XQA: A Cross-lingual Open-domain Question Answering Dataset

Jiahua Liu, Yankai Lin, Zhiyuan Liu, Maosong Sun[*]
Department of Computer Science and Technology,
Institute for Artificial Intelligence,
State Key Lab on Intelligent Technology and Systems,
Tsinghua University, Beijing, China
alphaf52@gmail.com,linyk14@mails.tsinghua.edu.cn
{liuzy,sms}@tsinghua.edu.cn

Abstract

Open-domain question answering (OpenQA) aims to answer questions through text retrieval and reading comprehension. Recently, lots of neural network-based models have been proposed and achieved promising results in OpenQA. However, the success of these models relies on a massive volume of training data (usually in English), which is not available in many other languages, especially for those low-resource languages. Therefore, it is essential to investigate cross-lingual OpenQA. In this paper, we construct a novel dataset XQA for cross-lingual OpenQA research. It consists of a training set in English as well as development and test sets in eight other languages. Besides, we provide several baseline systems for cross-lingual OpenQA, including two machine translation-based methods and one zero-shot cross-lingual method (multilingual BERT). Experimental results show that the multilingual BERT model achieves the best results in almost all target languages, while the performance of cross-lingual OpenQA is still much lower than that of English. Our analysis indicates that the performance of cross-lingual OpenQA is related to not only how similar the target language and English are, but also how difficult the question set of the target language is. The XQA dataset is publicly available at http://github.com/thunlp/XQA.

1 Introduction

In recent years, open-domain question answering (OpenQA), which aims to answer open-domain questions with a large-scale text corpus, has attracted lots of attention from natural language processing researchers. Chen et al. (2017) proposed DrQA model, which used a text retriever to obtain relevant documents from Wikipedia, and further applied a trained reading comprehension model to extract the answer from the retrieved documents. Moreover, researchers have introduced more sophisticated models, which either aggregate all informative evidence (Lin et al., 2018; Wang et al., 2018b) or filter out those noisy retrieved text (Clark and Gardner, 2018; Choi et al., 2017; Wang et al., 2018a) to better predict the answers for open-domain questions. Benefiting from the power of neural networks, these models have achieved remarkable results in OpenQA. However, these neural-based models must be trained with a huge volume of labeled data. Collecting and labeling large-size training data for each language is often intractable and unrealistic, especially for those low-resource languages. In this case, it is impossible to directly apply existing OpenQA models to many different languages.

To address this problem, an alternative approach is to build a cross-lingual OpenQA system. It is trained on data in one high-resource source language such as English, and predicts answers for open-domain questions in other target languages. In fact, cross-lingual OpenQA can be viewed as a particular task of cross-lingual language understanding (XLU). Recently, XLU has been applied to many natural language processing tasks such as cross-lingual document classification (Schwenk and Li, 2018), cross-lingual natural language inference (Conneau et al., 2018b), and machine translation (Lample et al., 2018). Most cross-lingual models focus on word or sentence level understanding, while the interaction between questions and documents as well as the overall understanding of the documents are essential to OpenQA. To the best of our knowledge, there is still no dataset for cross-lingual OpenQA.

In this paper, we introduce a cross-lingual OpenQA dataset called XQA. It consists of a training set in English, and development and test sets in English, French, German, Portuguese, Polish,

[*] Corresponding author: Maosong Sun

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2358–2368
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Language	Question	Answer
English	Do you know that the <Query> is the largest stingray in the Atlantic Ocean, at up to across and weighing?	Roughtail stingray
Chinese	你知道<Query>可以在美国无限期居住和工作，并持有称为“绿卡”的证件?	美国永久居民
French	Le saviez-vous le <Query> est une forme de danse classique indienne originaire du sud de l’Inde ?	Bharata natyam
German	Schon gewusst die ersten <Query> entstanden in den 1960er Jahren durch Kreuzungsversuche und zeichneten sich durch einen intensiven Duft aus?	Englische Rosen
Polish	Czy wiesz <Query> w Wojewódzkim Parku Kultury i Wypoczynku w Chorzowie i Katowicach to najdłuższa nizinna kolej linowa w Europie?	Kolej linowa „Elka”
Portuguese	Sabia que no curso da história, <Query> foi destruída duas vezes, sitiada 23 vezes, atacada 52 vezes, e capturada e recapturada 44 vezes?	Jerusalém
Russian	термин <Query> был введен в 1981 для обозначения усиления слабого сигнала при наложении шума	Стохастический резонанс
Tamil	ந்தாம் நூற்றாண்டில் இந்தியாவின் பீகார் மாநிலத்தில் துவங்கப்பட்ட <Query> (படம்) உலகின் மிகப்பழமையான பல்கலைக்கழகங்களுள் ஒன்று.	நாளந்தா பல்கலைக்கழகம்
Ukrainian	22 жовтня 2006 року на гран-прі Бразилії семиразовий чемпіон світу з автоперегонів «Формула-1» <Query> закінчив кар’єру гонщика. Гран-прі Бразилії 2006 стало 250-им гран-прі в кар’єрі гонщика за 16 років виступів.	Міхаель Шумахер

