Diverse and Non-redundant Answer Set Extraction on Community QA based on DPPs

Shogo Fujita[1]*, **Tomohide Shibata**[2] **and Manabu Okumura**[1]
[1]Tokyo Institute of Technology
[2]Yahoo Japan Corporation
{fujisyo,oku}@lr.pi.titech.ac.jp
tomshiba@yahoo-corp.jp

Abstract

In community-based question answering (CQA) platforms, it takes time for a user to get useful information from among many answers. Although one solution is an answer ranking method, the user still needs to read through the top-ranked answers carefully. This paper proposes a new task of selecting a diverse and non-redundant answer set rather than ranking the answers. Our method is based on determinantal point processes (DPPs), and it calculates the answer importance and similarity between answers by using BERT. We built a dataset focusing on a Japanese CQA site, and the experiments on this dataset demonstrated that the proposed method outperformed several baseline methods.

1 Introduction

Community-based question answering (CQA) platforms have been used by many people to solve problems that cannot be solved simply by searching in books or web pages. On a CQA platform, the questioner posts a question and waits for others to post answers. The posted answers are viewed by not only the questioner herself but also other users.

A single question may receive several to dozens, even hundreds, of answers, and it would take a lot of time to see all the answers. The questioner chooses the most helpful answer, called the best answer, and users can view this answer as a representative answer. However, there may be more than one appropriate answer to a question that asks for reasons, advice, and opinions, while there is often only one appropriate answer for a question that asks for facts, etc. For example, Table 1 shows example answers to the question *"Why do birds turn their heads backwards when they sleep?"*, which asks for reasons. The best answer A1 says it is to keep out the cold and A4 says it is for balance, which is also an appropriate answer. In such a case, if users view only the best answer, they miss other appropriate answers.

One possible way to deal with this problem is to use ranking methods (Wang et al., 2009; Xue et al., 2008; Bian et al., 2008; Zhao et al., 2017; Tay et al., 2017; Zhang et al., 2017; Laskar et al., 2020). Since the studies use the best answers as training data, similar answers are ranked higher although the existing methods try to reduce redundancy by using the similarity between a question and an answer or between answers. As a result, users need to read through carefully the answers that are ranked higher to obtain the information they want to know.

We propose a new task of selecting a diverse and non-redundant answer set from all the answers, instead of ranking the answers. We treat the set of answers as the input, from which the system determines both the appropriate number of answers to be included in the output set and the answers themselves. The datasets used in the conventional ranking task of CQA are not suitable for our task because they contain only a ranked list of answers, whereas we need to determine the number of answers to be included in the output set in addition to the answers themselves. Therefore, we have constructed a new dataset for our task that contains diverse and non-redundant annotated answer sets.

We use determinantal point processes (DPPs) (Macchi, 1975; Aho and Ullman, 1972) to select a diverse and non-redundant answer set. DPPs are probabilistic models that are defined over the possible

*This work was done as an intern at Yahoo Japan Corporation.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5309–5320
Barcelona, Spain (Online), December 8-13, 2020

Q	Why do birds turn their heads backwards when they sleep?	
	answer	gold
A1	**...Because it's cold. They keep their beaks in their feathers to protect them from getting cold.**	✓
A2	This is because it reduces the surface area of the body to prepare for a drop in body temperature and to prevent physical exhaustion.	
A3	Well balanced.	
A4	They look back to stabilize their bodies by shifting their center of gravity on the vertical line of their feet to stabilize their bodies.	✓
A5	I think it's to use their feather as pillows.	

Table 1: Example of a question and its answers where multiple answers are considered to be appropriate. The answer in bold is the best answer. Sentences in the same color indicate they have the same content. Gold indicates answers included in the reference answer set.

subsets of a given dataset and give a higher probability mass to more diverse and non-redundant subsets. To estimate the answer importance and the similarity between answers, we use BERT (Devlin et al., 2019), which has achieved high accuracy in various tasks.

