Multilingual Neural Question Answering on MS-MARCO and DuReader
Team Name: AngryFrog

Zihua Liu
zihual@andrew.cmu.edu

Huiming Jin
huimingj@andrew.cmu.edu

Hui Luo
huiluo@andrew.cmu.edu

Yihui Peng
yihuip@andrew.cmu.edu

1 Motivations

Low-resource natural language processing remains many interests for many tasks. Neural models define the state of the art for Machine Reading Comprehension (MRC) and Question Answering (QA), however, those models are in general very data-hungry, and do not reach good performances in low-resource settings. Living in a multilingual world surrounded by thousands of tongues and dialects, we always wonder if we can build a bridge that connects worlds and civilizations, a core system that works regardless of language types, since adequate annotation is high-cost for most languages. Multi-task learning is one solution for this problem. With this desire in mind, we would like to build a general question answering system that works on multiple datasets in different languages, where the knowledge for machine reading comprehension can be shared among those languages. We expect that although Chinese and English are two quite different languages, the nature of MRC and QA will keep the same.

We choose MS-MARCO dataset in English and DuReader dataset in Chinese. These two datasets have similar structures and features so that it is convenient to transfer models and conduct reasonable experiments. They both consist of real-world generated questions in various domains and have linguistic diversity, which helps preserve original language features. Moreover, they have reliable answers written by real humans with special care, so it could be beneficial to develop new models and perform evaluations. The datasets are large scale, allowing us to train rather complicated neural network models and keep on exploring.

2 Dataset Description

2.1 MS-MARCO

Microsoft Machine Reading Comprehension dataset was first released at NIPS 2016 (Nguyen et al., 2016) and then expanded to 1,010,916 queries in the current version in 2018. All questions were generated from Bing usage logs, and they are classified into different categories according to answer topics, including numeric, entity, location, person and description. From the perspective of question types, most of them are what questions, while there are also how, yes/no, where, when ones. It is worth pointing out that all queries have 10 passages retrieved by Bing that are relevant to the query, among which both URLs and texts present. In addition, all queries have human-written answers, and even revised answers for selected queries. The data format includes query id, the query itself, passages, query type, and answers given by real humans. Table 1 shows an example query of MS-MARCO.

Item	Content
query_id	1102432
query	what is a corporation?
query_type	DESCRIPTION
passages	
is_selected	0
url	http://www.wisegeek.com/what-is-a-corporation.htm
passage_text	A company is ... or indirectly.
answers	A corporation is a company or group of people authorized to act as a single entity and recognized as such in law.

Table 1: An example of MS-MARCO. Note that while only one passage and one answer are presented as the example here, there are multiple passages and might be more than one answers provided in the dataset.

2.2 DuReader

DuReader (He et al., 2018) is a large-scale Chinese MRC dataset extracted from real-world and is released by Baidu Inc, a Chinese search engine company. Similar to MS-MACRO, each entry of DuReader contains question text, question type,

	Amount	**Average Length**
Question	301,574	26
-Description-Fact	34.6%	–
-Description-Opinion	17.8%	–
-Entity-Fact	23.4%	–
-Entity-Opinion	8.5%	–
-Yes_No-Fact	8.2%	–
-Yes_No-Opinion	7.5%	–
Document	1,431,429	1,793
Answer	665,723	299

Table 2: Data statistics of DuReader (He et al., 2018).

evidence documents, and reference answers. The contents of DuReader are all real, where the questions are from Baidu web user queries, the evidence documents are retrieved by Baidu search engine, and the answers are human-generated. The questions of DuReader are categorized into six types by two axes, where one axis contains Description, Entity, Yes_No, and the other one contains Fact and Opinion. Table 2 shows the statistics about the data, and Table 3 lists examples of different question categories.

DuReader is organized in a similar way as MS-MARCO, which makes it feasible to make use both of them. However, compared to MS-MARCO and many other MRC datasets, DuReader emphasizes more on Yes_No and Opinion questions which are more difficult to handle but are more close to reality.

Category	**Question**
Description-Fact	消防车为什么是红的 (Why are firetrucks red)
Description-Opinion	丰田卡罗拉怎么样 (How is Toyota Carola)
Entity-Fact	iphone哪天发布 (On which day will iphone be released)
Entity-Opinion	2017最好看的十部电影 (Top 10 movies of 2017)
Yes_No-Fact	39.5度算高烧吗 (Is 39.5 degree a high fever)
Yes_No-Opinion	学围棋能开发智力吗 (Does learning to play go improve intelligence)

Table 3: Examples of different question categories in DuReader (He et al., 2018). Note that the English translations are not provided.

3 Baseline Models

We plan to implement two typical state-of-the-art neural models BiDAF (Seo et al., 2016) and Match-LSTM (Wang and Jiang, 2016) as baselines, both of which are designed for MRC. We will run each model on both datasets and compare their performance as well as their generalization ability on different languages. For our initial sentence and passage encoder, we will use the 300-D pre-trained Glove embeddings (Pennington et al., 2014) for MS-MARCO dataset and we will follow the preprocessing steps described in He et al. (2018) for DuReader.

