Computer Speech and Language 80 (2023) 101486

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

PQuAD: A Persian question answering dataset

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, Saeedeh Momtazi *

Computer Engineering Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran

ARTICLE INFO

Dataset link: https://github.com/AUT-NLP/PQuAD

Keywords:
Machine reading comprehension
Natural language processing
Persian dataset
Question answering

ABSTRACT

We present the Persian Question Answering Dataset (PQuAD), a crowdsourced reading comprehension dataset on Persian Wikipedia articles. It includes 80,000 questions along with their answers, with 25% of the questions being adversarially unanswerable. We examine various properties of the dataset to show the diversity and the level of its difficulty as a MRC benchmark. By releasing this dataset, we aim to ease research on Persian reading comprehension and the development of Persian question answering systems. Our experiments on different state-of-the-art pre-trained contextualized language models show 74.8% Exact Match (EM) and 87.6% F1-score that can be used as the baseline results for further research on Persian QA.

1. Introduction

Machine Reading Comprehension (MRC) is one of the main question answering tasks in natural language understanding which requires a system to read a passage and then answer the given questions from the passage. Developing a machine that has the comprehension ability helps Artificial Intelligence (AI) to reach the goal of competing with humans. Developing such a machine that can comprehend the available data in the text format would be highly beneficial, because in many jobs human workers are required to get information from text and having a high performance MRC model will help to reduce manual tasks significantly.

Besides various question answering systems over text and knowledge-based (Abbasiantaeb and Momtazi, 2021; Momtazi and Abbasiantaeb, 2022), MRC has undergone significant advancements in recent years due to the availability of many large-scale annotated datasets, including MCTest (Richardson et al., 2013), BookTest (Bajgar et al., 2016), SQuAD (Rajpurkar et al., 2016), SearchQA (Dunn et al., 2017), NewsQA (Trischler et al., 2017), ReCoRD (Zhang et al., 2018) and ReCO (Wang et al., 2020). These datasets provide the opportunity to train data-intensive deep learning models and achieve performance comparable to humans.

The main MRC datasets, however, are particularly in English, while low-resource languages such as Persian require further efforts in this regard. Recently, some progress has been made towards this goal, namely PersianQA (Ayoubi, 2021) and ParSQuAD (Abadani et al., 2021), which are gathered automatically by different methodologies. These datasets, however, either lack the sufficient number of examples to be used in a supervised learning framework for today's large models or are not as high quality as manual ones. ParSQuAD (Abadani et al., 2021) is collected by translating the SQuAD dataset (Rajpurkar et al., 2016) into Persian by google-translate and PersianQA (Ayoubi, 2021) is a small size MRC dataset consisting of 10,000 questions, which is not suitable for training the large models.

In this paper, we introduce PQuAD,[1] a large-scale and high quality Persian span-extraction MRC dataset consisting of 80,000 human-annotated questions. PQuAD's questions are based on Persian Wikipedia articles and cover a wide variety of subjects. The dataset has been used for training the state-of-the-art models in the field. The main contributions of this paper are as follows: (1) containing a high volume of questions compared to other corpora in low-resource languages, (2) collecting and labeling data

* Corresponding author.
E-mail address: momtazi@aut.ac.ir (S. Momtazi).

[1] https://github.com/AUT-NLP/PQuAD

https://doi.org/10.1016/j.csl.2023.101486
Received 7 February 2022; Received in revised form 7 September 2022; Accepted 15 January 2023
Available online 25 January 2023
0885-2308/© 2023 Elsevier Ltd. All rights reserved.

manually rather than translating existing corpora from other languages, (3) considering 25% of questions as unanswerable to improve the reading comprehension ability of QA systems, (4) specifying multiple answers to questions in the validation and test sets for more accurate evaluation.

The rest of the paper is organized as follows: in Section 2, we briefly overview related MRC datasets. Section 3 describes the data collection process of PQuAD and methodologies used to ensure a high quality dataset. In Section 4, we analyze various properties of the dataset to show the level of its difficulty as a MRC benchmark. Section 5 is devoted to introduction of baseline models and their results on PQuAD. Finally, Section 6 concludes the paper.