Focusing on a Japanese CQA site, we built training and evaluation datasets through crowdsourcing and conducted experiments demonstrating that the proposed method outperformed several baseline methods.

2 Related Work

2.1 Work on CQA

CQA has been actively studied (Nakov et al., 2017; Nakov et al., 2016; Nakov et al., 2015). The most relevant task to our study is *answer selection*. There are two approaches to the answer selection task. One approach is to predict the best answers (Tian et al., 2013; Dong et al., 2015). The other is to rank the entire list of potential answers (Wang et al., 2009; Xue et al., 2008; Bian et al., 2008; Zhao et al., 2017; Tay et al., 2017; Zhang et al., 2017; Laskar et al., 2020). The problem with these ranking methods is that they do not take account of the similarity between answers. To solve this problem, Omari et al. (2016) proposed a method for ranking answers based on the similarity between questions and answers and between answers, and Harel et al. (2019) proposed a method for ranking answers using a neural model that was learned to select answers that are similar to the best answer and less similar to the answers already selected. These studies are related to our method in that they take account of diversity in selecting from the answers in CQA. Moreover, they address the task of selecting one appropriate answer or sorting them in appropriate order, which is different from our task.

A number of methods of summarizing answers in CQA (Chowdhury and Chakraborty, 2019; Song et al., 2017) also take diversity into account. These methods extract sentences of a fixed length, and therefore, cannot determine an appropriate output length. Our method is different in that it can determine the appropriate output length by using DPPs. Some methods of question retrieval in CQA also take diversity into account (Cao et al., 2010; Szpektor et al., 2013).

2.2 Work on Diversity

Taking account of diversity is important in information retrieval and summarization, and methods using maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998) have been proposed (Boudin et al., 2008; Guo and Sanner, 2010).

DPPs (Macchi, 1975; Aho and Ullman, 1972) has also been proposed as a way for taking account of diversity. This method performs better than MMR in these fields. DPPs have been applied to recommendation problems as a way of taking diversity into account (Wilhelm et al., 2018; Chen et al., 2018; Mariet et al., 2019). Multi-document summarization models combine DPPs and neural networks to take account of diversity (Cho et al., 2019a; Cho et al., 2019b). DPPs have been also used in studies on clustering verbs (Reichart and Korhonen, 2013) and extracting example sentences with specific words (Tolmachev and Kurohashi, 2017).

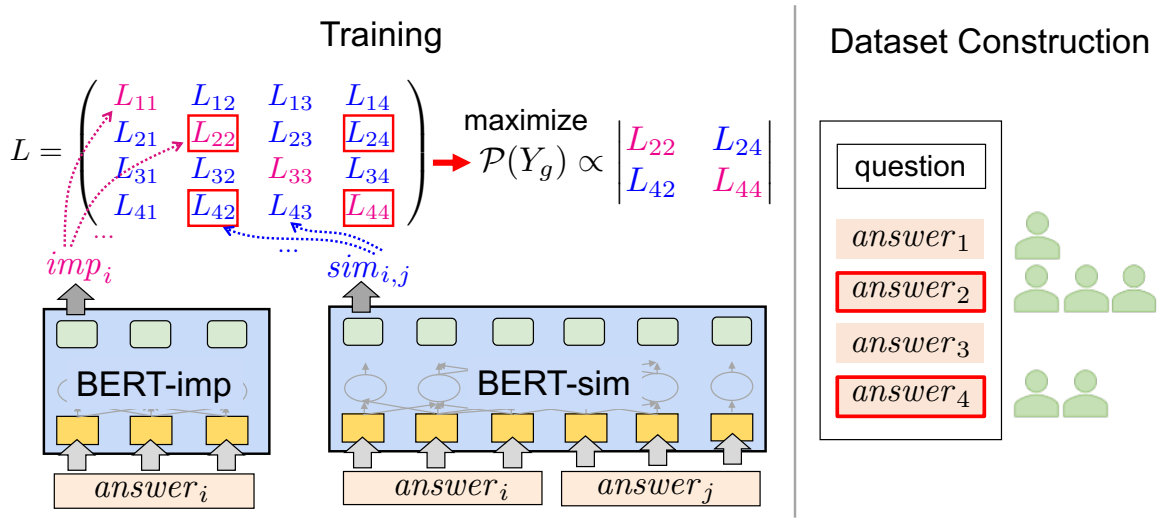

Figure 1: Overview of the proposed method.

