# Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension

**Fei Yuan**[1*] **Linjun Shou**[2†] **Xuanyu Bai**[2] **Ming Gong**[2] **Yaobo Liang**[3]
**Nan Duan**[3] **Yan Fu**[1] **Daxin Jiang**[2]
[1]University of Electronic Science and Technology of China
[2]STCA NLP Group, Microsoft, Beijing, China
[3]Microsoft Research Asia, Beijing, China
feiyuan@std.uestc.edu.cn
{lisho,xub,migon,yalia,nanduan,djiang}@microsoft.com
fuyan@uestc.edu.cn

## Abstract

Multilingual pre-trained models could leverage the training data from a rich source language (such as English) to improve the performance on low resource languages. However, the transfer effectiveness on the multilingual Machine Reading Comprehension (MRC) task is substantially poorer than that for sentence classification tasks, mainly due to the requirement of MRC to detect the word level answer boundary. In this paper, we propose two auxiliary tasks to introduce additional phrase boundary supervision in the fine-tuning stage: (1) a mixed MRC task, which translates the question or passage to other languages and builds cross-lingual question-passage pairs; and (2) a language-agnostic knowledge masking task by leveraging knowledge phrases mined from the Web. Extensive experiments on two cross-lingual MRC datasets show the effectiveness of our proposed approach.

## 1 Introduction

Machine Reading Comprehension (MRC) plays a critical role in the assessment of how well a machine could understand natural language. Among various types of MRC tasks, the span extractive reading comprehension task (like SQuAD (Rajpurkar et al., 2016)) has been become very popular. Promising achievements have been made with neural network based approaches (Seo et al., 2017; Wang et al., 2017; Xiong et al., 2018; Yu et al., 2018; Hu et al., 2017), especially those built on pre-trained language models such as BERT (Devlin et al., 2018), due to the availability of large-scale annotated corpora (Hermann et al., 2015; Rajpurkar et al., 2016; Joshi et al., 2017). However, these large-scale annotated corpora are mostly exclusive to English, while research about MRC on languages other than English (i.e. multilingual MRC) has been limited due to the absence of sufficient training data.

| Language | MRC EM | MRC Gap to en | NLI ACC | NLI Gap to en |
|---|---|---|---|---|
| en | 62.4 | – | 85.0 | – |
| es | 49.8 | -12.6 | 78.9 | -6.1 |
| de | 47.6 | -14.8 | 77.8 | -7.2 |
| ar | 36.3 | -26.1 | 73.1 | -11.9 |
| hi | 27.3 | -35.1 | 69.6 | -15.4 |
| vi | 41.8 | -20.6 | 76.1 | -8.9 |
| zh | 39.6 | -22.8 | 76.5 | -8.5 |

Table 1: The gap between target languages and English on Machine Reading Comprehension (MRC) (Lewis et al., 2019) is significantly larger than sentence level classification task like Natural Language Inference (NLI) (Conneau et al., 2018). In this experiment, we fine-tune XLM (Conneau and Lample, 2019) on English and directly test on other languages.

To alleviate the scarcity of training data for multilingual MRC, the translation based data augmentation approaches were firstly proposed. For example, (question $q$, passage $p$, answer $a$) in English SQuAD can be translated into ($q'$, $p'$, $a'$) in other languages (Asai et al., 2018) to enrich the non-English MRC training data. However, these approaches are limited by the quality of the translators, especially for those low resource languages.

Most recently, approaches based on multilingual/cross-lingual pre-trained models (Devlin et al., 2018; Lample and Conneau, 2019; Huang et al., 2019; Yang et al., 2019) have proved very effective on several cross-lingual NLU tasks. These approaches learn language-agnostic features and align language representations in vector space during multilingual pre-training process (Wang et al., 2019; Castellucci et al., 2019; Keung et al., 2019;

[*]Work is done during internship at STCA NLP Group, Microsoft.
[†]Correspondence author.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 925–934
July 5 - 10, 2020. 

| |
|---|
| **[Question]:** who were the kings of the southern kingdom<br>**[Passage]:** In the southern kingdom there was only one dynasty, that of king David, except usurper Athaliah from the northern kingdom, who by marriage, [. . . ]<br>**[Answer - ground truth]:** king David<br>**[Answer - model predication:]** David, except usurper Athaliah |
| **[Question]:** What is the suggested initial does dosage of chlordiazepoxide<br>**[Passage]:** If the drug is administered orally, the suggested initial dose is 50 to 100 mg, to be followed by repeated doses as needed until agitation is controlled – up to 300 mg per day. [. . . ]<br>**[Answer - ground truth]:** 50 to 100 mg<br>**[Answer - model predication:]** 100 mg |

Table 2: Bad answer boundary detection cases of multilingual MRC model.