2. Related work

The importance of providing MRC datasets has increased as a result of both industries' and academics' growing interest in MRC. Although a large number of MRC datasets are available (Richardson et al., 2013; Bajgar et al., 2016; Rajpurkar et al., 2016; Dunn et al., 2017; Trischler et al., 2017; Zhang et al., 2018; Wang et al., 2020), new MRC datasets are still welcome, specially for low-resource languages.

Recently, there has been a great motivation towards collecting the MRC datasets for low-resource languages (Mozannar et al., 2019; Carrino et al., 2020; d'Hoffschmidt et al., 2020; Lim et al., 2019). The availability of the multilingual pre-trained language models, such as Multilingual BERT (Devlin et al., 2019), and XLM-RoBERTa (Conneau et al., 2020), can be considered as one of the main reasons behind this motivation. Most of these datasets, are collected from Wikipedia pages similar to SQuAD (Mozannar et al., 2019; d'Hoffschmidt et al., 2020; Lim et al., 2019) or are curated by translating the SQuAD dataset (Carrino et al., 2020).

MCTest (Richardson et al., 2013) is one of the earliest MRC datasets which includes fictional stories and multiple choice questions collected for each story. Having only 2000 questions, its size is smaller than the datasets with which large models are often trained.

SQuAD 1.1 (Rajpurkar et al., 2016) is a large-scale MRC dataset which is collected from Wikipedia pages. The answer of each question is represented as a span of text within the corresponding passage. In SQuAD 2.0 (Rajpurkar et al., 2018), more than 50,000 unanswerable questions are added. The unanswerable questions are very similar to the answerable questions and this feature makes the SQuAD 2.0 dataset a more challenging dataset. Nevertheless, the available studies on MRC using SQuAD 2.0 (Yamada et al., 2020; Yang et al., 2019; Joshi et al., 2020) have achieved a great performance over human evaluation. Inspired by SQuAD 2.0, PersianQA (Ayoubi, 2021) has been created with the help of crowdworkers and using Wikipedia articles. It includes about 10,000 questions, some of which are written in informal language.

NewsQA (Trischler et al., 2017) is a large-scale dataset which includes 119,633 questions gathered by human workers from 12,744 CNN's news articles. Among the available studies on the NewsQA dataset (Joshi et al., 2020; Tay et al., 2018; Kundu and Ng, 2018), the SpanBERT model (Joshi et al., 2020) has achieved a higher performance compared to the human judgment.

SearchQA (Dunn et al., 2017) is collected by retrieving the correct answer in a real question answering system. Different from other datasets in which a question is posed given the passage, the question is posed and then the relevant passages are retrieved for the given question. SearchQA includes more than 140000 question–answer pairs and each question–answer pair is linked to about 50 text snippets.

ReCoRD (Zhang et al., 2018) is collected automatically from news articles. The distinctive feature of this dataset is that despite the other MRC datasets, such as SQuAD and NewsQA, it needs commonsense reasoning over several sentences for comprehending the answer. A small portion of the questions of this dataset can be answered by paraphrasing.

There are several MRC datasets for other languages than English, including ParSQuAD (Abadani et al., 2021) in Persian, ReCo (Wang et al., 2020) in Japanese, FQuAD (d'Hoffschmidt et al., 2020) in French, ARCD (Mozannar et al., 2019) in Arabic, KorQuAD1.0 (Lim et al., 2019) in Korea, and Spanish translation of SQuAD (Carrino et al., 2020).

ParSQuAD (Abadani et al., 2021) in Persian is gathered by translating the SQuAD into Persian. ARCD (Mozannar et al., 2019) consists of Arabic translation of SQuAD and questions posed by crowd workers. For Spanish SQuAD (Carrino et al., 2020), a trained neural machine translation and a trained unsupervised word alignment model are developed for automatically translating the SQuAD dataset to Spanish. FQuAD (d'Hoffschmidt et al., 2020) is collected from French Wikipedia pages. ReCO (Wang et al., 2020) is a very large opinion-based MRC dataset including 300,000 Japanese questions which includes both factoid and non-factoid questions.

