLTT with XLM-R denotes that XLTT adopts XLM-R as a multilingual encoder to represent the input

F1/EM	English	Spanish	German	Arabic	Hindi	Vietnamese	Mandarin	Avg
BERT-Large [25]	80.2/67.4							
Multi-BERT [19]	80.2/67.0	67.4/49.2	59.0/43.8	52.3/34.6	50.2/35.3	61.2/40.7	59.6/38.6	61.4/44.2
XLM [25]	74.9/62.4	68.0/49.8	62.2/47.6	54.8/36.3	48.8/27.3	61.4/41.8	61.1/39.6	61.6/43.5
XLM-R [19]	83.5/70.6	74.1/56.6	70.1/54.9	66.6/47.1	70.6/53.1	74.0/52.9	67.1/44.4	72.3/54.2
Translate-Test, BERT-Large [25]		65.4/44.0	57.9/41.8	33.6/20.4	23.8/18.9	58.2/33.2	44.2/20.3	47.2/29.8
Translate-Train, Multi-BERT [25]		53.9/37.4	62.0/47.5	51.8/33.2	55.0/40.0	62.0/43.1	61.4/39.5	57.7/40.1
Translate-Train, XLM [19]		65.2/47.8	61.4/46.7	54.0/34.4	50.7/33.4	59.3/39.4	59.8/37.9	58.4/39.9
XLTT with Multi-BERT (ours)	81.0/68.1	67.1/49.2	60.0/46.2	54.8/36.7	50.7/35.7	63.9/44.2	65.1/44.0	63.2/46.3
XLTT with XLM-R (ours)	83.9/71.2	76.7/59.4	72.2/56.8	75.7/54	77.5/59.1	80.8/59.7	74.7/51.1	77.4/58.7

Table 2: Size of Models

Models	Model Size (million)
BERT-Large	340
Multilingual BERT	110
XLM	270
XLM-R	550

5.3 Baselines

We turn to briefly introduce baseline models on MLQA and XQuAD to prove effectiveness of our approach, the size of models detailed in Table 2.

Translate-Train: This baseline model includes Translate-Train + Multi-BERT and Translate-Train + XLM. The training datasets such as SQuAD v1.1 are first translated into test languages to train Multi-BERT and XLM, and then evaluate them on test data.

Translate-Test: This baseline model represents Translate-Test + BERT-Large. Firstly, BERT-Large is trained using the SQuAD v1.1 English training data and the test data is translated into English. Then evaluate BERT_Large on the translated test set and translate the predicted answer into test languages for final evaluation.

Multi-BERT: Multi-BERT is a transformer-based language model pretrained on the Wikipedia of 104 languages using masked language modelling.

XLM: XLM uses a similar pretraining objective as multilingual BERT with a larger model, a larger shared vocabulary, and trained on the same Wikipedia data covering 100 languages, but XLM introduces additional translation language model.

XLM-R: XLM-R is similar to XLM but is trained on more than a magnitude more data from the web.

5.4 Evaluation

As far as we know, most of extractive RC tasks adopt EM and F1 metrics as evaluation measures. If prediction answer exactly matches the golden answer span, EM is equal to 1, otherwise 0. As for F1, a harmonic mean of precision and recall, each of which is calculated over instances within a language. The EM and F1 scores for each instance are averaged within each target language to obtain final F1 and EM scores, respectively. We take advantage of the multilingual evaluation script for the MLQA and XQuAD tasks to perform answer preprocessing operations such as stripping Unicode punctuation, standalone article stripping.

5.5 Results and Analysis

Table 1 and Table 3 show the results on the MLQA and XQuAD test sets, respectively. Overall, our XLTT with XLM-R as input multilingual encoder has the best generalization performance comparing the six baselines. To be specific, when comparing with the previous best model, our model achieves an average improvement of 5.1 in F1 and 4.5 in EM on the MLQA test set, and obtains an average improvement of 2.7 in F1 and 3.6 in EM on the XQuAD test set. The experimental results on the two multilingual ERC demonstrate effectiveness of our proposed multilingual ERC approach.

