Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6083–6092
Marseille, 20-25 June 2022

PerCQA: Persian Community Question Answering Dataset

Naghme Jamali, Yadollah Yaghoobzadeh, Heshaam Faili
School of Computer Science, Institute for Research in Fundamental Sciences (IPM)
School of Electrical and Computer Engineering, College of Engineering, University of Tehran
School of Electrical and Computer Engineering, College of Engineering, University of Tehran
Tehran, Iran
naghme.jamali@ipm.ir, y.yaghoobzadeh@ut.ac.ir, hfaili@ut.ac.ir

Abstract

Community Question Answering (CQA) forums provide answers to many real-life questions. These forums are trendy among machine learning researchers due to their large size. Automatic answer selection, answer ranking, question retrieval, expert finding, and fact-checking are example learning tasks performed using CQA data. This paper presents **PerCQA**, the first Persian dataset for CQA. This dataset contains the questions and answers crawled from the most well-known Persian forum. After data acquisition, we provide rigorous annotation guidelines in an iterative process and then the annotation of question-answer pairs in SemEvalCQA format. PerCQA contains 989 questions and 21,915 annotated answers. We make PerCQA publicly available to encourage more research in Persian CQA. We also build strong benchmarks for the task of answer selection in PerCQA by using mono- and multi-lingual pre-trained language models.

Keywords: community question answering, dataset, answer selection, Persian Language

1. Introduction

A Community Question Answering (CQA) platform allows web users to get answers to their questions from domain experts. With the advent of websites such as Yahoo! Answers[1], Stack Exchange[2], and Quora[3], CQA has attracted a lot of attention. CQA is considered a reliable source for acquiring the required knowledge to solve problems that cannot be solved directly by searching on web pages. Due to the significant expansion of these forums in terms of topics and the number of users in different languages, the administrators' manual checking, and verification of contents are very challenging. Therefore, automatic solutions are required primarily to direct users to the most appropriate answers for each question.

The ability to automatically find relevant questions to reuse their existing answers (i.e., question retrieval) and searching for relevant answers among many responses for a given question (i.e., answer selection) are famous tasks in CQA. Most current work in this area is conducted using publicly available datasets, such as SemEvalCQA (Nakov et al., 2015; Nakov et al., 2016a; Nakov et al., 2017), TREC QA (Wang et al., 2007), Wiki QA (Yang et al., 2015), and InsuranceQA (Feng et al., 2015). Here, we focus on datasets in the Persian language. One significant limitation for developing automatic systems to manage the Persian CQA forums is the lack of such datasets in the Persian language. This affects the quality and usability of these forums. By researching these forums, their quality can be improved because the users are able to get the answers they desire more efficiently.

In this study, we build a dataset for Persian CQA called *PerCQA*. It consists of questions and answers posted by the most popular Persian forum users, named Ninisite[4]. Our primary focus is on data gathering, preparation, and annotation. A web scraping tool is developed for extracting useful information from Ninisite. The most crucial step in preparing a high-quality dataset is the development of a detailed and consistent annotation guideline. Our annotation guideline affects the agreement between annotators significantly. We also build an annotation tool to reduce the labeling time.

Data annotation in PerCQA consists of two stages: question annotation for selecting appropriate questions and answer annotation with three labels ("Good", "Bad", "Potential"). As a result, *PerCQA* is built with 989 questions and 21,915 answers. This dataset is structurally similar to SemEval 2015 English CQA dataset (Nakov et al., 2015), which has 3,229 questions and 20,162 answers.

Furthermore, to evaluate PerCQA and analyze the answer selection task, several extensive experiments are performed with non-contextualized, namely *Word2vec* (Mikolov et al., 2013), *fastText* (Bojanowski et al., 2017) and contextualized word embeddings learned by pre-trained language models (PLMs), namely *BERT* (Devlin et al., 2019), *ParsBERT* (Farahani et al., 2021). We further improve our results by transferring knowledge from English datasets using multilingual PLMs such as *mBERT* (Devlin et al., 2019) and XLM-RoBERTa (*XLM-R*) (Conneau et al., 2020).