Table 1: Some examples in various languages from the XQA corpus.

Chinese, Russian, Ukrainian, and Tamil. The training set contains $56,279$ English question-answer pairs along with relevant documents. The development and test sets contain a total amount of $17,358$ and $16,973$ question-answer pairs respectively. All questions are naturally produced by native speakers, and potentially reflect cultural differences in different languages.

Moreover, we build several baseline systems that use the information of multilingual data from publicly available corpora for cross-lingual OpenQA, including two translation-based methods that translate training data and test data respectively and one zero-shot cross-lingual method (multilingual BERT (Devlin et al., 2019)). We evaluate the performance of the proposed baselines in terms of text retrieval and reading comprehension for different target languages on the XQA dataset.

The experimental results demonstrate that there is a gap between the performance in English and that in cross-lingual setting. The multilingual BERT model achieves the best performance in almost all target languages, while translation-based methods suffer from the problem of translating name entities. We show that the performance on the XQA dataset depends on not only how similar the target language and English are, but also how difficult the question set of the target language is. Based on the results, we further discuss potential improvement for cross-lingual OpenQA systems.

We will release the dataset and baseline systems online with the hope that this could contribute to the research of cross-lingual OpenQA and overall cross-lingual language understanding.

2 Related Work

2.1 Open-domain Question Answering

OpenQA, first proposed by Green et al. (1986), aims to answer an open-domain question by utilizing external resources. In the past years, most work in this area has focused on using documents (Voorhees et al., 1999), online webpages (Kwok et al., 2001), and structured knowledge graphs (Bordes et al., 2015). Recently, with the advancement of reading comprehension technique (Chen et al., 2016; Dhingra et al., 2017; Cui et al., 2017), Chen et al. (2017) utilized both the information retrieval and reading comprehension techniques to answer open-domain questions. However, it usually suffers from the noise problem since the data is constructed under the distant supervision assumption. Hence researchers have made various attempts to alleviate the noise problem in OpenQA. Wang et al. (2018a) and Choi et al. (2017) performed paragraph selection before extracting answer of the question. Min et al. (2018) proposed to select a minimal set of sentences with sufficient information to answer the questions, while Lin et al. (2018) and Wang et al. (2018b) took all informative paragraphs into consideration by aggregating evidence in multiple paragraphs. Moreover, Clark and Gardner (2018) applied a shared-normalization learning objective on sampling paragraphs. All the models mentioned above were only verified in a single language (usually in English) with vast volumes of labeled data, and cannot be easily extended to the cross-lingual scenario.

2.2 Cross-lingual Language Understanding

Recent years, plenty of work has focused on multilingual word representation learning, including learning from parallel corpus (Gouws et al., 2015; Luong et al., 2015), with a bilingual dictionary (Zhang et al., 2016; Artetxe et al., 2018), and even in a fully unsupervised manner (Conneau et al., 2018a). These multilingual word representation models could be easily extended to multilingual sentence representation by averaging the representations of all words (Klementiev et al., 2012). Nevertheless, this method does not take into account the structure information of sentences. To address this issue, much effort has been devoted to using the context vector of NMT system as multilingual sentence representation (Schwenk and Douze, 2017; Espana-Bonet et al., 2017). Recently, Artetxe and Schwenk (2018) proposed to utilize a single encoder to learn joint multilingual sentence representations for 93 languages. Besides, Devlin et al. (2019) also released a multilingual version of BERT which encoded over 100 languages with a unified encoder. These models have shown their effectiveness in several cross-lingual NLP tasks such as document classification (Klementiev et al., 2012), textual similarity (Cer et al., 2017), natural language inference (Conneau et al., 2018b), and dialog system (Schuster et al., 2019). However, there is still no existing benchmark for cross-lingual OpenQA.