Jing et al., 2019; Cui et al., 2019). On top of these cross-lingual pre-trained models, zero-shot learning with English data only, or few-shot learning with an additional small set of non-English data derived from either translation or human annotation, can be conducted. Although these methods achieved significant improvement in sentence level multilingual tasks (like XNLI task (Conneau et al., 2018), the effectiveness on phrase level multilingual tasks is still limited. As shown in Table 1, MRC has bigger gap compared with sentence level classification tasks, in terms of the gap between non-English languages and English. To be specific, the EM metrics for non-English languages have 20+ points gap with the counterpart of English on average.

For extractive MRC, the EM metric is very critical since it indicates the answer boundary detection capability, i.e. the accuracy for extractive answer spans. In Table 2, there are two multilingual MRC cases with wrong boundary detection. In real scenarios, these bad extractive answers will bring negative impact to user experience. One interesting finding after case study is that the multilingual MRC model could roughly locate the correct span but still fail to predict the precise boundary (e.g. missing or adding some words in the spans as the cases in Table 2). For example, an error analysis of XLM on MLQA (Lewis et al., 2019) showed about 49% errors come from answers that partially overlap with golden span. Another finding is that a large amount ($\sim$ 70% according to MLQA) of the extractive spans are language-specific phrases (kind of broad knowledge, such as entities or N-grams noun phrases). We call such phrases *knowledge phrase* in the rest of paper, and will leverage them as prior knowledge in our model.

Motivated by the above observations, we propose two auxiliary tasks to enhance boundary detection for multilingual MRC, especially for low-resource languages. First, we design a cross-lingual MRC task with mixed-languages $\langle$question, passage$\rangle$ pairs to better align the language representation. We then propose a knowledge phrase masking task as well as a language-agnostic method to generate per-language knowledge phrases from the Web. Extensive experiments on two multilingual MRC datasets show that our proposed tasks could substantially boost the model performance on answer span boundary detection. The main contributions of our paper can be summarized as follows.

- We design two novel auxiliary tasks in multi-task fine-tuning to help improve the accuracy of answer span boundary detection for multilingual MRC model.
- We propose a language-agnostic method to mine language-specific knowledge phrase from search engines. This method is light-weight and easy to scale to any language.
- We conduct extensive experiments to prove the effectiveness of our proposed approach. In addition to an open benchmark dataset, we also create a new multilingual MRC dataset from real-scenario together with fine-grained answer type labels the in-depth impact analysis.

## 2 Related Work

### 2.1 Multilingual Natural Language Understanding (NLU)

A straightforward approach is leveraging translation to translate training data in rich resource language to low resource language. Asai et al. (2018) proposed to use run-time machine translation for multilingual extractive reading comprehension. Cui et al. (2019) developed several back-translation methods for cross-lingual MRC. Singh et al. (2019) introduced a translation-based data augmentation mechanism for question answering. However, these methods highly depend on the availability and quality of translation systems.

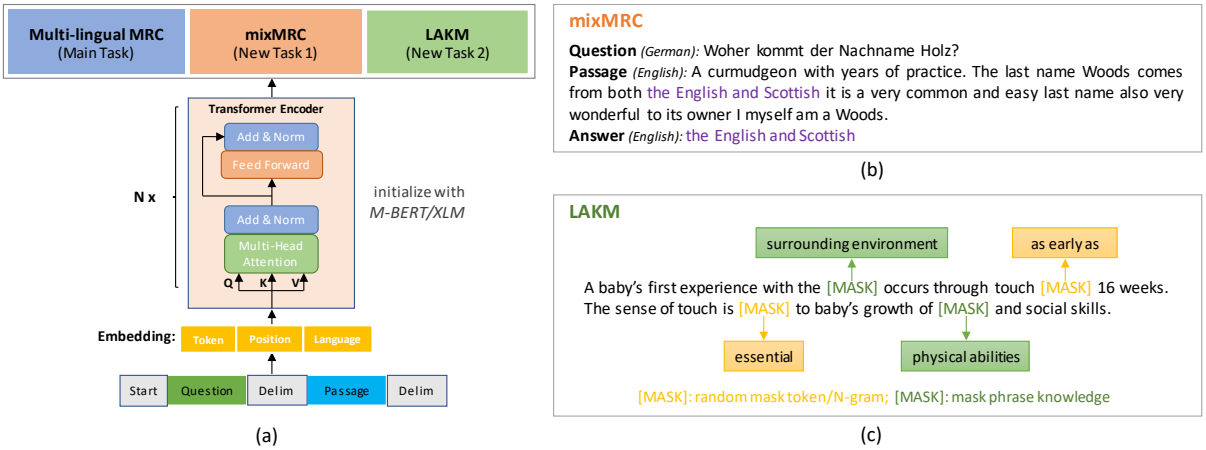


Figure 1: Overview of enhancing answer boundary detection work for multilingual machine reading comprehension. Our approach consists of three tasks: (a) Main task: multilingual MRC model requires to read text material and answer the question based on given context; (b) mixMRC task: cross-lingual MRC task with mix-language ⟨question, passage⟩ pairs; (c) LAKM task: A language-agnostic knowledge masking task by leveraging language-specific knowledge mined from web.