Table 3: F1 and EM scores on the XQuAD test set

F1/EM	English	Arabic	German	Greek	Spanish	Hindi	Russian	Thai	Turkish	Vietnamese	Mandarin	Avg
Multi-BERT [19]	83.5/72.2	61.5/45.1	70.6/54.0	62.6/44.9	75.5/56.9	59.2/46.0	71.3/53.3	42.7/33.5	55.4/40.1	69.5/49.6	58.0/48.3	58.7/49.4
XLM [19]	74.2/62.1	61.4/44.7	66.0/49.7	57.5/39.1	68.2/49.8	56.6/40.3	65.3/48.2	35.4/24.5	57.9/41.2	65.8/47.6	49.7/39.7	59.8/44.3
XLM-R [19]	86.5/75.7	68.6/49.0	80.4/63.4	79.8/61.7	82.0/63.9	76.7/59.7	80.1/64.3	74.2/62.8	75.9/59.3	79.1/59.0	59.3/50.0	76.6/60.8
Translate-train Multi-BERT [19]	83.5/72.2	68.0/51.1	75.6/60.7	70.0/53.0	80.2/63.1	69.6/55.4	75.0/59.7	36.9/33.5	68.9/54.8	75.6/56.2	66.2/56.6	70.0/56.0
Translate-train [19]	86.0/74.5	71.0/54.1	78.8/63.9	74.2/56.1	82.4/66.2	71.3/56.2	78.1/63.0	38.1/34.5	70.6/55.7	78.5/58.8	67.7/58.7	72.4/58.3
Translate-train BERT-Large [19]	87.9/77.1	73.7/58.8	79.8/66.7	79.4/65.5	82.0/68.4	74.9/60.1	79.9/66.7	64.6/50.0	67.4/49.6	76.3/61.5	73.7/59.1	76.3/62.1
XLTT with Multi-BERT (ours)	85.2/73.9	63.7/46.6	72.3/55.7	64.2/46.5	77.1/58.4	61.0/47.7	72.9/55.1	44.9/35.8	57.7/42.4	71.1/51.3	60.2/50.4	66.4/51.3
XLTT with XLM-R (ours)	87.7/78.2	71.5/51.3	83.1/65.2	82.5/65.7	84.8/65.7	79.4/61.8	83.0/66.1	77.1/64.5	78.9/61.1	81.7/61.3	62.1/69.8	79.3/64.4

According to our preliminary simple analysis, we consider the improvement on MLQA and XQuAD owing to our proposed multilingual adaptive attention approach, data augmentation of constructed multilingual parallel corpora and multiple existing training datasets.

Table 4: Ablation study for considering the effect of training datasets on XLTT mode

Pretrained XLTT (AVG F1/EM)	MLQA	XQuAD
XLTT with XLM-R (w/ SQuAD, CMRC, DRCD)	77.4/58.7	79.3/64.4
XLTT (w/o CMRC)	-1.3/-1.0	-0.9/-0.8
XLTT (w/o DRCD)	-1.1/-0.9	-0.8/-0.8
XLTT (w/o SQuAD)	-2.4/-1.8	-2.1/-1.4
XLTT (w/o CMRC, DRCD)	-1.5/-1.2	-1.4/-1.1
XLTT (w/o CMRC, SQuAD)	-2.9/-2.3	-2.4/-1.7
XLTT (w/o SQuAD, DRCD)	-2.6/-2.1	-2.2/-1.6

5.6 Ablation Study

To further determine the specific source of the contribution. We perform two ablation studies as follows. Table 4 shows the effect of training datasets on the improvement. We find that all three training datasets can improve the performance with different degrees using the new training framework.

To confirm the contribution of the multilingual parallel corpora and multilingual adaptive attention approach, we use SQuAD v1.1, CMRC2018 and DRCD to train XLM-R in cascading way and evaluate XLM-R on MLQA and XQuAD. The results in Table 5 show that multilingual parallel corpora can also improve the performance, but XLM-R is still behind our XLTT, which prove multilingual adaptive attention can also improve the performance.