In addition, another line of research attempts to answer questions in one language using documents in other languages (Magnini et al., 2004; Vallin et al., 2005; Magnini et al., 2006). Different from their setting, we emphasize on building question answering systems for other languages using labeled data from a rich source language such as English, while the documents are in the same language as the questions.

3 Cross-lingual Open-domain Question Answering

Existing OpenQA models usually first retrieve documents related to the question from the large-scale text corpus using information retrieval module, and then predict the answer from these retrieved documents through reading comprehension module. Formally, given a question Q, the OpenQA system first retrieves m documents (paragraphs) $P = \{p_1, p_2, \cdots, p_m\}$ corresponding to the question Q through information retrieval system, and then models the probability distribution of the answer given the question and the documents $\Pr(A|Q, P)$.

In cross-lingual OpenQA task, we are given a source language $D_s = \{(Q_i^s, A_i^s, P_i^s)\}_{i=1}^{n_s}$ with n_s labeled examples, and a target language $D_t = \{(Q_i^t, P_i^t)\}_{i=1}^{n_t}$ with n_t unlabeled examples. The cross-lingual OpenQA system aims to learn language independent features, and then build an answer predictor that is able to model the answer prediction probability $\Pr_t(A^t|Q_i^t, P_i^t)$ for target language under the supervision from source language.

In the following part of this section, we will introduce our baseline systems for cross-lingual OpenQA, including two translation-based methods and one zero-shot cross-lingual method.

3.1 Translation-Based Methods

The most straightforward solution for cross-lingual OpenQA is to combine the machine translation system and the monolingual OpenQA system. In this paper, we consider two ways to use the machine translation system: first, **Translate-Train** which translates the training dataset from the source language into target languages, and then trains standard OpenQA system on the trans-

Language	English	Chinese	French	German	Polish	Portuguese	Russian	Tamil	Ukrainian
Avg. question len	18.82	36.83	20.09	14.61	14.49	17.66	14.21	13.29	16.73
Avg. document len	735.91	1159.28	913.72	450.65	256.87	482.74	503.28	200.45	584.93
Avg. paragraph num	10.54	8.66	25.95	8.85	5.34	8.42	10.36	13.78	25.09

Table 2: Average length of questions and documents (number of characters for Chinese, and number of words for other languages) and average number of paragraphs in various languages.

Language	Train	Dev	Test
English	56,279	2,926	2,924
Chinese	-	2,532	2,535
French	-	1,946	1,749
German	-	3,895	3,804
Polish	-	924	922
Portuguese	-	359	348
Russian	-	3,590	3,490
Tamil	-	597	586
Ukrainian	-	589	615

Table 3: Statistics of the XQA dataset.

lated data; second, **Translate-Test** in which an OpenQA system is built with the training data in the source language, and questions and retrieved articles are translated from target languages into the source language.

For the OpenQA model, we select two state-of-the-art models, including:

Document-QA model, proposed by (Clark and Gardner, 2018), is a multi-layer neural network which consists of a shared bi-directional GRU layer, a bi-directional attention layer, and a self-attention layer to obtain the question and paragraph representations. To produce well-calibrated answer scores on each paragraph, Document-QA samples multiple paragraphs and applies a shared-normalization learning objective to them.

BERT model (short for Bidirectional Encoder Representations from Transformers), proposed by (Devlin et al., 2019), aims to pre-train deep bidirectional representations by jointly conditioning on the context information in all layers. We use BERT to encode questions and paragraphs, and also adopt the shared-normalization learning objective on top to generate well-calibrated answer scores for it.

These two translation-based methods are simple and effective, but still have some drawbacks. Both translate-train and translate-test methods rely heavily on the quality of the machine translation system. However, the quality of the machine translation system varies in different language pairs, depending on the size of parallel data and the similarity of the language pair.