We use *PV-Cnt* (Yang et al., 2015), *BiLSTM-attention* (Tan et al., 2015), *RCNN* (Zhou et al., 2018), and *CETE* (Laskar et al., 2020) as our baseline systems that employ word embeddings. Our experimental results demonstrate that ParsBERT and XLM-R embeddings using by CETE system outperform all other base-

[1] https://answers.yahoo.com/
[2] https://stackexchange.com/.
[3] https://www.quora.com/.
[4] www.ninisite.com

lines. We find that XLM-R overtakes ParsBERT by up to about +3% macro F1-score on PerCQA by transferring knowledge from SemEvalCQA English datasets. Our main contributions can be summarized as the followings:

- We build and release PerCQA[5], the first Persian dataset for CQA, to enhance the research and applications of CQA tasks in Persian.
- We apply several state-of-the-art methods to our dataset and set strong baselines for the task of answer selection in PerCQA.

Some previous research and existing CQA datasets are reviewed in the next section. Section 3 describes the process of creating our dataset in detail. The structure of PerCQA and its statistics is presented in Section 4. In Section 5 experimental results and analysis are presented. Finally, Section 6 is the concluding remarks.

2. Related Work

In this section, we look at existing CQA datasets that are widely used for evaluating the CQA tasks and introduce some of the tasks in these forums and describe some various models in answer selection task.

Dataset	Set	#Questions	#Answers
SemEval 2015	Train	2600	16541
	Dev	300	1654
	Test	329	1976
SemEval 2016	Train	4879	36198
	Dev	244	2440
	Test	327	3270
SemEval 2017	Train	4879	36198
	Dev	244	2440
	Test	293	2930
Insurance QA	Train	12889	21325
	Dev	2000	3354
	Test	2000	3308
Yahoo CQA	Train	50112	253440
	Dev	6289	31680
	Test	6283	31680
WikiQA (RAW)	Train	2118	20360
	Dev	296	2733
	Test	633	6165
TREC-QA (RAW)	Train	1229	53417
	Dev	82	1148
	Test	100	1517

Table 1: Statistics of various CQA datasets.

2.1. CQA Datasets

There are many CQA datasets released in different languages to date. The following datasets are available in English: SemEval (Nakov et al., 2015) (Nakov et al., 2017), TREC QA (Wang et al., 2007), WikiQA (Yang et al., 2015), Insurance QA (Feng et al., 2015), and Yahoo! Answers (Qiu and Huang, 2015). There are three versions of the SemEval dataset (2015, 2016, and 2017) crawled from the Qatar Living forum, and each question has attributes such as question category, question type, and question date. The TREC-QA dataset, provided by TREC-QA track 8-13, is the most widely used benchmark for testing QA and CQA models. Another popular dataset for evaluating answer selection systems is WikiQA. This dataset is collected from Bing query logs. The Insurance QA dataset is a non-factoid QA dataset from the insurance domain. The Yahoo! Answers dataset was generated by (Qiu and Huang, 2015), using the Computer and Internet category is resolved questions in Yahoo! Answers.
Other datasets from CQA are also released in other languages like *Arabic* or *Chinese* in addition to English. For Arabic, "SemEval-2015 CQA-subtask A" provided a dataset that was crawled from the *Fatwa*[6] website for its answer selection shared task (Nakov et al., 2015) and also in (Nakov et al., 2017) "SemEval-2017 CQA-subtask D" only was used one of the existing websites,namely Altibbi[7]. JEC-QA (Zhong et al., 2020) is a Chinese question answering dataset of Chinese law forums for legal advice that contains a large amount of legal knowledge. It was collected from the National Judicial Examination of China and is available from[8]. Table 1 presents the statistics of some datasets in CQA, and according to the research we have done, there is no Persian dataset in CQA.

2.2. CQA Tasks

Following the creation of datasets in different languages, different types of research have been conducted on CQA platforms. The "Question similarity" task in CQA forums is to retrieve a collection of questions similar to the question that the user has asked in advance. In SemEval-2016/2017 (task3-Subtask B) (Nakov et al., 2017) (Nakov et al., 2016a), the "Question-Question Similarity" task has been included as a benchmark task. (Kunneman et al., 2019) demonstrated adjusting preprocessing and word similarity settings improved the result of identifying duplicate questions. The goal of "Question Retrieval" in CQA is to find existing and semantically equivalent questions. Answers to the queried questions are derived from the best answers to these similar questions. Numerous studies have been conducted in this field (Zhou et al., 2015), (Othman et al., 2017), (Othman et al., 2019), (Wang et al., 2020).
It is vital and valuable to find users who have the expertise to answer your questions. Therefore, "Expert Finding" is one of the essential tasks in CQA. To this aim, (Ghasemi et al., 2021)'s model, for extract-

[5] https://github.com/PerCQA

[6] http://fatwa.islamweb.net/

[7] http://www.altibbi.com

[8] https://jecqa.thunlp.org/

ing users' embeddings, applied node2vec (Grover and Leskovec, 2016) and matrix factorization-based embedding (Qiu et al., 2018). In "Answer Selection" task, the answers provided for each question are classified to determine relevant and non-relevant answers. "Semantic similarity" is used to rank the answers relevant to a question in the answer ranking task. This field is covered in the works of (Mihaylov and Nakov, 2016), (Nakov et al., 2016b), and (Omari et al., 2016).