Another type of the datasets is cloze datasets. In the cloze datasets, a word is omitted from the text and the goal is to detect the omitted word from the given text. Children's Book Test (CBT) (Hill et al., 2015) is a cloze dataset. Each sample in this dataset is 21 consecutive sentences and one word in the last sentence is omitted. The first 20 sentences are given as context and the missing word in the next sentence must be predicted. The BookTest dataset (Bajgar et al., 2016) is very similar to the CBT dataset but it is designed for training large models due to its 60 times larger size than CBT.

3. Dataset collection

Following the proposed structure in SQuAD (Rajpurkar et al., 2016), the data collection process of PQuAD is performed in three steps: passage curation, question–answer pair annotation, and additional answer collection.

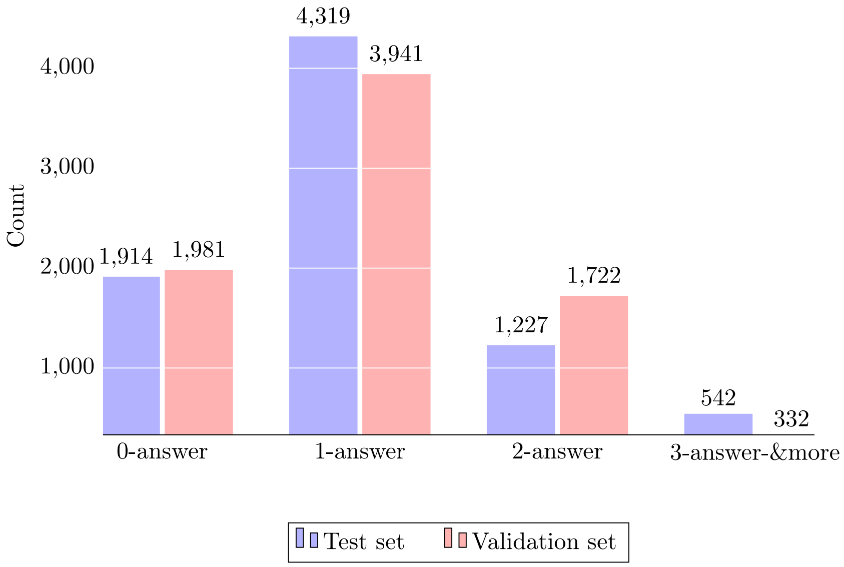


Fig. 1. Distribution of number of answers available for each question in test and validation sets.

3.1. Passage curation

The most important articles in the Persian Wikipedia are selected based on two criteria: (1) having an info-box, and (2) being among the top pages in the Persian Wikipedia based on the PageRank algorithm.

We believe that important Wikipedia pages, that have enough content for question collection, usually contain an info-box and we use it as the primary condition to retrieve 600,000 articles.

In the next step, we build a graph for these articles based on their links and rank the collected articles using the NetworkX library's[2] PageRank method with damping parameter set to 0.4. Starting from the highest rank, we choose the introduction section and the first 20 paragraphs of each article, whose length is 500–1100 characters. The selection of articles and paragraphs continued until 80,000 questions were obtained. 11,000 paragraphs which belong to 1,125 articles spanning a wide range of topics were selected in this process. These articles are then randomly divided into three subsets: train, validation, and test sets.

3.2. Question–answer annotation

Crowdworkers were instructed to spend 15 min on each paragraph and collect about five to ten questions as well as their answers, if they are answerable using the paragraph's content. At the same time, they discarded the paragraphs that were hard to understand or contained too many non-Persian words. To further regulate the process, we developed a toolkit for the annotation process. In this toolkit, crowdworkers typed each question in a text box and then either labeled the question as unanswerable or specified an answer by marking a span. For answerable questions, the crowdworkers selected the shortest text span that provides the answer to the question. The answers are not limited to full sentences; they could contain only one word or any other number of words.