3.2 Zero-shot Cross-lingual Method

Zero-shot cross-lingual method uses a unified model for both source and target languages, which is trained with labeled data in the source language and then applied directly to the target language. In this paper, we select the widely-used multilingual BERT model since it has already been proved successful on reading comprehension benchmarks such as SQuAD (Devlin et al., 2019).

Multilingual BERT is a multilingual version of BERT, which is trained with the Wikipedia dumps of the top 100 languages in Wikipedia. Similar to the monolingual OpenQA model, we also fine-tune the multilingual BERT model with the shared-normalization learning objective.

4 The XQA Dataset

In this paper, we collect a novel dataset called XQA to support the cross-lingual OpenQA task.

4.1 Data Collection

Wikipedia provides a daily "Did you know" box on the main page of various languages[1], which contains several factual questions from Wikipedia editors, with links to the corresponding answers. This serves as a good source for cross-lingual OpenQA.

We collect questions from this session, and use the entity name as well as its aliases from WikiData [2] knowledge base as golden answers. For each question, we retrieve top-10 Wikipedia articles ranked by BM25 as relevant documents. Examples in various languages are shown in Table 1.

In Wikipedia articles, the entity name almost always appears at the very beginning of the document. The model may trivially predict the first few words, ignoring the true evidence in relevant documents. In order to avoid this, we remove the first paragraph from each document.

In total, we collect 90,610 questions in nine languages. For English, We keep around 3000 ques-

[1] For English: `https://en.wikipedia.org/wiki/Main_Page`

[2] `https://www.wikidata.org`

Language	English	French	German	Russian	Tamil
1	human	human	human	human	human
2	taxon	taxon	taxon	taxon	literary work
3	film	commune of France	film	film	city
4	church	film	book	book	film
5	book	book	song	archaeological site	book
6	business enterprise	song	archaeological site	battle	chemical compound
7	song	album	business enterprise	painting	disease
8	album	sovereign state	painting	song	ethnic group
9	video game	fossil taxon	album	literary work	archaeological site
10	single	single	fossil taxon	single	chemical element

Table 4: Top answer types in some languages.

Language	zh-en	fr-en	de-en	pt-en	ru-en
THUMT	38.76	33.50	34.78	35.62	30.81
Google Trans	43.30	34.80	43.34	31.00	32.83

Table 5: BLEU score of some translation models.

tions for development and test set respectively, and use the other questions as the training set. For other languages, we evenly split the questions into development and test set. The detailed statistics in each language are shown in Table 3.

4.2 Dataset Analysis

We calculate the average length of questions and documents in different languages, and the results are shown in Table 2. The average question length for most languages falls in the range of 10 to 20. The average question length in all languages is 18.97.

The documents on the XQA dataset are considerable long, containing 703.62 tokens and 11.02 paragraphs on average. Documents in Tamil and Polish are among the shortest, with an average length of 200.45 and 256.87 respectively. Documents in French and Ukrainian contain much more paragraphs than documents in other languages.

To understand whether questions in different languages have different topic distributions, we match the answers in WikiData, and obtain their types accordingly (Note that many answers either cannot be matched to WikiData entity or do not have a type label in WikiData). The top answer types in some of the languages from WikiData are displayed in Table 4. As we can see, there are some common topics across all languages, with *human* ranking first, and *film* and *book* ranking high. Besides, many questions in French are related to *commune of France*, while the topic *battle* ranks high in Russian. This indicates that XQA captures different data distributions for different languages, which may be influenced by cultural differences to some extent.

5 Experiments

5.1 Implementation Details

In translate-test setting, we use our own translation system THUMT [3] (Zhang et al., 2017) to translate German, French, Portuguese, Russian, and Chinese data into English. Google Translate is used for Polish, Ukrainian, and Tamil as they are not supported by our translation system. Since it is very time-consuming to translate the large training data, we only perform the translate-train experiment for two selected languages, i.e., German and Chinese, using our translation system. To give an idea of the performance of the translation models, we report the BLEU scores in some public benchmarks in Table 5.

To handle multiple paragraphs for a single question, following Clark and Gardner (2018), we adopt shared-normalization as the training objective on sampling paragraphs as training object for all models. Documents are restructured by merging consecutive paragraphs up to 400 tokens. During testing, the model is run on top-5 restructured paragraphs separately, and the answer span with the highest score is chosen as the prediction.