2.3. Answer Selection Models

There are basically two main types of answer classification methods: feature-based methods and deep learning methods. Several early works used feature-based methods for explicitly modeling the semantic relation between the question and answer (Nakov et al., 2015), (Huang et al., 2007), (Agichtein et al., 2008), (Nicosia et al., 2015). JAIST (Tran et al., 2015), HITSZ-ICRC (Hou et al., 2015), and QCRI (Nicosia et al., 2015) utilize typical features such as special component features, word matching features, non-textual features, and topic-modeling-based features. A simple classifier such as Support Vector Machine (SVM) or KNN is applied to the features and easily were selected. In (Yang et al., 2015), a lexical-semantic feature method employing word/lemma matching is considered for baseline.

The use of deep learning based methods reduces feature engineering to a large extent, as they automatically learn all features through end-to-end training. In light of significant advancements in deep learning neural networks, considerable recent researches have applied deep learning-based methods to perform answer classification in CQA (Tan et al., 2015), (Xiang et al., 2017), (Xiang et al., 2016), (Wen et al., 2019), (Yang et al., 2019). Typically, they (Mihaylov and Nakov, 2016), (Nakov et al., 2016b), (Omari et al., 2016), (Zhou et al., 2015), (Othman et al., 2019) use a Convolutional Neural Network (CNN) or Long Short Term Memory (LSTM) network for matching the question and answer. *Word2vec* (Mikolov et al., 2013), *Glove* (Pennington et al., 2014), and *fastText* (Joulin et al., 2016) as non-contextualized word embedding provide fixed representation for each word and do not capture its context in different sentences.

Recently, for learning sentence representation, various attention models based on the transformer model have been proposed (Vaswani et al., 2017). Transformer networks also serve as an encoder or decoder for some models in different tasks (Cer et al., 2018), (Radford and Narasimhan, 2018). Nowadays, *BERT* (Devlin et al., 2019) and *RoBERTA* (Liu et al., 2019) are used widely as contextualized word embeddings. In (Laskar et al., 2020), a model is presented which integrates contextualized embeddings with the transformer encoder (CETE) for sentence similarity modeling. CETE is based on contextualized embeddings (BERT, RoBERTA, and ELMo (Embeddings from Language Models (Peters et al., 2018))). There are two approaches in CETE, namely, features-based and finetuning-based. We use the feature-base approach here.

Most of these models are geared towards English, leaving multilingual models with limited resources to cover other languages. Multilingual Language Models such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020) apply the power of pretraining to multiple languages. We use the standard zero-shot and cross-lingual transfer learning such as mBERT and XLM-RoBERTa (XLM-R) to achieve more accurate representations for our task. We also employ ParsBERT (Farahani et al., 2021), the Persian pre-trained language model, and compare it with the multi-lingual representation models. We utilize them as encoders and measure the similarity between the vector representation of two sentences produced by these models.

3. Dataset Construction

A CQA is a powerful mechanism that usually consists of two steps: (i) creating a free-form question and (ii) posting various answers, thus creating an extensive collection of the desired question by users in order to obtain specific answers to their questions. The new question can be related to one or more previous questions in the forum. The best answer is sometimes marked in the question-answer threads that the lists of replies are sorted chronologically. The meta-information includes the posting date, the user who asked/answered, a category question, and answer tags.

Several research CQA datasets are conducted in English as well as a few other languages. We build the first CQA dataset in Persian, called PerCQA, designed for research purposes. To construct the dataset, we initially study the standard CQA datasets in English. We follow the standard approach in CQA datasets and crawl user forums to construct our dataset. Table 2 illustrates a summary of the steps involved in building PerCQA. In the following subsections, we describe the process of creating the PerCQA dataset.

