Zero-Shot Cross-Lingual Discrete Reasoning

Stanford CS224N {Custom} Project

Yu Fan
Department of Computer Science
Stanford University
yuf@stanford.edu

Yang Tian
Department of Computer Science
Stanford University
yat@stanford.edu

Abstract

Discrete reasoning, including addition, subtraction, counting, sorting etc., remains a challenging task of machine reading comprehension. In addition, lack of parallel MRC data in languages other than English leads to increasing research interest on cross-lingual transfer learning. In light of studies from both sides, we tackle the task of zero-shot cross-lingual discrete reasoning using DROP data set and their manual translations in German and Chinese languages, and show that 1) multilingual BERT model can be configured to solve discrete reasoning tasks, and 2) the knowledge of discrete reasoning can be transferred cross-lingually in German and Chinese languages to certain extent, even without any available parallel training data.

1 Key Information to include

- Mentor: (As of Winter 2020: John Hewitt)
- External Collaborators (if you have any):
- Sharing project: None

2 Introduction

In the last years, research on machine reading comprehension (MRC), i.e., automated question answering by reading and interpreting a single passage of unstructured text, has experienced significant improvement. Related to this is the emergence of various benchmarking datasets, including SQuAD [1, 2], CoQA [3], etc., on which machines have achieved super-human performance. However, answering more complex questions, in particular the ones whose answers require discrete reasoning, including addition, subtraction, counting, sorting etc., still remains a challenging task for these machines. Thus, researchers have developed new data set and models to tackle this task, i.e. the DROP data set [4], where the questions force a structured analysis of the texts that permits discrete reasoning.

Moreover, the task of discrete reasoning mainly requires, in addition to the capability of information extraction, a machine's awareness of numeric entities as well as the ability of basic calculation, which is supposed to be independent of the input language in which such numeric entities are embedded. Thus, one may expect that a model's capability of solving the complex MRC task could be transferred into foreign languages. Although the textual data regarding this task, i.e. the DROP data set, is only available in English language, one remarkable feature of the data, compared to previous QA data sets, is that the answers of most questions are numbers. In so far, manually translating passages and their corresponding QA pairs does not require additional human annotation of correct answers. Furthermore, we find that DROP dataset is especially suitable for cross-lingual transfer for several reasons. First, it is more "sample-efficient", as an average passage entails significantly more questions than conventional MRC datasets, e.g. SQuAD. Second, DROP entails a variety of tasks related to natural language understanding, which allows us to experiment from different facets of cross-lingual transfer.

In our study, we attempt to tackle the problem of cross-lingual discrete reasoning using English samples from DROP data set and their paired manual translations in German and Chinese languages, to test the cross-lingual transferablity of number-awareness and the ability of calculation as well as comparison, while content and difficulty of the passages, together with their QA pairs, are taken into account. In particular, we apply Multilingual BERT model, proposed by Delvin et al.bert, to explore the ability of a system to handle discrete reasoning with multilinguality. First, we pre-train the Multilingual BERT on the English DROP dataset. Second, we fix the model and evaluate on a set of translated German and Chinese data. The results from our experiments for zero-shot learning show that a model's knowledge of discrete reasoning can be transferred into foreign languages to certain extent, even if there is no corresponding training samples in the target language available. More specifically, the model performs significantly better with German test samples than with Chinese, indicating the importance of language similarity in the transfer learning for the task.

3 Related Work

3.1 Question Answering Datasets

In the research field of reading comprehension, researchers have published various datasets in order to promote the ability of systems to complete question-answering tasks. A widely used dataset is Stanford Question Answering Dataset (SQuAD) [1, 2], where a system should select a text or a span from the passage as the answer. Unlike previous dataset, where a list of answer choices are considered as answer candidates for the question, the span-based answer in SQuAD is easier to evaluate. Whereas some datasets, like NewsQA [5] and TriviaQA [6], expects the answers to be spans as well, some other datasets, like CoQA [3] and MS MARCO [7] do not have this restriction and the answer could be free-form text. Specificaly, the MultiRC dataset [8] requires the system to answer a question correctly based on multiple sentences in the passage, and the HotpotQA [9] dataset targets on the performance on multi-hop reasoning on account of multiple paragraphs of the models.