For DocumentQA model, we use the official implementation[4] and follow the setting for TriviaQA-Wiki in (Clark and Gardner, 2018). We use GloVe 300-dimensional word vector in Translate-Test setting, and 300-dimensional Skip-gram word vector trained on Chinese/German Wikipedia dumps in Translate-Train setting.

Our BERT model is similar to the BERT model for SQuAD in (Devlin et al., 2019), but we use shared-normalization on sampling paragraphs during training. We use the BASE setting

[3] `http://thumt.thunlp.org`

[4] `https://github.com/allenai/document-qa`

Model	Translate-Test				Translate-Train				Zero-shot	
	DocQA		BERT		DocQA		BERT		Multilingual BERT	
Languages	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
English	32.32	38.29	**33.72**	**40.51**	32.32	38.29	**33.72**	**40.51**	30.85	38.11
Chinese	7.17	17.20	9.81	23.05	7.45	18.73	18.93	31.50	**25.88**	**39.53**
French	11.19	18.97	15.42	26.13	-	-	-	-	**23.34**	**31.08**
German	12.98	19.15	16.84	23.65	11.23	15.08	19.06	24.33	**21.42**	**26.87**
Polish	9.73	16.51	13.62	**22.18**	-	-	-	-	**16.27**	21.87
Portuguese	10.03	15.86	13.75	21.27	-	-	-	-	**18.97**	**23.95**
Russian	5.01	9.62	7.34	13.61	-	-	-	-	**10.38**	**13.44**
Tamil	2.20	6.41	4.58	10.15	-	-	-	-	**10.07**	**14.25**
Ukrainian	7.94	14.07	10.53	17.72	-	-	-	-	**15.12**	**20.82**

Table 6: Overall results on the XQA dataset.

Language	Top-1	Top-5	Top-10
English	57.98	73.28	77.48
Chinese	51.21	66.35	70.52
French	49.58	69.12	74.59
German	41.86	55.90	60.14
Polish	31.52	46.75	52.60
Portuguese	35.21	51.34	57.57
Russian	28.88	43.87	49.77
Tamil	43.95	56.72	60.44
Ukrainian	43.85	60.22	65.12

Table 7: Retrieval performance on the XQA dataset.

with a maximum sequence length of 512. The translate-test model is initialized with the public released "BERT-Base, Cased" pretrained model, while translate-train and multilingual BERT models are initialized with the "BERT-Base, Multilingual Cased" model.

The widely accepted exact match (EM) and F1 over tokens in the answer(s) are used as the evaluation metrics. In translate-test setting, we translate the golden answers from the target languages into English, and report the results based on the translated answers.

5.2 Retrieval Results

First, we show the retrieval performance for different languages in Table 7. As we can see, the retrieval performance varies for questions from different language sets. The retrieval results for questions from English, French and Chinese set are among the best, while answers to questions from Portuguese, Polish and Russian set are much harder to retrieve.

Figure 1 suggests that as the question length increases, the retrieval performance in all languages grows. This is not difficult to understand, because longer questions will provide more information and make the retrieval problem easier.

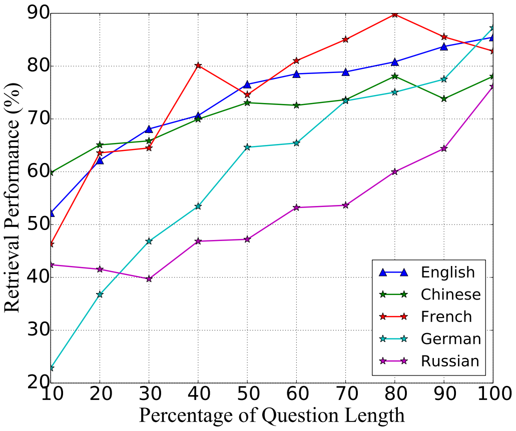


Figure 1: Retrieval performance over different question lengths.

5.3 Overall Results

Table 6 shows the overall results for different methods in different languages. There is a large gap between the performance of English and that of other target languages, which implies that the task of cross-lingual OpenQA is difficult.

In the English test set, the performance of the multilingual BERT model is worse than that of the monolingual BERT model. In almost all target languages, however, the multilingual model achieves the best result, manifesting its ability in capturing answers for questions across various languages.