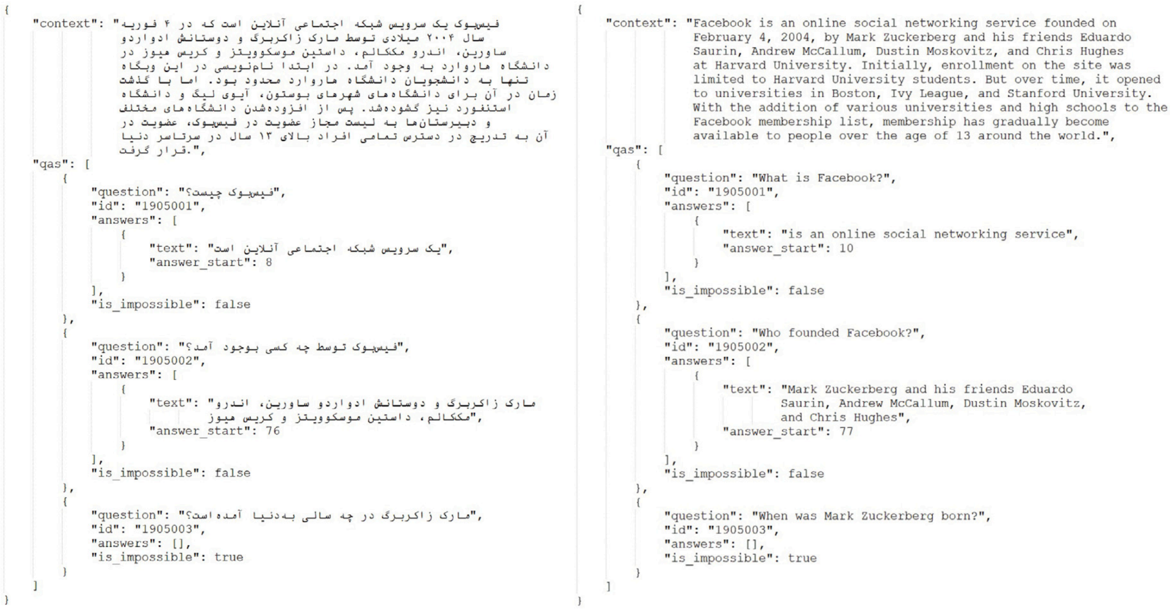
Workers were encouraged to use paraphrased sentences in their questions to address vocabulary gap (Momtazi and Kalkow, 2015) when developing a system over this dataset. They also asked to avoid choosing the answers comprising non-Persian words. Considering the challenges of Persian scripts in corpus development and text processing (Ghayoomi et al., 2010; Ghayoomi and Momtazi, 2009), workers are asked to follow the Persian standard dictation.

Unanswerable questions roughly form 25% of all questions and were collected in such a way that are in the same content of the paragraph, but they could not be answered just by reading the paragraph.

3.3. Additional answer collection

To have a better evaluation of models, we further check the question–answer pairs in the test and validation sets while additional answers are being collected. A new group of crowdworkers answer the questions in these two sets. The crowdworkers were encouraged to include any other correct answer spans with minor differences to the original answers. Moreover, if they found that the original answer is incorrect, they resolved the incorrect answer either by choosing the more appropriate answer or removing the corresponding question from the passage. The distribution of the number of answers per question is represented in Fig. 1.