3 Proposed Method

The overview of our method is shown in Figure 1. Our method uses DPPs to select a diverse and non-redundant subset from a set of answers. The DPPs assign a higher probability to a subset in which each answer is important and the answers are diverse. BERT (Devlin et al., 2019) is used for estimating the importance of an answer and the similarity between answers. In the example shown in Figure 1, answers 2 and 4 are annotated as a gold answer set. The probability of choosing these gold answers is defined by the DPPs. Intuitively, BERT is fine-tuned so that the importance of answers 2 and 4 is higher and the similarity between answers 2 and 4 is lower.

3.1 DPPs

DPPs define the probability measure $\mathcal{P}$ over all subsets (2^N) of a ground set containing all the answers $\mathcal{Y} = \{1, 2, ..., N\}$ where N denotes the number of answers. Our method computes the probability of choosing a subset of answers based on the kernel matrix $L \in \mathbb{R}^{N \times N}$ that contains the importance of the answers and the similarities between answers, and it selects the subset $Y \subseteq \mathcal{Y}$ with the highest probability. The probability measure $\mathcal{P}$ is defined as

$$\mathcal{P}(Y; L) = \frac{det(L_Y)}{\sum_{Y'} det(L_{Y'})} \tag{1}$$

$$= \frac{det(L_Y)}{det(L + I)}, \tag{2}$$

where $L_Y \in \mathbb{R}^{|Y| \times |Y|}$ is a submatrix containing only entries indexed by elements of Y, $I \in \mathbb{R}^{N \times N}$ is the identity matrix, and $det(\cdot)$ is the determinant of a matrix. The kernel matrix L is a positive semi-definite matrix. The i, j-th component of the kernel matrix $L_{i,j}$ is defined as

$$L_{i,j} = imp_i \times imp_j \times sim_{i,j}, \tag{3}$$

where imp_i is the importance of the i-th answer and $sim_{i,j}$ is the similarity between the i-th and j-th answers. The similarity of the same answer $sim_{i,i}$ is set to 1. From the definition of the determinant, the DPPs select important answers and do not select similar answers at the same time. Following Wilhelm et al. (2018), the kernel matrix is decomposed into eigenvalues, and negative eigenvalues are replaced with tiny values to satisfy the semi-positive definiteness constraint.[1] In the training, the parameters of BERT are updated so that the following negative log-likelihood is minimized:

$$-log(\mathcal{P}(Y_g; L)) = -log(det(L_{Y_g})) + log(det(L + I)), \tag{4}$$

[1]While Wilhelm et al. (2018) replace the negative eigenvalues with zero, we replace them with tiny values to stabilize the learning.

where $Y_g \subseteq \mathcal{Y}$ is the gold answer set.

At inference time, our method computes the probability of all the candidate answer sets and selects the one with the highest probability. Note that since the dataset used in our experiments had a relatively small number of answers, we did not need to adopt an approximate method by sampling.

3.2 BERT Architecture

The architecture has two mechanisms based on BERT: BERT-imp, which uses the BERT mechanism to calculate the answer importance, and BERT-sim, which calculates the similarity between answers. Following the BERT paper, the `[CLS]` and `[SEP]` tokens are added to the beginning and end of a sequence of tokens, respectively.