When we compare DocumentQA to BERT, although they have similar performance in English, BERT consistently outperforms DocumentQA by a large margin in all target languages in both translate-test and translate-train settings. We conjecture that it is because the BERT model, which has been pretrained on large-scale unlabeled text data, has better generalization power, and could better handle the different distributions between

Languages	Translate-Test BERT		Multilingual BERT	
	EM	F1	EM	F1
Chinese	12.50	26.53	35.93	48.49
French	22.45	33.35	31.21	39.23
German	32.22	41.67	36.67	43.58
Polish	28.21	37.22	31.17	37.41
Portuguese	25.81	35.10	33.68	39.52
Russian	14.77	24.95	21.11	25.67
Tamil	5.20	14.30	16.95	22.65
Ukrainian	16.89	30.30	24.26	32.18

Table 8: Reading comprehension performance.

Languages	Genetic dist.	Pct. of easy	EM
German	30.8	19.09	36.67
Chinese	82.4	33.24	35.93
Portuguese	59.8	29.03	33.68
French	48.7	23.37	31.21
Polish	66.9	17.70	31.17
Ukrainian	60.3	21.18	24.26
Russian	60.3	18.56	21.11
Tamil	96.5	17.63	16.95

Table 9: Performance with respect to language distance and percentage of "easy" questions.

the original English training data and the machine translated test data.

Translate-train methods outperform translate-test methods in all cases except for DocumentQA in German. This may be due to the fact that DocumentQA uses space-tokenized words as basic units. In German, there is no space between compound words, resulting in countless possible combinations. Therefore, many of the words in translate-train German data do not have pretrained word vectors. On the contrary, using WordPiece tokenizer, BERT is not influenced by this.

6 Discussion

6.1 Reading Comprehension Results across Different Languages

To remove the influence of retrieval, and compare the reading comprehension performance across different target languages, we conduct experiments on a subset of questions whose answers can be found in the top-10 retrieved documents. As BERT consistently outperforms DocumentQA in translation-based methods, we only report the result of BERT model in Table 8.

We assume that the reading comprehension performance in the target language depends on two factors, the degree of similarity between the target language and the source language (i.e. English), and the intrinsic difficulty of the question set in the target language.

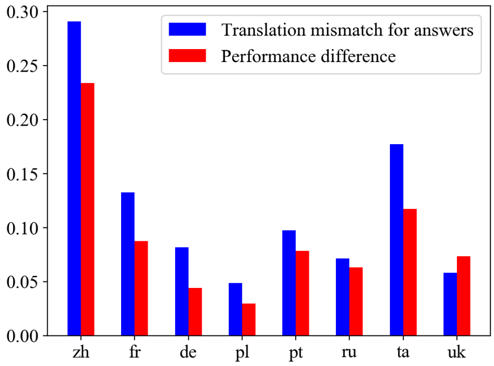


Figure 2: Performance difference (EM) between translate-test BERT and multilingual BERT, along with the percentage of translation mismatch for answers.

To quantify the intrinsic difficulty of the question sets in different languages, we calculate the percentage of questions whose answers can be found in the sentence that shares the most words with the question. We refer those questions as "easy" questions, and use the percentage of those questions as a rough indicator of how hard the subset is.

To measure the degree of similarity between the target language and English, we use the genetic distance of the language pair given by eLinguistics.net [5]. In their model, the score calculation for two languages is based on the comparison of the consonants in certain well-chosen words. The quantification of the consonant relationship is established partially with data from (Brown et al., 2013). The larger the distance is, the less similar English and the target language are.

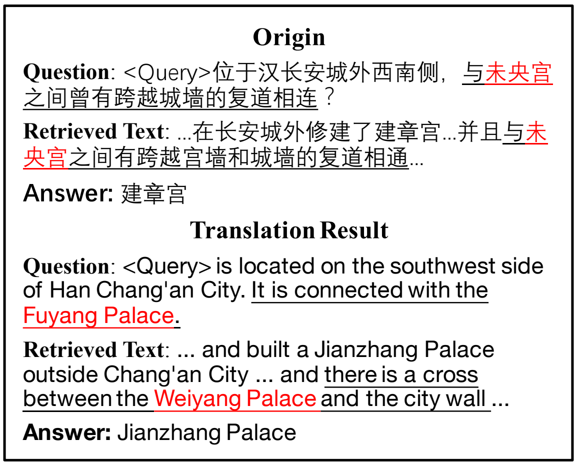
The results in Table 9 verify our assumption. The performance of different languages generally decreases as the genetic distance grows. The exceptions are Chinese and Portuguese since the percentages of "easy" questions in them are significantly higher than those in other languages. For languages that have similar genetic distances with English (i.e. Russian, Ukrainian, and Portuguese), the performance increases as the percentage of "easy" questions grows.