Another approach to Multilingual NLU extracts language-independent features to address multilingual NLU tasks. Some works (Keung et al., 2019; Jia and Liang, 2017; Chen et al., 2019) apply adversarial technology to learn language-invariant features and achieve significant performance gains. More recently, there has been an increasing trend to design cross-lingual pre-trained models, such as multilingual BERT (Devlin et al., 2018), XLM (Lample and Conneau, 2019), and Unicoder (Huang et al., 2019), which showed promising results due to the capability of cross-lingual representations in a shared contextual space (Pires et al., 2019). In this paper, we propose two novel sub-tasks in fine-tuning cross-lingual models for MRC.

### 2.2 Knowledge based MRC

Prior works (Yang and Mitchell, 2017; Mihaylov and Frank, 2018; Weissenborn et al., 2017; Sun et al., 2018) mostly focus on leveraging structured knowledge from knowledge bases (KBs) to enhance MRC models following a retrieve-then-encode paradigm, i.e., relevant knowledge from KB are retrieved first and sequence modeling methods are used to capture complex knowledge features. However, such a paradigm often suffers from the sparseness of knowledge graphs.

Recently, some works fuse knowledge into pre-trained models to get knowledge enhanced language representation. Zhang et al. (2019) uses both large-scale textual corpora and knowledge graphs to train an enhanced language representation. Sun et al. (2019) construct unsupervised pre-trained tasks with large scale data and prior knowledge to help the model efficiently learn the lexical, syntactic and semantic representations, which significantly outperforms BERT on MRC.

Most previous works on knowledge-based MRC are limited to English only. Meanwhile the requirement of acquiring large-scale prior knowledge (such as entity linking, NER models) may be challenging to meet for non-English languages. In this work, we propose a light-weight language-agnostic knowledge phrase mining approach and design a knowledge phrase masking task to boost the model performance for *multilingual* MRC.

## 3 Approach

In this section, we first introduce the overall training procedure, and then introduce two new tasks, namely, Mixed Machine Reading Comprehension (mixMRC) and Language-agnostic Knowledge Phrase Masking (LAKM), respectively.

The overview of our training procedure is shown at Figure 1. Our approach is built on top of popular multilingual pre-trained models (such as multilingual BERT and XLM). We concatenate passage, question (optional) together with special tokens `[Start]` and `[Delim]` as the input sequence of our model, and transform word embedding into contextually-encoded token representations using transformer. Finally, this contextual

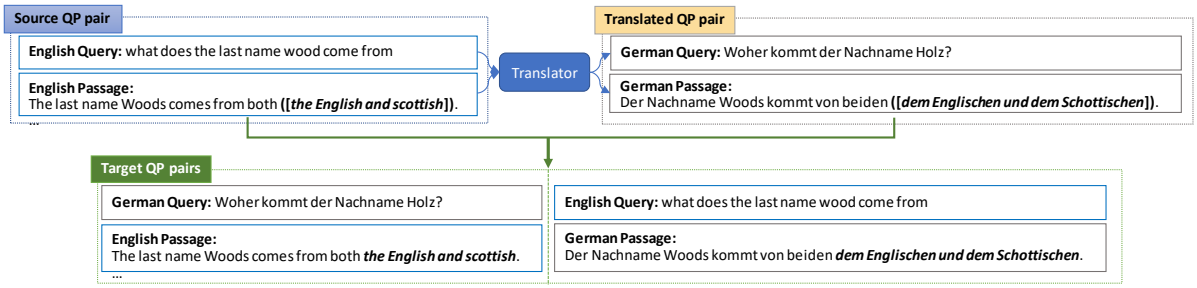


Figure 2: MixMRC data generation process. Given source (English) QP pair, we translate QP pair from English into non-English. Then the target mix-language pair can be divided into two forms: translated question-source passage and source question and translated passage pair.

representation is used for all three tasks introduced as following.

The first task, also our main task, is multilingual MRC, which aims to extract answers spans from the context passage according to the question. In this task, each language has its own data. However, only English has human labeled training data, and the other languages use machine translated training data from English. During training, the MRC training data in all languages will be used together for fine-tuning.

In the following, we introduce our new proposed tasks which will jointly train with our main task to boost multilingual MRC performance.