3.2 Question Answering Systems

Researchers also have been developed models to complete tasks of reading comprehension. As an example, QANet is a question answering architecture proposed by Yu et al.qanet, which consists of only convolutional models and self-attention models. In the previous experiments of machine reading comprehension, recurrent neural networks were frequently applied in the models, for example Bidirectional Attention Flow (BiDAF) model [10] and FastQA model [11], which process sequential inputs, and the attention mechanism additionally deals with long term interactions. However, the recurrent model often leads to slow training and inference. The QANet removes the recurrent models and merely uses convolutions and self-attentions for the encoder, which captures the local structure of data and the global interaction between words, respectively. In the experiments on both SQuAD and TriviaQA datasets, the results of speedup comparison of QANet and various RNN models show that QANet is indeed significantly faster than all the RNN models. However, both question answering datasets and models merely concentrate on spans as answers, and still do not enable models to have a higher level of reading comprehension.

3.3 Multilingual BERT

The BERT model, which stands for Bidirectional Encoder Representations from Transformers, is currently a state-of-the-art pre-trained language model introduced by Devlin et al.bert. The model architecture of BERT is a multi-layer bidirectional Transformer and it is designed to be pre-trained to tackle a wide range of NLP tasks. In addition to the basic BERT model, its multilingual version is released as well. Multilingual BERT is pre-trained from monolingual corpora in 104 languages [1]. Pires et al.ml-bert prove that multilingual BERT model performs very well at zero-shot cross-lingual model transfer. However, Karthikeyan et al.cross-ml argue against the standpoint from Pires et al. that shared word-pieces between different languages play an important role in the good performance of Multilingual BERT, and state that the model is cross-lingual even when there is no word-piece overlap. In addition, various researchers implement specific experiments e.g. machine translation[12], text generation[13], reading comprehension[14] with Multilingual BERT as well.

[1] https://github.com/google-research/bert/blob/master/multilingual.md

4 Approach

We apply Multilingual BERT model on the original English DROP dataset to build our baseline model. To accomplish this, the training and dev set were used unchanged, provided by AllenNLP for training the Multilingual BERT model.

In the DROP paper [4], the authors established several baseline systems including fine-tuning BERT-base model for SQuAD-style reading comprehension task using huggingface [2], and the BERT model performs with EM 30.10 and F1 33.36 on the dev set. We adopted this framework and preprocessed the DROP data to convert them into SQuAD-style question-answer pairs by means of AllenNLP library. Different from what the authors did, we switched from BERT-base to Multilingual BERT model to train the model for the following cross-lingual zero-shot transfer learning. The result of our model yields EM 23.85 and F1 27.31, which was attained with batch size of 6 and training epoch of 3.

To compare, the performance of Multilingual BERT is slightly weaker than the BERT-base model. A potential reason might be that our implementation suffered from the memory restriction of the virtual machine. Besides, while the Multilingual BERT model we applied is currently only available in small version with 110M parameters, the experiment with BERT-base on reading comprehension was done in large version with 340M parameters.

5 Experiments

5.1 Data

5.1.1 DROP dataset

In this paper, we focus on the ability of discrete reasoning of models using DROP (Discrete Reasoning Over Paragraph) dataset [4], which contains 96,567 adversarially-created question-answer pairs. The DROP dataset is proposed to promote the comprehensive analysis of texts of the systems, and requires specifically discrete reasoning to answer the questions, such as addition, subtraction or counting. With the goal of encouraging systems producing comprehensive analyses of paragraphs, the authors are the first ones that attempt to combine more complex questions, to which the answers should be discovered based on multiple occurrences of an event in the questions, and in a form of spans in the passage, spans in the question, numbers, or dates, with paragraph understanding, rather than extracting the positions of answers given certain questions.

5.1.2 Data Conversion

As BERT model requires its text input as an answer in addition to the passage index of the first character of the answer, which is a different format of input from the answers in DROP dataset, we need to convert DROP to adapt BERT model. We used AllenNLP library[3] to convert the plain text in DROP. As for the numbers and dates, we firstly converted them to new fields in the SQuAD JSON file, and then enables BERT model to accept the JSON files as inputs and into its framework.