[2] https://networkx.org/



```
{
    "context": "فیسبوک یک سرویس شبکه اجتماعی آنلاین است که در ۴ فوریه
                سال ۲۰۰۴ میلادی توسط مارک زاکربرگ و دوستانش ادواردو
                ساورین، اندرو مککالم، داستین موسکوویتز و کریس هیوز در
                دانشگاه هاروارد به وجود آمد. در ابتدا نام‌نویسی در این وبگاه
                تنها به دانشجویان دانشگاه هاروارد محدود بود. اما با گذشت
                زمان در آن برای دانشگاه‌های شهرهای بوستون، آیوی لیگ و دانشگاه
                استنفورد نیز گشوده‌شد. پس از افزوده‌شدن دانشگاه‌های مختلف
                و دبیرستان‌ها به لیست مجاز عضویت در فیسبوک، عضویت در
                آن به تدریج در دسترس تمامی افراد بالای ۱۳ سال در سرتاسر دنیا
                قرار گرفت.",
    "qas": [
        {
            "question": "فیسبوک چیست؟",
            "id": "1905001",
            "answers": [
                {
                    "text": "یک سرویس شبکه اجتماعی آنلاین است",
                    "answer_start": 8
                }
            ],
            "is_impossible": false
        },
        {
            "question": "فیسبوک توسط چه کسی بوجود آمد؟",
            "id": "1905002",
            "answers": [
                {
                    "text": "مارک زاکربرگ و دوستانش ادواردو ساورین، اندرو
                            مککالم، داستین موسکوویتز و کریس هیوز",
                    "answer_start": 76
                }
            ],
            "is_impossible": false
        },
        {
            "question": "مارک زاکربرگ در چه سالی به‌دنیا آمده‌است؟",
            "id": "1905003",
            "answers": [],
            "is_impossible": true
        }
    ]
}
```

```
{
    "context": "Facebook is an online social networking service founded on
                February 4, 2004, by Mark Zuckerberg and his friends Eduardo
                Saurin, Andrew McCallum, Dustin Moskovitz, and Chris Hughes
                at Harvard University. Initially, enrollment on the site was
                limited to Harvard University students. But over time, it opened
                to universities in Boston, Ivy League, and Stanford University.
                With the addition of various universities and high schools to the
                Facebook membership list, membership has gradually become
                available to people over the age of 13 around the world.",
    "qas": [
        {
            "question": "What is Facebook?",
            "id": "1905001",
            "answers": [
                {
                    "text": "is an online social networking service",
                    "answer_start": 10
                }
            ],
            "is_impossible": false
        },
        {
            "question": "Who founded Facebook?",
            "id": "1905002",
            "answers": [
                {
                    "text": "Mark Zuckerberg and his friends Eduardo
                            Saurin, Andrew McCallum, Dustin Moskovitz,
                            and Chris Hughes",
                    "answer_start": 77
                }
            ],
            "is_impossible": false
        },
        {
            "question": "When was Mark Zuckerberg born?",
            "id": "1905003",
            "answers": [],
            "is_impossible": true
        }
    ]
}
```

Fig. 2. An example of a passage with corresponding questions. The left part is the original Persian data, and the right part is the English translation. The third question is an example of an unanswerable question. Answers are represented by the text of answer and the index of the start token of the answer.

Table 1
Statistics of PQuAD dataset.

	Train	Validation	Test	Total
Total questions	63994	7976	8002	79972
Unanswerable questions	15721	1981	1914	19616
Mean # of paragraph tokens	125	121	124	125
Mean # of question tokens	10	11	11	10
Mean # of answer tokens	5	6	5	5

4. Dataset analysis

We investigate PQuAD in order to demonstrate the level of its challenge and its various properties. Types of answers and some statistics about the different domains available in the dataset, as well as general information, are included in this analysis.

4.1. Dataset structure

PQuAD is stored in the JSON format and consists of passages where each passage is linked to a set of questions. The unanswerable questions are marked as unanswerable and the answer of the answerable questions is specified with answer's span. An example of a passage with corresponding question and answers is shown in Fig. 2. This example includes three different questions. The first two questions have answers, while the last question (When was Mark Zuckerberg born?) has no answer. The content of the unanswerable question is similar to the content of the paragraph, but the answer is not available in this paragraph.