3.1. Data Acquisition

We perform a complete analysis of various Persian forums and examine the ranking websites in Alexa Internet [9] and Similarweb.com[10]. A website was chosen for its question source based on the following characteristics: (i) having the highest rank in Iran, (ii) having numerous users, (iii) containing irrelevant answers in the sequence of replies to a question, (iv) having more than %80 traffic from search in alexa.com, (v) ordinary people are using it, not necessarily specialists in different fields.

[9] https://www.alexa.com/siteinfo/ninisite.com

[10] https://www.similarweb.com/website/ninisite.com

1. Data acquisition	a. Exploring various CQA Persian websites. b. Crawling the desired website(www.ninisite.com).
2. Data annotation	a. Setting Annotation Guidelines(AG) for the questions(Labels: Valid, Invalid). b. Setting AG for the answers(labels: Good, Bad, Potential). c. Developing an application for manual labeling. d. Crowdsourcing: question labeling(4people), answer labeling(12people). e. Evaluation of labeling(by Cohen's Kappa coefficient)(4people). f. Repeating the previous steps with improving the quality of AG and labeling answers again.

Table 2: The steps of construction of PerCQA.

On alexa.com, Ninisite is ranked approximately 20th in Iran and about 2,049 in the world, and also it has all the above characteristics. Therefore, `www.ninisite.com` is used as our question source, and from May 5th, 2017, until December 21st, 2020, about 1400 questions and the corresponding logs are crawled. However, we do not use all of the questions, mainly because some of them are advertisements or polls. Therefore, we need to evaluate the types of questions first to filter out some of them. The description of *Ninisite*'s structure is given in the following subsection.

3.2. Structure of Ninisite

Ninisite is the first and largest online forum in Persian for children, mothers and families. This website includes diverse sections. There is a section called "/tba:dol næzær/" (idea exchange) with 20 categories. Furthermore, there are specific forums called "/ta:la:r/" and any question asked in each forum is called "/ta:pik/". Webmasters with this slogan invite their users to participate in the forum. "Please participate in the Ninisite exchange and connect with thousands of other mothers." We examine the frequently asked questions about women, children, and medical subjects.

3.3. Data Annotation

Data annotation is necessary for this dataset because we aim to run strong supervised learning methods. It is the most crucial component of our work. There are three steps: question annotation, answer annotation, and evaluation of labeling quality. We develop a tool for data annotation which selects appropriate questions and answers. This tool drastically speeds up collecting datasets and also is an efficient way to collaborate on annotation projects. We crowdsource the annotation through a particular platform with 20 users.

Question annotation We review about a hundred questions and set up an annotation guideline for them. Table 3 shows annotation guidelines for the questions. For example, "Hiring a hairdresser, if you want to learn hairdressing, send me a message and I will tell you the conditions." or introducing a book with this post, "This topic makes your life 180 degrees better" [11] are marked invalid questions. It is not included in the dataset because there is no helpful answer to this question.

In our data annotation tool, three stages of the cascaded UI are designed. The first step displays the questions. "Valid" and "Invalid" labels are assigned to the questions in the question selection method. Valid questions are entered into the database and transferred to the next labeling phase, and invalid ones are deleted. Then, the UI moves on to the next step. After choosing appropriate questions, the system enters the second stage, labeling the answers.

Type of questions	**Label**
Advertisement	Invalid
Polls	Invalid
News	Invalid
Less than 3 answers and more than 300 answers	Invalid
Collaboration announcement	Invalid
Otherwise	Valid

Table 3: "Annotation guidelines" for questions.

Type of answers	**Label**
Dialogue, Advertisement, Not Persian, Greeting, Sympathy, stickers, Acknowledgment, Persian typed in English, and other comments.	*Bad*
Referred to other sources such as a related link or site (URLs) or a special page in social media and so on.	*Potential*
Partial or complete relevant answer to the given question	*Good*

Table 4: "Annotation guidelines" for answers.

Answer annotation After internal labeling of a trial dataset (the 100 selected questions) by several independent annotators, we prepare detailed annotation guidelines for labeling answers. The answers are classified as "Good", "Bad", and "Potential". "Good" labels are assigned to relevant answers, and potentially useful answers are labeled "Potential". The irrelevant answers

[11] `https://www.ninisite.com/discussion/topic/7764925`

Question: What is your suggestion to heal Cervical Disc? It bothers me a lot. (Valid Question)
Answer 1: Dr. Mohammad Kamali is a professional Surgeon who can treat you without surgery. I had a rigid Cervical Disc with a lot of pain. After some therapy sessions, his pain disappeared (Good Answer)
Answer 2: I had a cervical disc you should take a careful care and never bend your neck. (Good Answer)
Answer 3: For more information about Cervical Disc, check Dr. Samadian's website and type your questions https://drsamadian.com/cervical-disc/ (Potential Answer)
Answer 4: What medicines have been prescribed? (Bad Answer)
Answer5: The doctor told my mother that her Cervical Disc is in a lousy condition. The nerves are torn. She should pass 15 sessions of physiotherapy, but my mother can't move at all. (Bad Answers)

Table 5: The translation of a question and some of its answers in PerCQA.