3.1 BiDAF

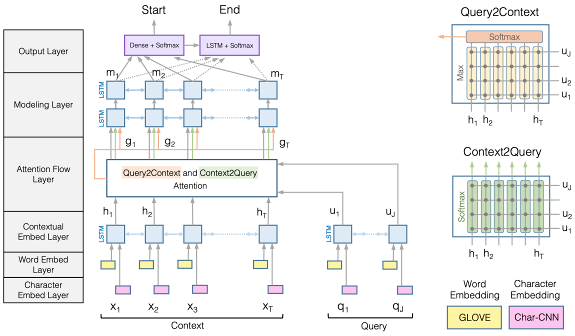
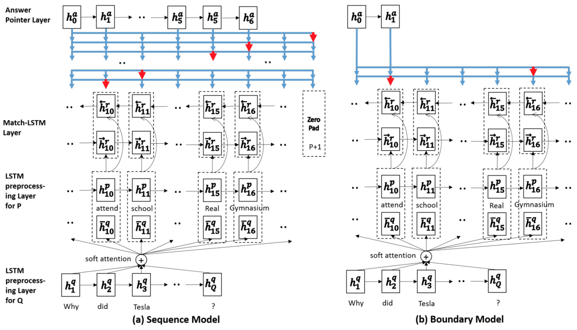


Figure 1: BiDAF model (Seo et al., 2016)

The attention mechanism has been wildly applied to MRC tasks and has achieved fairly good performance. Unlike the previous models which only employ unidirectional attention, BiDAF uses both context-to-query attention and query-to-context attention to highlight the important components in both queries and questions. The calculation of the attentions are independent on time series and will flow into the following layer, thus avoid information loss due to early summarizing. After the calculation of the attentions, the following attention flow layer is used to employ all useful information to get a vector representation for each position.

3.2 Match-LSTM

Match-LSTM is another promising MRC model. It is the first end-to-end neural model on SQuAD dataset (Rajpurkar et al., 2016) and it achieves a better performance than the original non-neural approach proposed when SQuAD was released. To find an answer in a passage, Match-LSTM goes through the passage sequentially and dynamically aggregates the matching of an attention-weighted question representation to each token in the passage. Finally, it uses an answer pointer layer to



(a) Sequence Model

(b) Boundary Model

Figure 2: Match-LSTM model (Wang and Jiang, 2016)

choose tokens within the passage to form an answer span.

4 Proposed Hypothesis

- **Build a general model using Transfer Learning regardless of the language.** We hope to use appropriate language models and word embedding methods to map the Chinese dataset and English dataset into similar data distributions, so that we can use a general model to train these two datasets. We expect that the model could sufficiently share the knowledge of reading comprehension among different languages. Because Chinese and English are very different, we could try to use shared or non-shared language models since the resource for training language models is relatively easier to obtain than MRC datasets. We also try to investigate the following questions:
  - What does the neural model learn from the other language?
  - Is sharing a language model important for the transfer?
  - How much of the transfer learning can be reduced to a regularization effect achieved by multi-task learning?
- **Apply pre-trained BERT for word embedding to improve the performance.** Pre-trained representations can either be context-free or contextual, and contextual representations can be unidirectional or bidirectional. Unlike some context-free models such as word2vec, BERT represents a word using both its previous and next context, making it deeply bidirectional, so it will perform better on word embedding and improve our baseline model.
- **Add an answer boundary prediction module to decrease the search space.** To extract the answer span from passages, we compute the probability of each word to be the start or end position of the span, so that we can predict the boundary of the answer in every passage and decrease the search space.
- **Apply cross passage answer verification.** The boundary model focus on extracting the answer within a single passage respectively, with little consideration of the cross-passage information. Therefore, we can enable the answer candidates from different passages to exchange information and verify each other through the cross-passage answer verification process.

5 Evaluation

Evaluation of our system will be done using the MS-MARCO leaderboard standard BLEU and ROUGH-L, which enables us to compare the performance of our model to the most state-of-the-art result.

6 Schedule and Work Distribution

6.1 Tentative Timeline

- Feb 13^{th} - Feb 25^{th}: Explore the datasets and implement two baseline models.
- Mar 4^{th} - Mar 11^{th}: Analyze the performance of the models on two datasets, recognize the difference of language features between English and Chinese.
- Mar 18^{th} - Apr 1^{st}: Design and train language models specifically for our datasets. Implement our proposed hypothesis.
- Apr 8^{th} - Apr 15^{th}: Do error analysis and fine tuning on our proposed system.
- Apr 22^{th} - End of the semester: Summarize the whole semester's work and finish a final report.

6.2 Individual Work

- **Zihua**: Implement BiDAF baseline and train the model on MS-MARCO. Apply transfer learning to test the model performance on

DuReader. Visualize the attention to check whether the model locates the correct position in the passage. Implement the answer boundary prediction layer to narrow down the search space in the passages for a given answer.

- **Huiming**: Implement Match-LSTM baseline and test the model on both datasets. Compare the performance of sequence model and boundary model to see which model is more suitable for our dataset. Build the cross-passage answer verification model to check if it can help the system to generate more reliable answers.

- **Hui**: Preprocess MS-MARCO dataset, apply 300-D pre-trained Glove embeddings for question and passage encodings and build a more robust English language model for the dataset if necessary. Use pre-trained BERT for embeddings and compare the performance with Glove embeddings.

- **Yihui**: Preprocess DuReader dataset, find an appropriate Chinese tokenizer for sentence segmentation and word recognition. Build a Chinese language model for DuReader. Apply pre-trained BERT-Base Chinese language model for encoding of DuReader.

References

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

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

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 Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.