5.1.3 German and Chinese DROP data

The original DROP dataset was released in English. To complete the cross-lingual zero-shot learning task, we picked 12 passages from its training set and manually translated them and their question-answer pairs in German and Chinese for evaluation, that is, the 12 paired passages and QA pairs are used for zero-shot learning. The selection of the passages are supposed to be possibly diverse to ensure that all types of questions and topics are represented.

5.2 Evaluation method

We use both exact match score and (macro-averaged) F1 score to evaluate our model's performance on discrete reasoning in reading comprehension, which are applied in previous related papers [4, 1].

[2] https://github.com/huggingface/transformers

[3] https://github.com/allenai/allennlp

Answer Type	Passage (shortened)	Question	Answer
Number	The Vikings took their 2010 bye week in Week 4 of the season - their earliest bye week since the 2004 season. There were mixed results for their divisional rivals, with the Green Bay Packers beating the Detroit Lions by 2 points, while the Chicago Bears were beaten **17-3** by the New York Giants.	How many points did the Chicago Bears lose by?	14
Span	The Bengals started the season at home against the Ravens. **In the first quarter, the Ravens scored as Justin Tucker kicked a 25-yard field goal to make it 3-0. They would increase their lead in the second quarter when Joe Flacco found Jeremy Maclin on a 48-yard pass to make it 10-0.** This would be followed up by a 2-yard touchdown run by Terrence West to make the score 17-0 at halftime.	Did the Ravens score fewer points in the first or second quarter?	first
Multiple Spans	As of the 2010 United States Census, there were 1,951,269 people, 715,365 households, and 467,916 families residing in the county. The population density was. There were 840,343 housing units at an average density of. **The racial makeup of the county was 60.9% white, 10.5% black or African American, 8.7% Asian, 0.7% Pacific islander, 0.7% American Indian, 13.5% from other races, and 5.1% from two or more races.** Those of Hispanic or Latino origin made up 29.1% of the population.	Which two racial groups make up 0.7% of the population?	Pacific islander, American Indian
Date	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Sibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack **the next day**.	What date did the JNA form a battlegroup to counter-attack after the village of Nos Kalik was captured?	3 March 1992

Table 1: Examples of passage and question-answer pairs as well as answer type from the DROP dataset

5.3 Experimental details

In order to reveal how well cross-lingual zero-shot transfer works, we need to set a reference for the performance of the model on different languages. Therefore, as test set, 12 passages with 335 QA pairs from the English DROP data set were extracted for reference, i.e. its translated German and Chinese version shall be used in the zero-shot task.

Our main task is to accomplish zero-shot learning for cross-lingual discrete reasoning, i.e., we take the baseline model that was pre-trained and optimised on the English data, and then make evaluation on the German and Chinese passages and QA pairs to investigate how well the system transfers discrete reasoning knowledge across languages. As mentioned before, the corresponding 12 passages in German and Chinese version of the English test set utilized as reference was evaluated, since we are supposed to use paired dev set to make comparison of model's performance on different languages.

5.4 Results

The results of zero-shot experiments are demonstrated in Table 2. As revealed in the table, our Multilingual BERT model performs well on zero-shot learning regarding discrete reasoning over

	EM	F1
English (reference)	27.16	30.99
German	**11.94**	**17.34**
Chinese	**6.87**	**7.56**

Table 2: Performance of Multilingual BERT (evaluated on test set) on original English DROP and on Zero-shot Learning with German and Chinese

		English		German		Chinese	
Answer Type	%	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
Number	61.79	3.38	4.03	2.42	3.71	0	0
Span	31.04	**78.85**	**82.81**	**32.69**	**42.15**	**22.12**	**22.12**
Multiple Spans	6.57	0	33.36	0	25.36	0	10.59
Date	0.60	**100**	**100**	**50**	**50**	0	0

Table 3: Test set performance breakdown by different answer types

paragraph. While the reference performance is EM 27.16 and F1 30.99, the model yields EM 11.94 and F1 17.34 on German and EM 6.87 and F1 7.56 on Chinese data. We thus obtain a significantly better score with German samples, possibly because German and English have bigger lexical similarity and similarity in grammatical rules than Chinese and English, which is easier for the model to capture the information from German texts.

6 Analysis

6.1 Error analysis: English Test Samples

The model's performance on English test set shows that it works particularly well in extractive questions, in which it is typically asked to compare and list the order of different events, or find the date of certain event (e.g., Q: "Which player had the longest scoring play of the game?", A: "Donald Driver"), with an F1 score of 82.81, as opposed to numeric questions, which are answered almost randomly, with an F1 score of 4.02.