(bad, dialog, non-English, other) take "Bad" labels. Table 4 shows "Annotation guidelines" for answers, a labeling guide to increase agreement between the annotators. Twelve annotators tag the answers using their unique user IDs in this phase. We ask three workers to tag each answer and select the correct label by majority voting.

Quality assessment Assessing the labeling quality of answers is the third stage of our data annotation. We used Cohen Kappa criteria to evaluate the quality. Cohen's kappa coefficient is a statistic that is used to measure inter-rater reliability for categorical items. Consequently, 45.28% of the total data are re-tagged for analysis by two groups consisting of two individuals. The judges' harmony levels with the final labels and the number of agreements on the matrix's main diagonal is shown in Table6. The Cohen's kappa (percentage of agreement) is 80% (Unweighted kappa: 0.802).

Labels	**Good**	**Bad**	**Potential**	**Total**
Good	4195	59	128	4382
Bad	412	4216	368	4996
Potential	79	63	384	526
Total	4686	4358	880	9924

Table 6: The number of agreements and disagreements in the labeling process to calculate the kappa criterion.

4. PerCQA Dataset

There are, in total, 989 questions and 21,915 corresponding answers in PerCQA. The content of questions and answers is kept chiefly unchanged. In this section, the structure of the dataset, its features, and statistics are explained.

4.1. Structure

The translation of a sample question and a subset of its answers in PerCQA are illustrated in Table 5. Figure 1 indicates the original version of the question and some of its answers in the proposed dataset. There is one of the latest questions on the subject of "Cervical Disc" in this link[12]. As you can see, as mentioned before, it is asked and answered in a very informal way.

It is a sequence of questions with the following attributions: (i) question identifier (QID); (ii) the questioner Username (QUsername); (iii) the day and date the question is posted on the forum (QDate); (iv) the questioners consider a subject for their question and write it (QSubject); (v) the whole question (QBody).

A list of comments follows each question, and each answer contains the following attributes: (i) comment identifier (CID); (ii) identifier of the user posting the comment (CUserID); (iii) Username of the respondent (CUsername); (iv) comment body (CBody); (v) an annotator rating of whether the comment is "Good", "Bad" and "Potential" (CGOLD).

At testing time, CGOLD is hidden, and the system is instructed to predict CGOLD. It can be seen that questions and answers in PerCQA are informal and can be considered a problem in building a suitable language model. Other challenges include incorrect capitalization and punctuation, misspelling, as well as slang and elongations. To learn word embeddings, we collect unlabeled texts, including 10000 questions and their comments that are posted on the site.

4.2. Statistics

The statistics of the dataset are given in Table 7. Since one of the extremely important features in feature-based methods is "the length of questions and answers", its mean and median are reported in this table for a more comprehensive review. The existence of one-word questions and answers, as well as very long ones in the dataset, motivates us to report the median in addition to the mean. A report on the distribution of dataset labels is shown in Table 8.

We manually extracted some simple features to perform feature engineering. Although, because of our pre-trained model, we do not require to perform feature extraction, we extracted a few simple features and depicted their correlation with the three classes. This is performed to get a better understanding of the underlying latent relationships in data. The number of URLs, Emojis, Numbers, and Non-Persian words are

[12] https://www.ninisite.com/discussion/topic/6365515

پرسش: دیسک گردن، خیلی اذیتم میکنه. برای درمانش چه پیشنهادی دارین؟ (پرسش معتبر)
پاسخ ۱ : دکتر محمد کمالی عالیه بدون جراحی درستت میکنه. من بابام دیسک گردن شدید داشت چند جلسه رفت پیشش الان دیسکش رفع شده. (پاسخ خوب)
پاسخ ۲ : من دیسک گردن داشتم واقعا باید رعایت کنید و دیگه سر رو خم نکنید. اصلا. (پاسخ خوب)
پاسخ ۳ : برای کسب اطلاعات بیشتر درباره دیسک گردن به سایت دکتر صمدیان مراجعه کنید و سوالات خود را با ایشان در میان بگذارید . https://drsamadian.com/cervical-disc (پاسخ بالقوه).
پاسخ ٤ : رفتی دکتر قرص چی بهت داد؟ (پاسخ بد).
پاسخ ٥ : به مامانم گفته دیسک گردنش شدیده و عصب ها پاره شده و ۱۵ جلسه فیزیوتراپی اصلا نمیتونه بلندشه. (پاسخ بد).