### 3.1 Mixed Machine Reading Comprehension (mixMRC)

We propose a task, named mixMRC, to detect answer boundaries even when ⟨question, passage⟩ are in different languages, which is shown in Figure 1 (b). It is mainly motivated by the strategy of data augmentation (Singh et al., 2019). In detail, we utilize the mixMRC to derive more accurate answer span boundaries according to the constructed ⟨question, passage⟩ pairs.

The way to obtain ⟨question, passage⟩ pairs consists of two steps: 1) translate training data from English into non-English; 2) construct mix-language training data for mix-MRC task. We show the entire data generation process in the Figure 2.

**Step 1: Data Translation** When using machine translation system to translate paragraphs and questions from English into non-English, the key challenge is how to address the answer span in translation.

To solve this problem, we enclose the answer text of source passage in special token pair "([" and "])", similar to (Lee et al., 2018). After translation, we discard training the instances where the translation model does not map the answer into a span well. Some skip data can still be recalled by finding the translated answer in the translated passage. The statistics of translated data are shown in Table 3.

Formally, given a monolingual dataset $D = \{(q_i, p_i, a_i)\}$ where $q_i$, $p_i$ and $a_i$ mean the query, passage and answer of language $i$ respectively. We apply a public translator and create a translated dataset $D' = \{(\tilde{q_j}, \tilde{p_j}, \tilde{a_j})\}$, where $\tilde{q_j}$ is the translation of $q_i$, and $\tilde{a_j}$ is the answer span boundary in $\tilde{p_j}$.

| | MTQA | | MLQA | |
|---|---|---|---|---|
| | # instance | skip ratio | # instance | skip ratio |
| en | 56616 | - | 87599 | - |
| fr | 52502 | 0.0727 | - | - |
| de | 51326 | 0.0934 | 80284 | 0.0835 |
| es | - | - | 87134 | 0.0053 |

Table 3: The statistics of translated data. The skip ratio is the percentage of those cases which are discarded.

**Step 2: Mix Language** After translation, we create a mixed-language dataset $D'' = \{(\tilde{q_k}, \tilde{p_l}, \tilde{a_l})\}$ where $l \neq k$. This could encourage MRC model to distinguish the phrases boundary by answer span selection and also keep the alignment of the underlying representations between two languages. In this task, we use the same fine-tuning framework as in monolingual MRC task.

### 3.2 Language-agnostic Knowledge Phrase Masking (LAKM)

In this section, we first introduce the approach for mining knowledge phrases from the Web. We then introduce the masking task created with these knowledge phrases.

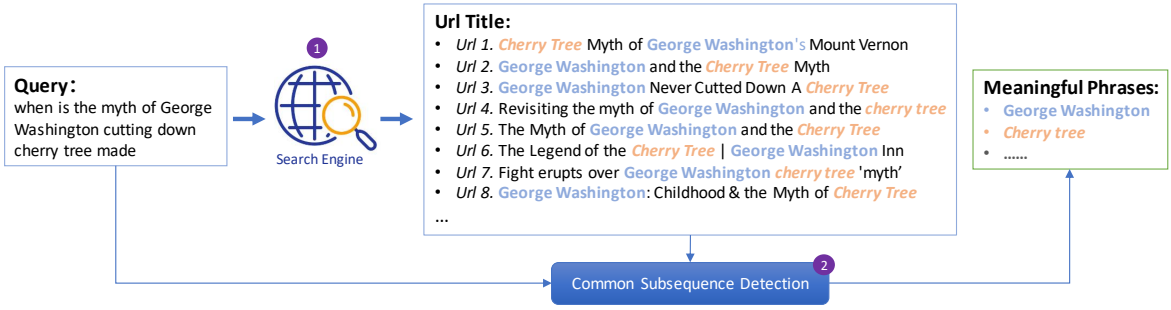


Figure 3: The process to generate knowledge data.

**Data Generation** In the following, we will describe our data generation method to collect large-scale phrase knowledge for different languages. The source data comes from a search engine, consisting of queries and the top N relevant documents. Let us take a running example of query {`when is the myth of George Washington cutting down cherry tree made`}. As shown in Figure 3, our mining pipeline consists of two main steps:

1. Phrase Candidates Generation: This step targets at high recall. We enumerate all the n-grams (n=2,3,4) of the given query as phrase candidates, such as `when is, the myth, George Washington, cherry tree, is the myth, etc.` We further filter the candidates with a stop word list. A manual analysis (by asking humans to identify all meaningful n-gram phrases in the given queries) shows that recall reaches $\sim 83\%$.