6.2 Error analysis: German Test Samples

The model's performance on German test set follows a similar pattern, with an F1 score of 42.15 on span questions and 3.71 on numeric questions. Moreover, it is slightly more frequent in German samples that the model gives a span as outputs when a number is practically asked than in English test samples, which indicates a lower level of understanding on queries (e.g., Q: "How many points did Denver score in the third quarter?", A: "play time").

6.3 Error analysis: Chinese Test Samples

For Chinese test samples, the model gets an F1 score of 22.12 for span-based questions, and 0 zero for numeric ones. In addition, the model systematically misunderstands the intent of numeric questions and in most cases outputs spans as its answer. One interesting finding in Chinese samples is that, although the model is not capable to pick the right answers in some cases, it correctly captured certain meaning of tokens (e.g., Q: "How many overnight stays were in either four or three star hotels?", A: "four star hotels", indicating a contextual misinterpretation, but a possibly meaningful understanding of the token 'or').

7 Conclusion

This work serves as our first attempt to conduct zero-shot learning of cross-lingual discrete reasoning, in which we show that 1) multilingual BERT model can be configured to solve discrete reasoning tasks, and 2) the knowledge of discrete reasoning can be transferred cross-lingually in German and Chinese languages to certain extent, even without any available parallel training data.

However, our study entails several significant limitations. To start with, we trained a multilingual BERT model, whose performance is considerably worse than the state-of-the-art number-aware models on the DROP leaderboard. Thus, in our future work, we plan to further explore a better optimized model that provides us with a more solid foundation of zero-shot transfer. Moreover, discrete reasoning is in its nature a set of various tasks, for which an evaluation metric for different type of questions is required. Thus, we plan to consider the task as a multi-task setup and design an adequate loss function that covers various type errors in our later work. In addition, we did not utilize our collected data in Chinese examples, which allowed us to make a comparison between 1) our model and a model fine-tuned on the Chinese samples (i.e. few-shot learning), and 2) ours and a model trained from scratch on Chinese samples. In our future work, we plan to further collect data in German language as well to make such comparisons. We also plan to make our code and data publicly available for further research.

References

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

[2] Pranav Rajpurkar, Jia Robin, and Liang Percy. Know what you don't know: Unanswerable questions for squad. In *arXiv preprint arXiv:1806.03822*, 2018.

[3] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge, 2018.

[4] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL*, 2019.

[5] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830, 2016.

[6] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[7] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016.

[8] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[9] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[10] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603, 2016.

[11] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. Fastqa: A simple and efficient neural architecture for question answering. *CoRR*, abs/1703.04816, 2017.

[12] Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *arXiv preprint arXiv:1909.00437*, 2019.

[13] Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. Is multilingual bert fluent in language generation? *arXiv preprint arXiv:1910.03806*, 2019.

[14] Momchil Hardalov, Ivan Koychev, and Preslav Nakov. Beyond english-only reading comprehension: Experiments in zero-shot multilingual transfer for bulgarian. *arXiv preprint arXiv:1908.01519*, 2019.

A Appendix (optional)

Fine-tuning Multilingual BERT on Chinese Samples

Apart from directly implementing zero-shot transfer based on the baseline model, we also tried to fine-tune our Multilingual BERT by training with new Chinese passages and their question-answer pairs in the next step, in order to explore a better performance of the model at zero-shot learning. The logic is reflected as training the Chinese data set and evaluating both on the English and Chinese test set, as designed in the previous part. We divided the 12 passages into 6 dev and 6 test sets, while the rest 30 passages were utilized as training data for fine-tuning.

Table 4 shows the performance of fine-tuned Multilingual BERT with Chinese training data. After evaluating the fine-tuned model on English and Chinese test sets, EM and F1 scores of both languages are relative low compared to the main zero-shot experiment, indicating that our fine-tuning process does not have positive effect on the zero-shot learning of Multilingual BERT.

	EM	**F1**
English (reference)	27.16	30.89
Chinese	**6.57**	**7.26**

Table 4: Performance of Fine-tuned Multilingual BERT (evaluated on test set) with Chinese training data