PQuAD includes about 11,000 passages and 80,000 questions where each passage includes about 7 questions on average. Final set of articles in PQuAD are randomly divided to form three sets: train, validation, and test sets. Some detailed statistics of PQuAD are presented in Table 1.

As mentioned before, PQuAD features a diverse range of topics. Table 2 categorizes topics based on the domains associated with the articles in the whole dataset.

4.2. Answer types

We categorize answer spans of answerable questions of the dataset based on their Part of Speech (POS) and Named Entity (NE) tags, similarly to Rajpurkar et al. (2016). For POS tagging and NE recognition, we trained two BERT-based models using available Persian datasets provided by Bijankhan (2004) and Momtazi and Torabi (2020), respectively. The results of categorization are shown in Table 3. Considering the average of all subsets, the most popular types of the answers are numerical, proper noun phrase, and common noun phrase accounting for 24.4%, 40.7%, and 24.3% of all answers, respectively. The remaining 10.5% is distributed across adjective phrases, verb phrases, and other types.

Table 2
Diversity of article topics present in PQuAD and sample titles of source Wikipedia pages.

Domain	Example	Percentage(%)
Person	Pablo Picasso	22.3
Geographical locations	Caspian Sea	20.1
Science & Tech	Database	6.5
Organization	United Nations	6.0
Sports	Olympic Games	5.8
Fields of specialty	Psychology	5.0
Plants & Animals	Starfish	4.0
Art	Pop music	3.3
Historical Eras	Precambrian	3.2
Books & Movies	Star Wars	2.9
Religious	God	2.7
Events	Nowruz	2.6
Groups	Vikings	2.5
Languages	Middle Persian	2.4
Chemistry & Biology	Hydrogen	2.4
Astronomy	Solar System	2.2
Diseases & Medicines	COVID-19 pandemic	2.0
Objects	Carpet	1.6
Others	Immigration	2.4

Table 3
Categorization of answers based on POS and NE tags.

POS	NE	Percentage(%)		
		Train	Val	Test
Numeric	Date	10.1	10.2	9.9
	Other numeric	14.3	13.4	14.0
Proper noun phrase	Person (Individual)	15.3	14.9	17.0
	Location	15.1	12.4	13.6
	Person (Group)	2.7	2.8	3.0
	Organization	2.4	2.3	3.0
	Field	1.1	1.1	0.8
	Language	1.2	1.3	1.1
	Other entity	3.4	3.7	3.3
Common noun phrase	–	23.8	27.0	24.2
Adjective phrase	–	2.0	1.9	1.3
Verb phrase	–	4.5	4.9	4.9
Other	–	4.0	4.1	3.8

5. Experiments

5.1. Human performance

To estimate the human performance on our PQuAD dataset, we asked a new group of crowdworkers to answer 1000 questions in the test set. These questions are from 15 randomly selected articles and all questions of each article are tasked to one of the workers. The percentage of the unanswerable questions in the samples is the same as in the dataset. Workers were informed about the existence of unanswerable questions and were asked to either mark the questions for which they found no answer as unanswerable questions or specify the answer span in the paragraph. Performance measures used in this experiment are F1 and Exact Match (EM) metrics, the same as measures used by Rajpurkar et al. (2016). The estimated human performance on the test set is 88.3% for F1 and 80.3% for EM.

5.2. Models

We have evaluated PQuAD using two pre-trained transformer-based language models, namely ParsBERT (Farahani et al., 2021) and XLM-RoBERTa (Conneau et al., 2020), as well as BiDAF (Levy et al., 2017) which is an attention-based model proposed for MRC.

- ParsBERT: ParsBERT is a transform-based language model which uses the architecture of the BERT$_{base}$ model. ParsBERT is specially developed for the Persian language and is trained on Persian texts gathered from diverse sources.
- XLM-RoBERTa: XLM-RoBERTa is a transformer-based language model. Similar to the monolingual RoBERTa language model, XLM-RoBERTa does not use the Next Sentence Prediction (NSP) task for training and it is only trained using the multilingual

Table 4
Scores of baseline models and human performance. HasAns and NoAns columns show the scores over positive and negative questions respectively.