2. Phrase Filtering: This step targets at high precision by removing useless phrases. For each candidate, we count its frequency in the titles of relevant documents. We only keep those frequent candidates. For example, phrases `George Washington, cherry tree` appear in every title. We name them as *knowledge phrases*. Our empirical study suggests a frequency of 0.7 results in a good balance between precision and recall, and we use this threshold in our approach.

Following this approach, large amount of meaningful phrases can be mined independent of languages. After this, we further extract the passages which contain the mined knowledge phrases from the documents (following similar passage creation approach proposed by Rajpurkar et al. (2016)), which is the input of the LAKM. For the purpose of fair comparisons, the number of passages in different languages is equal, and the total amount of training data in LAKM is the same as that of mixMRC. The statistics of the knowledge phrases are given in Table 4.

| | en | fr | de | es |
|---|---|---|---|---|
| # P | 99.7k | 91.2k | 93.8k | 78.8k |
| # K-phrases | 229k | 102k | 102k | 101k |
| Avg. K-words | 2.14 | 2.36 | 2.18 | 2.19 |
| Avg. K-phrases / P | 2.29 | 1.11 | 1.09 | 1.28 |

Table 4: Statistics of the knowledge data we used: (1) **# P**: the total number of passage, (2) **# K-phrases**: the total number of knowledge phrases, (3) **Avg. K-words**: the average number of words per knowledge phrase and (4) **Avg. K-phrases / P**: the average number of phrases per passage.

**Model Structure** Given a ⟨passage, knowledge phrases⟩ pair, denoted as $(X, Y)$, we formalize that $X = (x_1, x_2, \ldots, x_m)$ is a passage with $m$ tokens, $Y = (y_1, y_2, \ldots, y_n)$ is a set of language-specific knowledge phrases generated as before, where $y_i = (x_j, x_{j+1}, \ldots, x_{j+(l-1)})(1 \leq j \leq m)$, $l$ is the number of tokens in $y_i(1 \leq i \leq n)$. The representations $h_\theta$ can be easily obtained from transformer. To inject language-specific knowledge into multilingual MRC model, we use masked language model as the fine-tuning objective. This task-specific loss has an additional summation over the length of sequence:

$$p_t = \text{Softmax}(W h_\theta(x)_t + b) \quad (1)$$

$$L_{LAKM} = \sum_{k=1}^{m} -y_{kt}^T \log p_t \quad (2)$$

where $p_t$ is the prediction value of $t^{th}$ word, $m$ is the number of tokens in the input passage, $y_{kt}$ is the target word, $W, b$ are the output projections for the task-specific loss $L_{LAKM}$, and $h_\theta(x)_t$ refers to the pre-trained embedding of the $t^{th}$ word.

## 4 Experiments

In this section, we firstly describe the dataset and evaluation in Section 4.1; then introduce the baseline models in Section 4.2 and experiment setting in Section 4.3; thirdly the experimental results are shown in Section 4.4.

### 4.1 Dataset and Evaluation

#### 4.1.1 Dataset

To verify the effectiveness of our approach, we conduct experiments on two multilingual datasets: one open benchmark called MLQA (Lewis et al., 2019); the other newly constructed multilingual QA dataset with multiple fine-grained answer types (MTQA).

**MLQA.** A multilingual question answering benchmark (Lewis et al., 2019). MLQA contains QA instances in 7 languages. Due to resource limitation, we evaluate our models on three languages (*English, German, Spanish*) of the dataset.

**MTQA.** To further evaluate our approach on real-scenario as well as conduct in-depth analysis of the impact on different answer types (in Section 5.3), we construct a new QnA dataset with fine-grained answer types. The construction process is described as following:

1. ⟨question, passage⟩ pairs come from the question answering system of one commercial search engine. Specifically, questions are real user searched queries on one commercial search engine, which are more diverse, covering various answer types. For each question, a QA system is leveraged to rank the best passage from the top 10 URLs returned by search engine. For each question, only the best passage is selected.

2. To annotate the answer span in each passage, we leverage crowd sourcing annotators for the labeling. Annotators are asked to first select the best shortest span* in the passage which can answer the question and also assign an answer type according to the query and the answer span. Each case are labeled by three annotators and those instances which are labeled with consensus (no less than two annotators agree on the result) are finally selected. An English example is given in Table 5.

*Only single span is considered.

Detailed statistics of MTQA dataset are given in Table 6 as well as the distribution of answer types in our dataset shown in Figure 4.

| |
|---|
| **[Question]:** how many players in rugby-league team on field |
| **[Passage]:** A rugby league team consists of thirteen players on the field, with four substitutes on the bench, [. . . ] |
| **[subtype]:** numeric |
| **[Answers:]**"start":41,"end":49,"text":"thirteen" |

Table 5: An English example of MTQA.

| | en | fr | de |
|---|---|---|---|
| # of dev instances | 6156 | 4900 | 3975 |
| # of test instances | 3017 | 2413 | 1893 |
| # of dev answer type | 58 | 57 | 55 |
| # of test answer type | 54 | 51 | 53 |

Table 6: Statistics of the dataset MTQA.