Figure 1: A question and some of its answers in PerCQA.

	Number of Questions	Number of Answers	Mean length of questions	Mean length of Answers	Median length of questions	Median length of answers
Train (70%)	692	15,454	171	84	129	61
Dev (10%)	99	2,164	174	92	128	64
Test (20%)	198	4,297	176	93	130	64
Total	989	21,915	173	89	129	62

Table 7: PerCQA statistics.

Answers	#Good	#Bad	#potential
21915	10467	10700	748
100%	47.8%	48.7%	3.14%

Table 8: Label distribution in PerCQA.

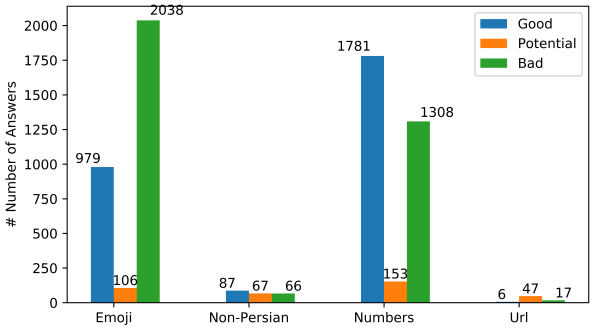

Figure 2: The number of Answers that have some features

reported in answers and help us gain new findings from the data. For example, there are more stickers in the answers that have a "Bad" labels. The number of "non-Persian" items in "Good" labels is higher than others, because the names of drugs, cosmetics, home appliances or dowry brands are in English. In addition, phone numbers and digits exist in the more answers with "Good" labels significantly and one-Tenth of the "Potential" tags have URLs. Figure 2 reveals these features. Furthermore, it is important to consider the length of the answers when assessing classification quality. The number of answers in different lengths according to their tags are shown in figure 3. As it can be seen, the answers with fewer than 25 characters are more likely to be labelled "Bad" and decrease sharply with increasing length.

5. Experiments

PerCQA offers us the opportunities to evaluate CQA systems on various tasks in Persian. In this section, we use PerCQA for answer selection, implement several baseline systems and evaluate and analyze their results.

5.1. Baseline Methods

For evaluating the dataset, we select the answer selection task and choose four baselines and apply them to PerCQA. We use one feature-based method (Yang et al., 2015) and three deep-learning models. We use the implementation of these baselines which are described below:

- **PV-Cnt** (Yang et al., 2015)[13]:

  Word Count and Weighted Word Count are two word-matching features. The paragraph Vector (Le and Mikolov, 2014) is the cosine similarity score between the question vector and the sentence vector. They combined PV and word matching features by training a logistic regression classifier, referring to PV-Cnt.

- **BiLSTM-attention** (Tan et al., 2015):

[13] gist.github.com/shagunsodhani/7cf3677ff2b0028a33e6702fbd260bc5

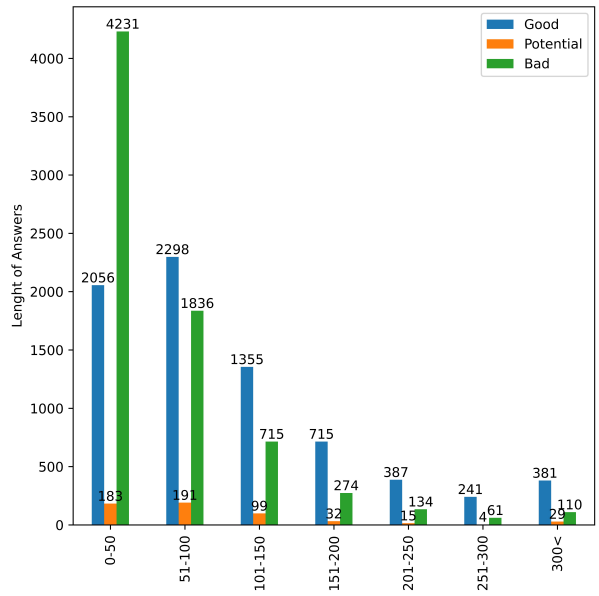

Figure 3: The distribution of answer labels based on their length

BiLSTM-attention[14] is a basic framework for building the embedings of questions and answers based on the BiLSTM model. BiLSTM generates distributed representations by the attention mechanism.