Model	EM	F1	HasAns_EM	HasAns_F1	NoAns_EM/F1
BNA	54.4	71.4	43.9	66.4	87.6
ParsBERT	68.1	82.0	61.5	79.8	89.0
XLM-RoBERTa	74.8	87.6	69.1	86.0	92.7
Human	80.3	88.3	74.9	85.6	96.8

Table 5
XLM-RoBERTa and human performance on different answer categories.

Answer Type	XLM-RoBERTa		Human	
	EM	F1	EM	F1
Numeric	83.51	92.64	86.9	92.7
Proper noun phrase	78.4	88.7	78.6	86.2
Common noun phrase	70.5	88.0	71.7	84.9
Adjective phrase	57.1	79.5	85.7	85.7
Verb phrase	53.3	83.4	62.0	84.7
Other	62.3	76.5	68.7	76.4

Masked Language Model (MLM). Size of the training data is very large and it is capable of training by 100 different languages. XLM-RoBERTa is a strong multilingual pre-trained language model, which has outperformed the Multilingual BERT (mBERT) on several cross-lingual tasks.

- BiDAF: BiDAF is a neural architecture which utilizes character-level, word-level, and contextualized embedding for text representation. Attention mechanism is used in the model for creating a query-aware representation of the input context. The model detects the span of answer from context by calculating the probability of being the begin and end token of the span for each word within the context.

5.3. Results

Table 4 shows the performance of baseline models on the test set. According to our experiments, monolingual ParsBERT underperforms multilingual XLM-RoBERTa which might be the result of a less effective pre-training process. Despite the close F1-scores of XLM-RoBERTa and humans, there is a 5.5% gap between their EM scores. Based on the detailed scores of answerable and unanswerable questions, this lower performance is mostly due to the model's inability to select the shortest span containing the answer for positive questions. Using the same model, and to have a better understanding of the model's weak points, we perform an analysis on the error rates based on answer types. As reported in Table 5, the model has comparatively high performance on noun phrases and numbers. However, differences between model and human performance are more obvious in adjective and verb phrases. These types of answer spans are more variable in their structures, especially in the Persian language, and specifying the exact start and end position of them is a challenging task for the model, suggesting a direction for future works.

6. Conclusion

In this paper, we present PQuAD, a Persian reading comprehension dataset consisting of 80,000 questions. This dataset is based on Wikipedia articles and its question–answer pairs are annotated by humans. Since unanswerable questions are also included in PQuAD, models are required to abstain from answering whenever no proper answer span is present in the given context. By releasing this dataset, we aim to ease research on Persian reading comprehension and the development of Persian question answering systems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The provided dataset is available via the following link: https://github.com/AUT-NLP/PQuAD.

Acknowledgments

PQuAD is developed with collaboration of Mabna Intelligent Computing at Amirkabir Science and Technology Park and is supported by the Vice Presidency for Scientific and Technology.

References

Abadani, N., Mozafari, J., Fatemi, A., Nematbakhsh, M.A., Kazemi, A., 2021. ParSQuAD: Machine translated squad dataset for Persian question answering. In: 2021 7th International Conference on Web Research. ICWR, IEEE, pp. 163–168.

Abbasiantaeb, Z., Momtazi, S., 2021. Text-based question answering from information retrieval and deep neural network perspectives: A survey. WIREs Data Min. Knowl. Discov..

Ayoubi, M.Y., 2021. PersianQA: a dataset for Persian question answering. https://github.com/SajjjadAyobi/PersianQA,

Bajgar, O., Kadlec, R., Kleindienst, J., 2016. Embracing data abundance: Booktest dataset for reading comprehension. arXiv preprint arXiv:1610.00956.

Bijankhan, M., 2004. The role of corpora in writing grammar. J. Linguist. 19 (2), 48–67.