6.2 Limitation of Translation-based Method

Our experiments demonstrate that translation-based methods do not perform well in cross-lingual OpenQA task. Particularly, we observe

[5] http://www.elinguistics.net

a large gap between the results of multilingual BERT and translate-test BERT for Chinese and Tamil. Through error analysis, we find that for a large portion of questions in Chinese and Tamil, the answers are translated into different forms under different conditions (i.e. with context and without context). This significantly decreases the metric numbers of translation-based systems in these languages. In Figure 2, we show the difference of reading comprehension performance (EM) between translate-test BERT and multilingual BERT, along with the percentage of questions whose answers are translated into different forms in the documents. As we can see, there is a correlation between the two variables.



Origin

Question: <Query>位于汉长安城外西南侧，与未央宫之间曾有跨越城墙的复道相连？

Retrieved Text: ...在长安城外修建了建章宫...并且与未央宫之间有跨越宫墙和城墙的复道相通...

Answer: 建章宫

Translation Result

Question: <Query> is located on the southwest side of Han Chang'an City. It is connected with the Fuyang Palace.

Retrieved Text: ... and built a Jianzhang Palace outside Chang'an City ... and there is a cross between the Weiyang Palace and the city wall ...

Answer: Jianzhang Palace

Figure 3: Example of translation error of name entity.

subset	English	Chinese	Δ
easy	58.30	52.48	-5.82 (-9.98%)
other	38.42	28.77	-9.65 (-25.11%)

Table 10: Reading comprehension performance for English and Chinese.

In fact, the performance of translation-based method depends heavily on the translation quality of name entities. As we know, name entities are critical for question answering. For many factual questions, the answers are either name entities themselves, or highly related to name entities (i.e. the property of a name entity). Translation error or inconsistency of name entities would significantly hurt the performance of translation-based cross-lingual OpenQA system. As shown in Figure 3, the name entity “未央宫(Weiyang Palace)” is incorrectly translated as “Fuyang Palace” in the question, while correctly translated in the retrieved document. In addition, as we can see from the underlined parts, highly similar expressions in the question and the retrieved document are translated into largely different ones.

Compared to other words or phrases which occur more frequently in the training corpus, name entities are more flexible and various, and thus have worse translation results from prevailing Neural Machine Translation systems (Li et al., 2018). While some work has focused on solving this problem (Hassan et al., 2007; Jiang et al., 2007; Grundkiewicz and Heafield, 2018; Li et al., 2018), it remains largely underresearched. With a translation system that handles name entities better, we can potentially obtain better results from translation-based methods.

6.3 Zero-shot Cross-lingual Method

Trained on pure English data without the involvement of machine translation systems, much effort has been saved using zero-shot cross-lingual methods. Moreover, a single model could be applied directly to various languages. Thus, compared to translation-based methods, zero-shot cross-lingual method seems to be a more practical way to build a cross-lingual OpenQA system.

Although trained and tested in different languages, the multilingual BERT model achieves relatively good results on the XQA dataset. This may indicate that multilingual BERT could transfer the ability of capturing some common interaction patterns between different text across different languages via pretraining a unified text encoder. To further investigate the cross-lingual transfer power of multilingual BERT, we examine the difference of reading comprehension performance between English and Chinese test sets, for “easy” questions and other questions respectively. Results in Table 10 show the performance gap between the source language and the target language for “easy” questions is much smaller than that for other questions. This may indicate that multilingual BERT better captures shallow matching information across different languages.

Despite multilingual BERT has been proved to have certain power in cross-lingual understanding, no parallel data is used in it. Another line of research extracts multilingual representation from the context vector of NMT models that are trained on parallel data (Schwenk and Douze, 2017; Artetxe and Schwenk, 2018), which may be complementary to multilingual BERT. Very recently, Lample and Conneau (2019) proposed a multilingual language model that leveraged both monolingual and parallel data. Incorporating monolingual and parallel data may help to improve the performance in cross-lingual OpenQA.