- **RCNN** (Zhou et al., 2018):

  Recurrent Convolutional Neural Network (RCNN) combines Convolutional Neural Network (CNN) with Recurrent Neural Network (RNN) to model the semantic relevance between questions and answers.

- **CETE** (Laskar et al., 2020)[15]:

  Contextualized Embeddings based Transformer Encoder (CETE) utilizes the pre-trained transformer encoder based models (BERT/RoBERTa) and integrates them for sentence similarity in the answer selection task. There are two approaches: i) feature-based approach, and ii) Fine-tuning-based approach.

5.2. Word Embeddings

We employ five pre-trained word embeddings. Three are contextual, including ParsBERT, XLM-R, and m-BERT. Two others are non-contextual or static, including Word2vec and Fasttext. For the contextual embeddings, we use the publicly available Pytorch models for each. We train the non-contextual embeddings on a corpus made from Ninisite forums. This corpus has about 2 billion tokens containing questions and answers from X threads. The static embeddings such as Word2vec and FastText are employed on the proposed dataset and the dimension of word embedding is adopted w=200. Besides, we compare the performance of a mono-lingual model (ParsBERT) versus multi-lingual models (mBERT and XLM-R).

5.3. Results

In the previous work, macro-averaged F1 is used to benchmark the answer selection task. Therefore, we make the comparison based on this measurement. Table 9 demonstrates the results of these methods on the PerCQA dataset. We combine various word embedding (Word2vec, FastText, ParsBERT, mBERT, XLM-R) with diverse baselines methods (PV-CNT, BiLSTM-attention, RCNN, CETE feature-based approach).
It is evident that PV-CNT achieves better results than BiLSTM-attention while both utilize Word2vec as word embedding. Therefore, we can conclude that lengthy and informal sentences may have contributed to the deep learning-based method's low performance. Models with contextual embeddings are significantly better than the ones with non-contextual embeddings. In PerCQA, there are often lengthy questions, so considering the context for representing each word makes it possible to achieve a better result. We also compare the results of mBERT and XLM-R, the multi-lingual contextual embeddings, with mono-lingual embeddings of ParsBERT. ParsBERT outperforms mBERT and XLM-R when training on the PerCQA data. However, we observe that preraining XLM-R on SemEval English datasets is very effective as its macro F1 improves from 50.71 to 61.14. Furthermore, another experiment sets the best result on our PerCQA dataset better than ParsBERT's 58.07 F1, and shows that cross-lingual transfer from English datasets to our PerCQA dataset is possible. Even the zero-shot cross-lingual results of XLM-R are descent (52.48) compared to training on the same language data (54.13).

6. Conclusion

We introduced PerCQA in the Persian Language for the task of answer selection in Community Question Answering (CQA). PerCQA contains 21,915 pairs of real questions and answers, which asked by a large number of users of various levels of literacy. We hope that PerCQA will promote the quality of Persian forums and enable further research in CQA tasks in the Persian language. We plan to release a new version of PerCQA in the future, which we expect to include data to perform more CQA tasks in Persian. We also hope that our experimental results will provide practical baselines for further research. Our dataset is available for download on GitHub[16].

[14] github.com/sachinbiradar9/Question-Answer-Selection

[15] https://github.com/tahmedge/CETE-LREC

[16] https://github.com/PerCQA

Word Embedding				Various models for Answer Selection	F1-Score
Non-Comtextualized	Word2vec			PV-CNT	41.03
				BiLSTM-attention	38.27
	FastText			RCNN	39.56
Contextualized	Mono-Lingual	**Pars-BERT**		CETE	**58.07**
	Multi-Lingual	m-BERT	Fine-tuned on PerCQA	CETE	50.71
		XLM-R		CETE	54.13
		XLM-R	Fine-tuned on SemEvalCQA datasets	CETE	52.48
			Fine-tuned on SemEvalCQA datasets + PerCQA	CETE	**61.14**

Table 9: Quantitative evaluation results on PerCQA, ordered by macro-averaged F1. Each row corresponds to the result of an answer selection model using one of the pretrained word embeddings.