#### 4.1.2 Experimental Evaluation

We use the same evaluation metrics in the SQuAD dataset (Rajpurkar et al., 2016), i.e., *F1* and *Exact Match*, to evaluate the model performance. Exact Match Score measures the percentage of predictions that exactly match any one of the ground truths. F1 score is used to measure the answer overlap between predictions and ground truth. We treat the predictions and ground truth as bags of words, and compute their F1 score. For a given question, we select the maximum value of F1 over all of the ground truths, and then we average over all of the questions.

### 4.2 Baseline Models

We use the following two multilingual pre-trained models to conduct experiments:

- **M-BERT:** Multilingual version of BERT released by (Devlin et al., 2018) which is pre-trained with monolingual corpora in 104 languages. This model proves to be very effective at zero-shot multilingual transferring between different languages (Pires et al., 2019).

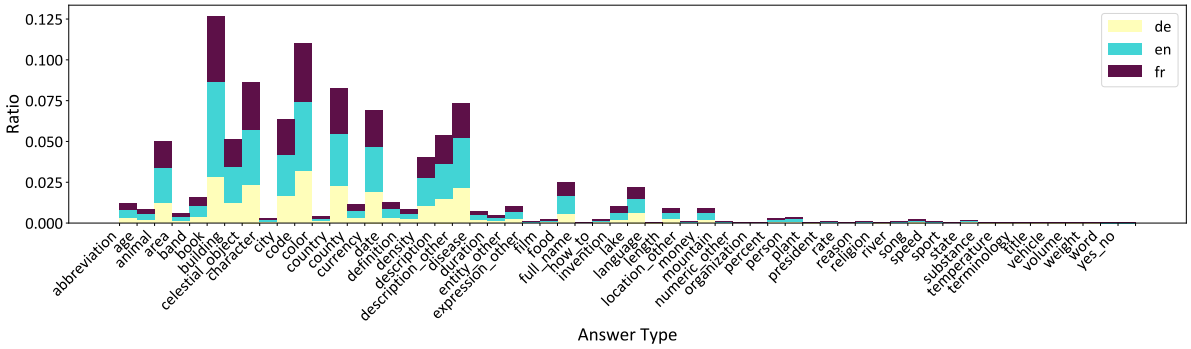


Figure 4: Answer type distribution in MTQA.

- **XLM:** A cross-lingual language model (15 languages) (Lample and Conneau, 2019) pre-trained with both monolingual data and cross-lingual data as well as cross-lingual tasks to enhance the transferring capacity among different languages.

For baseline, we directly fine-tune the pre-trained models using MRC training data only.

### 4.3 Experimental Setting

We use Adam optimizer with $\beta_1 = 0.9$ , $\beta_2 = 0.999$. The learning rate is set as 3e-5 for the mixMRC, LAKM and multilingual MRC tasks. The pre-trained model is configured with its default setting. Each of the tasks is trained until the metric of MRC task converges.

**mixMRC.** We jointly train mixMRC and multilingual MRC tasks using multi-task training at the batch level to extract the answer boundary in the given context. For both tasks, the max sequence length is 384.

**LAKM.** LAKM and multilingual MRC tasks are jointly trained using multi-task training. In terms of input, we randomly mask 15% of all WordPiece tokens in each sequence in a two step approach. Firstly, if the $i-th$ token belongs to a knowledge phrase, we replace the $i$- token with (1) the `[MASK]` token 80% of the time (2) a random token 10% of the time (3) the unchanged $i-th$ token 10% of the time. Secondly, if the proportion of knowledge phrase is less than 15%, we will further randomly mask other WordPiece tokens to make the total masked ratio to reach 15%. For LAKM, the max sequence length is set as 256.

**mixMRC + LAKM.** We jointly train mixMRC, LAKM and multilingual MRC tasks, take the gradients with respect to the multilingual MRC loss, mixMRC loss and LAKM loss, and apply the gradient updates sequentially at batch level. During the training, the max sequence length is 384 for multilingual MRC model, 256 for LAKM and 384 for mixMRC.

### 4.4 Experiment Results

The overall experimental results are shown in Table 7. Compared with M-BERT & XLM baselines, both mixMRC and LAKM have decent improvements in fr, es and de, and on-par performance in en in terms of both MLQA and MTQA datasets. This demonstrates the effectiveness of our models.

The combination of LAKM and mixMRC tasks gets the best results on both datasets. Take M-BERT and MLQA dataset as an example, mixMRC+LAKM have 1.7% and 4.7% EM improvements on es and de languages respectively, compared with baseline.