Carrino, C.P., Costa-jussà, M.R., Fonollosa, J.A.R., 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 5515–5523, URL https://aclanthology.org/2020.lrec-1.677.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 8440–8451. http://dx.doi.org/10.18653/v1/2020.acl-main.747, URL https://aclanthology.org/2020.acl-main.747.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

d'Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T., Vidal, M., 2020. FQuAD: French question answering dataset. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp. 1193–1208. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.107, URL https://aclanthology.org/2020.findings-emnlp.107.

Dunn, M., Sagun, L., Higgins, M., Guney, V.U., Cirik, V., Cho, K., 2017. Searchqa: A new q&a dataset augmented with context from a search engine. arXiv preprint arXiv:1704.05179.

Farahani, M., Gharachorloo, M., Farahani, M., Manthouri, M., 2021. Parsbert: Transformer-based model for persian language understanding. Neural Process. Lett. 53 (6), 3831–3847.

Ghayoomi, M., Momtazi, S., 2009. Challenges in developing Persian corpora from online resources. In: Proceedings of 2009 IEEE International Conference on Asian Language Processing. pp. 108–113.

Ghayoomi, M., Momtazi, S., Bijankhan, M., 2010. A study of corpus development for Persian. Int. J. Asian Lang. Process. 20 (1), 17–33.

Hill, F., Bordes, A., Chopra, S., Weston, J., 2015. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301.

Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O., 2020. Spanbert: Improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. 8, 64–77.

Kundu, S., Ng, H.T., 2018. A question-focused multi-factor attention network for question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Levy, O., Seo, M., Choi, E., Zettlemoyer, L., 2017. Zero-shot relation extraction via reading comprehension published in CoNLL 2017, Association for Computational Linguistics. arXiv preprint arXiv:1706.04115.

Lim, S., Kim, M., Lee, J., 2019. KorQuAD1. 0: Korean QA dataset for machine reading comprehension. arXiv preprint arXiv:1909.07005.

Momtazi, S., Abbasiantaeb, Z., 2022. Question Answering over Text and Knowledge Base. Springer.

Momtazi, S., Kalkow, D., 2015. Bridging the Vocabulary Gap between Questions and Answer Sentences. Inf. Process. Manage. 51.

Momtazi, S., Torabi, F., 2020. Named entity recognition in Persian text using deep learning. J. Signal and Data Process..

Mozannar, H., Hajal, K.E., Maamary, E., Hajj, H., 2019. Neural arabic question answering Published in Proceedings of the Fourth Arabic Natural Language Processing Workshop. pp. 108–118.

Rajpurkar, P., Jia, R., Liang, P., 2018. Know what you don't know: Unanswerable questions for squad. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Melbourne, Australia, pp. 784–789. http://dx.doi.org/10.18653/v1/P18-2124, URL https://aclanthology.org/P18-2124.

Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P., 2016. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp. 2383–2392. http://dx.doi.org/10.18653/v1/D16-1264, URL https://aclanthology.org/D16-1264.

Richardson, M., Burges, C.J., Renshaw, E., 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Seattle, Washington, USA, pp. 193–203, URL https://aclanthology.org/D13-1020.

Tay, Y., Tuan, L.A., Hui, S.C., Su, J., 2018. Densely connected attention propagation for reading comprehension. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 4911–4922.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K., 2017. NewsQA: A machine comprehension dataset. In: Proceedings of the 2nd Workshop on Representation Learning for NLP. Association for Computational Linguistics, Vancouver, Canada, pp. 191–200. http://dx.doi.org/10.18653/v1/W17-2623, URL https://aclanthology.org/W17-2623.

Wang, B., Yao, T., Zhang, Q., Xu, J., Wang, X., 2020. Reco: A large scale Chinese reading comprehension dataset on opinion. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9146–9153.

Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y., 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP, Association for Computational Linguistics, Online, pp. 6442–6454. http://dx.doi.org/10.18653/v1/2020.emnlp-main.523, URL https://aclanthology.org/2020.emnlp-main.523.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 32.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., Van Durme, B., 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885.