7. References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183–194.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, pages 135–146.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *CoRR*.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, pages 7059–7069.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Devlin, J., Chang, M.-W., Lee, K., and 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)*, pages 4171–4186.

Farahani, M., Gharachorloo, M., Farahani, M., and Manthouri, M. (2021). Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, pages 3831–3847.

Feng, M., Xiang, B., Glass, M. R., Wang, L., and Zhou, B. (2015). Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820.

Ghasemi, N., Fatourechi, R., and Momtazi, S. (2021). User embedding for expert finding in community question answering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, pages 1–16.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Hou, Y., Tan, C., Wang, X., Zhang, Y., Xu, J., and Chen, Q. (2015). HITSZ-ICRC: Exploiting classification approach for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 196–202.

Huang, J., Zhou, M., and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence*, pages 423–428.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kunneman, F., Ferreira, T. C., Krahmer, E., and van den Bosch, A. (2019). Question similarity in community question answering: A systematic exploration of preprocessing methods and models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 593–601.

Laskar, M. T. R., Huang, J. X., and Hoque, E. (2020). Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5505–5514.

Le, Q. and Mikolov, T. (2014). Distributed representa-

tions of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv e-prints*, pages arXiv–1907.

Mihaylov, T. and Nakov, P. (2016). Semanticz at semeval-2016 task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 879–886.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., and Randeree, B. (2015). SemEval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 269–281.

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. (2016a). SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545.

Nakov, P., Màrquez, L., and Guzmán, F. (2016b). It takes three to tango: Triangulation approach to answer ranking in community question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1597.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017). SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48.

Nicosia, M., Filice, S., Barrón-Cedeno, A., Saleh, I., Mubarak, H., Gao, W., Nakov, P., MARTINO, G. D. S., Moschitti, A., Darwish, K., et al. (2015). Qcri: Answer selection for community question answering-experiment for arabic and english. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*.

Omari, A., Carmel, D., Rokhlenko, O., and Szpektor, I. (2016). Novelty based ranking of human answers for community questions. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 215–224.

Othman, N., Faiz, R., and Smaili, K. (2017). A word embedding based method for question retrieval in community question answering. In *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*.

Othman, N., Faiz, R., and Smaïli, K. (2019). Manhattan siamese lstm for question retrieval in community question answering. *OTM Confederated International Conferences On the Move to Meaningful Internet Systems*, pages 661–677.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Qiu, X. and Huang, X. (2015). Convolutional neural tensor network architecture for community-based question answering. In *IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1305–1311.

Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J. (2018). Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 459–467.

Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Tan, M., dos Santos, C., Xiang, B., and Zhou, B. (2015). Lstm-based deep learning models for non-factoid answer selection. *arXiv e-prints*, pages arXiv–1511.

Tran, Q. H., Tran, V. D., Vu, T. T., Nguyen, M. L., and Pham, S. B. (2015). JAIST: Combining multiple features for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 215–219.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5998–6008.

Wang, M., Smith, N. A., and Mitamura, T. (2007). What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 22–32.

Wang, Z., Fan, Y., Guo, J., Yang, L., Zhang, R., Lan, Y., Cheng, X., Jiang, H., and Wang, X. (2020). Match²: A matching over matching model for sim-

ilar question identification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 559–568.

Wen, J., Tu, H., Cheng, X., Xie, R., and Yin, W. (2019). Joint modeling of users, questions and answers for answer selection in cqa. *Expert Systems With Applications*, pages 563–572.

Xiang, Y., Zhou, X., Chen, Q., Zheng, Z., Tang, B., Wang, X., and Qin, Y. (2016). Incorporating label dependency for answer quality tagging in community question answering via cnn-lstm-crf. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1231–1241.

Xiang, Y., Chen, Q., Wang, X., and Qin, Y. (2017). Answer selection in community question answering via attentive neural networks. *IEEE Signal Processing Letters*, pages 505–509.

Yang, Y., tau Yih, W., and Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.

Yang, M., Tu, W., Qu, Q., Zhou, W., Liu, Q., and Zhu, J. (2019). Advanced community question answering by leveraging external knowledge and multi-task learning. *Knowledge Based Systems*, pages 106–119.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). Jec-qa: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9701–9708.

Zhou, G., He, T., Zhao, J., and Hu, P. (2015). Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 250–259.

Zhou, X., Hu, B., Chen, Q., and Wang, X. (2018). Recurrent convolutional neural network for answer selection in community question answering. *Neurocomputing*, pages 8–18.