7 Conclusion

In this paper, we discuss the problem of cross-lingual open-domain question answering, and present a novel dataset XQA, which consists of a total amount of 90k question-answer pairs in nine languages.

We further examine the performance of two translation-based methods and one zero-shot cross-lingual method on the XQA dataset. The experimental results show that multilingual BERT achieves the best result in almost all target languages. The performance of translation-based methods can be increased by applying machine translation system that better translates name entities, while the multilingual BERT model may be improved by incorporating parallel data with monolingual data.

We hope our work could contribute to the development of cross-lingual OpenQA systems and further promote the research of overall cross-lingual language understanding.

Acknowledgement

This research is jointly supported by the NSFC project under the grant no. 61661146007 and the NExT++ project, the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of AAAI*, pages 5012–5019.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Cecil H Brown, Eric W Holman, and Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language*, pages 4–29.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval-2017*, pages 1–14, Vancouver, Canada.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of ACL*, pages 2358–2367, Berlin, Germany.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of ACL*, pages 1870–1879, Vancouver, Canada.

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *Proceedings of ACL*, pages 209–220, Vancouver, Canada.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of ACL*, pages 845–855, Melbourne, Australia.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of ICLR*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485, Brussels, Belgium.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of ACL*, pages 593–602, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of ACL*, pages 1832–1846, Vancouver, Canada.

Cristina Espana-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1340–1350.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of ICML*, volume 37, pages 748–756, Lille, France.

B Green, A Wolf, C Chomsky, and K Laughery. 1986. Readings in natural language processing. pages 545–549, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Roman Grundkiewicz and Kenneth Heafield. 2018. Neural machine translation techniques for named entity transliteration. In *Proceedings of the Seventh Named Entities Workshop*, pages 89–94, Melbourne, Australia. Association for Computational Linguistics.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In *Proceedings of Workshop in AMML.*

Long Jiang, Ming Zhou, Lee-Feng Chien, and Cheng Niu. 2007. Named entity translation with web mining and transliteration. In *Proceedings of IJCAI*, pages 1629–1634, San Francisco, CA, USA.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING*, pages 1459–1474, Mumbai, India.

Cody Kwok, Oren Etzioni, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. *ACM Transactions on Information Systems*, 19(3):242–262.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR*.

Zhongwei Li, Xuancong Wang, AiTi Aw, Eng Siong Chng, and Haizhou Li. 2018. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of ACL*, pages 1736–1745, Melbourne, Australia.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu, and Richard Sutcliffe. 2006. Overview of the clef 2006 multilingual question answering track. In *Proceedings of Workshop of CLEF*, pages 223–256. Springer.

Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten De Rijke, Paulo Rocha, Kiril Simov, and Richard Sutcliffe. 2004. Overview of the clef 2004 multilingual question answering track. In *Proceedings of Workshop of CLEF*, pages 371–391. Springer.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of ACL*, pages 1725–1735, Melbourne, Australia.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL*, pages 3795–3805, Minneapolis, Minnesota.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of LREC*, Miyazaki, Japan.

Alessandro Vallin, Bernardo Magnini, Danilo Giampiccolo, Lili Aunimo, Christelle Ayache, Petya Osenova, Anselmo Peñas, Maarten De Rijke, Bogdan Sacaleanu, Diana Santos, et al. 2005. Overview of the clef 2005 multilingual question answering track. In *Proceedings of Workshop of CLEF*, pages 307–331. Springer.

Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *Proceedings of TREC*, pages 77–82.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018a. R^3: Reinforced ranker-reader for open-domain question answering. In *Proceedings of AAAI*, pages 5981–5988.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2018b. Evidence aggregation for answer re-ranking in open-domain question answering. In *Proceedings of ICLR*.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017. THUMT: An open source toolkit for neural machine translation. *arXiv preprint arXiv:1706.06415*.

Meng Zhang, Yang Liu, Huanbo Luan, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2016. Building earth mover's distance on bilingual word embeddings for machine translation. In *Proceedings of AAAI*, pages 2870–2876. AAAI Press.