In terms of LAKM task, there are decent gains for all languages, including English. However, the gains are bigger on low resource languages compared with English performance. Take XLM and MLQA dataset as an example, LAKM gets 1.8% and 3.2% EM improvements on es and de, while the improvement on en is about 0.5%. The intuition behind en gains is that LAKM brings extra data with knowledge to en as well.

In terms of mixMRC task, there are slight regression on en compared with decent gains on es, de and fr. Take XLM and MTQA dataset for illustrations, mixMRC has 0.6% EM regression on en versus 1.4% and 0.5% EM gains on fr and de languages. This shows that mixMRC mainly improves the transferring capability from rich resource language to low resource language.

| Model | Methods | MLQA (EM / F1) en | MLQA (EM / F1) es | MLQA (EM / F1) de | MTQA (EM / F1) en | MTQA (EM / F1) fr | MTQA (EM / F1) de |
|---|---|---|---|---|---|---|---|
| M-BERT | Lewis et al. (2019) | 65.2 / 77.7 | 37.4 / 53.9 | 47.5 / 62.0 | - | - | - |
| | Baseline | 65.4 / 79.0 | 50.4 / 68.5 | 46.2 / 60.6 | 67.0 / 86.9 | 52.9 / 78.2 | 59.8 / 81.4 |
| | LAKM | **66.9** / 80.1 | 51.5 / 69.5 | 49.9 / 64.4 | **68.8** / 87.6 | 56.8 / 78.8 | 62.4 / 81.9 |
| | mixMRC | 65.4 / 79.4 | 50.5 / 69.1 | 49.1 / 64.0 | 67.9 / 86.8 | 56.4 / 77.8 | 62.4 / 81.0 |
| | mixMRC + LAKM | 64.7 / 79.2 | **52.1** / 70.4 | **50.9** / 65.6 | 68.6 / 87.0 | **57.5** / 78.5 | **62.9** / 81.3 |
| XLM | Lewis et al. (2019) | 62.4 / 74.9 | 47.8 / 65.2 | 46.7 / 61.4 | - | - | - |
| | Baseline | 64.1 / 77.6 | 50.4 / 68.4 | 47.4 / 62.0 | 67.1 / 86.8 | 51.5 / 75.8 | 61.6 / 81.3 |
| | LAKM | **64.6** / 79.0 | 52.2 / 70.2 | 50.6 / 65.4 | **68.3** / 87.3 | 52.5 / 75.9 | 61.9 / 81.2 |
| | mixMRC | 63.8 / 78.0 | 52.1 / 69.9 | 49.8 / 64.8 | 66.5 / 85.9 | 52.9 / 75.0 | 62.1 / 80.5 |
| | mixMRC + LAKM | 64.4 / 79.1 | **52.2** / 70.3 | **51.2** / 66.0 | 68.2 / 86.8 | **53.6** / 75.9 | **62.5** / 80.9 |

Table 7: Experimental results on MLQA and MTQA dataset under translation condition (%).

## 5 Analysis

In this section, we ablate important components in LAKM to explicitly demonstrate its effectiveness.

### 5.1 Random N-gram Masking vs LAKM

To study the effectiveness of LAKM, we compare LAKM with *Random N-gram Masking*[†] based on XLM and MTQA dataset. LAKM and Random N-gram Masking refer to fine-tuning XLM with the language-specific knowledge masking strategy and random n-gram masking strategy respectively. As shown in Table 8, without the language-agnostic knowledge masking strategy, the EM metrics drops by 0.2% - 0.87%, which proves the necessity of LAKM.

| Setting (EM) | en | fr | de |
|---|---|---|---|
| Random N-gram Masking | 67.5 | 51.8 | 61.7 |
| LAKM | 68.3 | 52.5 | 61.9 |

Table 8: Ablation study on MTQA (%).

### 5.2 Zero Shot Fine-tuning w/ vs w/o LAKM

To illustrate the effectiveness of the auxiliary tasks, an extreme scenario is considered when only English training data is available and there is no translation data. That means that we are unable to use mixMRC task to driver more accurate answer span boundaries. At this point, we only leverage LAKM to enhance answer boundary detection and compares the performance of M-BERT baseline with our model in Table 9.