Table 5: Effect of multilingual parallel corpora for the improvement

Models (Avg F1/EM)	MLQA	XQuAD
XLM-R (w/ SQuAD)	72.3/54.2	76.6/60.8
XLM-R (w/ SQuAD, CMRC2018, DRCD)	73.9/55.4	77.9/61.7
XLTT with XLM-R (w/ SQuAD, CMRC2018, DRCD)	77.4/58.7	79.3/64.4

6 CONCLUSION

In this paper, we propose a multilingual extractive RC modeling approach named XLTT, by a cross-lingual transposition rethinking approach for low-resource extractive reading comprehension using multilingual adaptive attention to model existing ERC training datasets in a multilingual context. We adopt self-adaptive attention with self-attention and inter-attention to capture linguistics-related

knowledge from parallel instances in different languages. Furthermore, we propose multilingual attention to learn general knowledge from different language families. Experimental results demonstrate the effectiveness of our multilingual ERC modeling approach.

ACKNOWLEDGMENTS

We would like to thank all reviewers for their thorough reviewing and providing constructive comments to improve this paper. This work is mainly funded by the Science and Technology Project of State Grid Corporation of China “Research and application of knowledge service and skill drill system for substation operation and Maintenance” (No. SGGSTS00RZJS2000136). This work was also supported by Beijing MoreHealth Technology Group Co. Ltd.

REFERENCES

- [1] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4623–4637. <https://doi.org/10.18653/v1/2020.acl-main.421>
- [2] Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *ArXiv preprint abs/1809.03275* (2018). <https://arxiv.org/abs/1809.03275>
- [3] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net. <https://openreview.net/forum?id=SJgVHkrYDH>
- [4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*. Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1409.0473>
- [5] 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)*. Association for Computational Linguistics, Vancouver, Canada, 1870–1879. <https://doi.org/10.18653/v1/P17-1171>
- [6] 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)*. Association for Computational Linguistics, Melbourne, Australia, 845–855. <https://doi.org/10.18653/v1/P18-1078>
- [7] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics* 8 (2020), 454–470. https://doi.org/10.1162/tacl_a_00317
- [8] Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.
- [9] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-Lingual 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)*. Association for Computational Linguistics, Hong Kong, China, 1586–1595. <https://doi.org/10.18653/v1/D19-1169>
- [10] 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)*. Association for Computational Linguistics, Hong Kong, China, 5883–5889. <https://doi.org/10.18653/v1/D19-1600>
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 20–25 June 2009, Miami, Florida, USA. IEEE Computer Society, 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

- Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [13] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive Graph for Multi-Hop Reading Comprehension at Scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2694–2703. <https://doi.org/10.18653/v1/P19-1259>
 - [14] Matthew Dryer, David Gil, and Martin Haspelmath. 2005. *The world atlas of language structures*. Oxford University Press.
 - [15] Matthew S Dryer. 2011. Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig (2011).
 - [16] Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.), 2946–2955. <https://proceedings.neurips.cc/paper/2018/hash/d63fbf8c3173730f82b150c5ef38b8ff-Abstract.html>
 - [17] Na-Rae Han and Shijong Ryu. 2005. Guidelines for Penn Korean treebank version 2.0. *IRCS Technical Reports Series* (2005), 7.
 - [18] Karl Moritz Hermann, Tomáš Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.), 1693–1701. <https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd047fd0f3c96-Abstract.html>
 - [19] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 4411–4421. <http://proceedings.mlr.press/v119/hu20b.html>
 - [20] 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)*. Association for Computational Linguistics, Vancouver, Canada, 1601–1611. <https://doi.org/10.18653/v1/P17-1147>
 - [21] Stefan Kaiser, Yasuko Ichikawa, Noriko Kobayashi, and Hilofumi Yamamoto. 2013. *Japanese: A comprehensive grammar*. Routledge.
 - [22] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=3JU4ayYgl>
 - [23] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276
 - [24] Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking Paragraphs for Improving Answer Recall in Open-Domain Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 565–569. <https://doi.org/10.18653/v1/D18-1053>
 - [25] Patrick Lewis, Barlas Oguz, Rutu Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7315–7330. <https://doi.org/10.18653/v1/2020.acl-main.653>
 - [26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint abs/1907.11692* (2019). <https://arxiv.org/abs/1907.11692>
 - [27] ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. 2018. Deep contextualized word representations. *NAACL-HLT*.
 - [28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).
 - [29] 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)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. <https://doi.org/10.18653/v1/P18-2124>
 - [30] 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*. Association for Computational Linguistics, Austin, Texas, 2383–2392. <https://doi.org/10.18653/v1/D16-1264>
 - [31] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1073–1083. <https://doi.org/10.18653/v1/P17-1099>
 - [32] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=HJ0UKP9ge>
 - [33] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiyi Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *ArXiv preprint abs/1806.00920* (2018). <https://arxiv.org/abs/1806.00920>
 - [34] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordani, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Vancouver, Canada, 191–200. <https://doi.org/10.18653/v1/W17-2623>
 - [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
 - [36] Shuohang Wang and Jing Jiang. 2017. Machine Comprehension Using Match-LSTM and Answer Pointer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=B1-q5Pqxl>
 - [37] 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)*. Association for Computational Linguistics, Vancouver, Canada, 189–198. <https://doi.org/10.18653/v1/P17-1018>
 - [38] Gaochen Wu, Bin Xu, Dejie Chang, and Bangchang Liu. 2021. A Multilingual Modeling Method for Span-Extraction Reading Comprehension. *ArXiv preprint abs/2105.14880* (2021). <https://arxiv.org/abs/2105.14880>
 - [39] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint abs/1609.08144* (2016). <https://arxiv.org/abs/1609.08144>
 - [40] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 72–77. <https://doi.org/10.18653/v1/N19-4013>
 - [41] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2013–2018. <https://doi.org/10.18653/v1/D15-1237>
 - [42] Z Yang, Z Dai, Y Yang, J Carbonell, R Salakhutdinov, and XLNet Le Q. 1906. generalized autoregressive pretraining for language understanding. *arXiv* 2019; 1906.08237.
 - [43] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. <https://doi.org/10.18653/v1/D18-1259>
 - [44] W Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V QANet Le. 2018. Combining local convolution with global self-attention for reading comprehension. *ArXiv preprint abs/1804.09541* (2018). <https://arxiv.org/abs/1804.09541>
 - [45] Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. 2018. Variational Reasoning for Question Answering With Knowledge Graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, New Orleans, Louisiana, USA, February 2-7, 2018, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 6069–6076. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16983>
 - [46] Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-XH: Multi-hop question answering with extra hop attention.

- [47] Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An Interpretable Reasoning Network for Multi-Relation Question Answering. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2010–2022. <https://aclanthology.org/C18-1171>

A DATASETS

The statistics of datasets we adopted in this paper are shown in Table 6, Table 7 and Table 8:

Table 6: Number of instances in MLQA

MLQA	Development	Test	Total
English (en)	1,148	11,590	12,738
German (de)	512	4,517	5,029
Spanish (es)	500	5,253	5,753
Arabic (ar)	517	5,335	5,852
Chinese (zh)	504	5,137	5,641
Vietnamese (vi)	511	5,495	6,006
Hindi (hi)	507	4,918	5,425

Table 7: Statistics of SQuAD v1.1

SQuAD v1.1	Training	Development	Test
Num of Instances	87,599	10,577	9,533

Table 8: Statistics of CMRC2018 and DRCD

Dataset	Training	Development	Test
CMRC2018	10,321	3,219	4,895
DRCD	26,936	3,524	3,493

B SIZE OF MODELS

The size of models we used in this paper as detailed in Table 9

Table 9: Size of Models

Models	Model Size (million)
BERT-Large	340
Multilingual BERT	110
XLNet	270
XLNet-R	550