From the experimental results, zero shot fine-tuning with LAKM is significantly better than M-BERT baseline. On MTQA, our model gets 2%, 3.3%, 3.8% EM improvements on English, French and German respectively. On MLQA, we get 1.6%, 1.4%, 1.2% EM improvements on English, Spanish and German.

| | MLQA (EM / F1) en | MLQA (EM / F1) es | MLQA (EM / F1) de |
|---|---|---|---|
| Baseline | 65.2 / 77.7 | 46.6 / 64.3 | 44.3 / 57.9 |
| LAKM | **66.8 / 80.0** | **48.0 / 65.9** | **45.5 / 60.5** |

| | MTQA (EM / F1) en | MTQA (EM / F1) fr | MTQA (EM / F1) de |
|---|---|---|---|
| Baseline | 65.8 / 86.6 | 41.3 / 70.9 | 50.7 / 76.2 |
| LAKM | **67.8 / 87.2** | **44.6 / 72.1** | **54.5 / 77.8** |

Table 9: Zero Shot experimental results on MLQA and MTQA datasets (%). We only use English MRC training data and don't use translation data.

### 5.3 Extensive Analysis on Fine-grained Answer Types

To have an insight that how the new tasks (LAKM/mixMRC) affect the multilingual MRC task, we further analyze model performance on various answer types, as shown in Figure 5.

The comparison with baseline indicates that in most of the answer types (like `color`, `description`, `money`), both LAKM and mixMRC can enhance the answer boundary detection for multilingual MRC task.

One interesting finding is that in terms of `animal`, `full_name`, LAKM outperforms mixMRC by a great margin, which are 9.1% and 14.3% respectively. One possible explanation is that the knowledge phrases of LAKM can cover some entity related phrases like animals and names, leading to the significant EM boost.

In terms of those numerical answer types (like `money`, `numeric`, `length`), the per-

[†]Random N-gram Masking shows gains in English SQuAD.

formance between mixMRC and LAKM are similar. The intuition behind this is that these numerical answers may be easier to transfer between different languages since answers like length are similar across different languages.

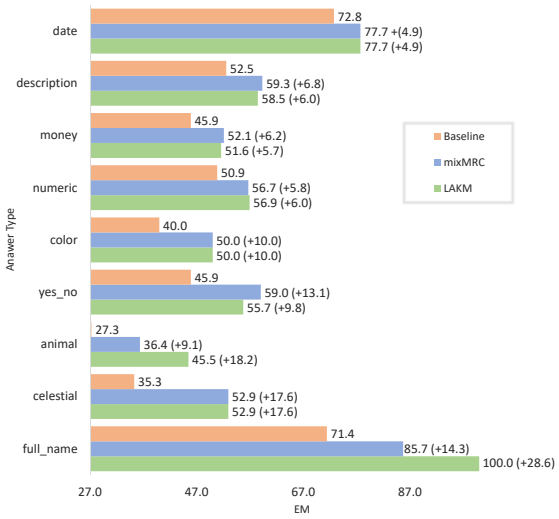


Figure 5: EM results comparison on M-BERT (MTQA French test set) for the different answer types.

## 6 Conclusion

This paper proposes two auxiliary tasks (mixMRC and LAKM) in the multilingual MRC fine-tuning stage to enhance answer boundary detection especially for low resource languages. Extensive experiments on two multilingual MRC datasets have been conducted to prove the effective of our proposed approach. Meanwhile, we further analyze the model performance on fine-grained answer types, which shows interesting insights.

## References


Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multilingual intent detection and slot filling in a joint bert-based model. *arXiv preprint arXiv:1907.02884*.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. *arXiv preprint arXiv:1909.00361*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karl Moritz Hermann, Tomas Kocisky, 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*, pages 1693–1701.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2017. Reinforced mnemonic reader for machine reading comprehension. *arXiv preprint arXiv:1705.02798*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Yimin Jing, Deyi Xiong, and Yan Zhen. 2019. Bipar: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. *arXiv preprint arXiv:1910.05040*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and ner. *arXiv preprint arXiv:1909.00153*.

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

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised Training Data Generation for Multilingual Question Answering. page 5.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. *arXiv preprint arXiv:1805.07858*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hananneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Yibo Sun, Daya Guo, Duyu Tang, Nan Duan, Zhao Yan, Xiaocheng Feng, and Bing Qin. 2018. Knowledge based machine reading comprehension. *arXiv preprint arXiv:1809.04267*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019. Ernie 2.0: A continual pre-training framework for language understanding.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual bert transformation for zero-shot dependency parsing. *arXiv preprint arXiv:1909.06775*.

Zhiguo Wang, Haitao Mi, Wael Hamza, and Radu Florian. 2017. Multi-perspective context matching for machine comprehension. *arXiv preprint arXiv:1612.04211*.

Dirk Weissenborn, Tomáš Kočiskỳ, and Chris Dyer. 2017. Dynamic integration of background knowledge in neural nlu systems. *arXiv preprint arXiv:1706.02596*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2018. Dynamic coattention network for question answering. US Patent App. 15/421,193.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in LSTMs for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1436–1446, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.