3.2.1 BERT-imp

BERT-imp is a mechanism for predicting the importance of an answer. imp_i is calculated as follows:

$$imp_i = \exp(\mathrm{MLP}^{imp}(\mathrm{BERT}^{imp}_{CLS}(\mathrm{input}^{imp}(a_i)))), \tag{5}$$

$$\mathrm{input}^{imp}(a_i) = \texttt{[CLS]}\, a_i\, \texttt{[SEP]}, \tag{6}$$

where a_i is the i-th answer, and $\mathrm{BERT}^{imp}_{\mathrm{CLS}}$ is a function that returns the hidden state vector corresponding to `[CLS]` by vectorizing the input using the BERT mechanism. MLP^{imp} is a feed-forward network with a single hidden layer that outputs the importance of an answer. Since DPPs are log-linear models, the $\exp$ function is applied after the output of the last layer.

3.2.2 BERT-sim

BERT-sim is a mechanism to compute the similarity between answers. Following standard practice, we separate the two answers with `[SEP]` tokens. $sim_{i,j}$ is calculated as follows:

$$sim_{i,j} = \sigma(\mathrm{MLP}^{sim}(\mathrm{BERT}^{sim}_{\mathrm{CLS}}(\mathrm{input}^{sim}(a_i, a_j)))), \tag{7}$$

$$\mathrm{input}^{sim}(a_i, a_j) = \texttt{[CLS]}\, a_i\, \texttt{[SEP]}\, a_j\, \texttt{[SEP]}, \tag{8}$$

where $\mathrm{BERT}^{sim}_{\mathrm{CLS}}$ is the same mechanism as $\mathrm{BERT}^{imp}_{\mathrm{CLS}}$, and MLP^{imp} is the same as MLP^{sim}. The σ function is applied to make the value in the range $[0, 1]$.

Our method is similar to the DPPs-based summarization model (Cho et al., 2019b). Cho et al. (2019b) use a single document summary dataset to train BERT-imp, and BERT-sim and the feature weight for BERT-imp is learned during DPP training. In our method, because we did not have the appropriate data to pre-train BERT-imp and BERT-sim, BERT-imp and BERT-sim are fine-tuned in an end-to-end fashion using only the training data for the target task.

4 Experiments

4.1 Dataset Construction

Since there was no dataset for the task of selecting a diverse and non-redundant answer set in the CQA field, we constructed one for this study. We focused on Yahoo! Chiebukuro,[2] the largest CQA service in Japan. From among categories with many questions, we excluded categories, in which almost all the answers are different, such as cooking, and categories, in which most of answers are too long, such as human relationships, and we selected questions in the healthcare, diet, and pet categories. We excluded questions and answers that were too short and uninformative in each category; in particular, we excluded those questions whose best answers were less than 110 characters in length and those questions with less than five answers.

We used a crowdsourcing service called Yahoo! Crowdsourcing for annotation. For a total of approximately 10,000 questions in the above three categories, the questions and answers were presented to 10 workers, and they were asked to select a diverse and non-redundant answer set. When crowdsourcing is

[2] https://chiebukuro.yahoo.co.jp/

	# of workers who chose the same answer set			
	4	3	2	1
A1	✓	✓	✓	✓
A2		✓		
A3			✓	
A4				✓
A5				

Table 2: Example of annotation results.

	# of answers in gold answer sets					
	1	2	3	4	5	6
# of questions	2552	923	155	34	10	6
ratio	0.69	0.25	0.06			

Table 3: Distribution of the number of answers in gold answer sets.

Require: $V = \{(S_1, N_1), (S_2, N_2), ..., \}$
$G \Leftarrow \{\}$
$min\_num \Leftarrow 2$
$end\_num \Leftarrow 0$
if Maximum number of workers matched is less than 4 **then**
    This instance is not used.
end if
for each S_i, N_i in V **do**
    if $N_i < min\_num$ or $N_i < end\_num$ **then**
        break
    end if
    if $G \subseteq S_i$ **then**
        $G \Leftarrow G \cup S_i$
    else
        $end\_num \Leftarrow N_i$
    end if
end for

Figure 2: Algorithm for selecting the gold answer set G.

used, only instances that most workers agree on should be chosen for training and evaluation datasets. In our task, the chosen answer set may vary from person to person, and the number of the chosen answers tends to be small (usually one). As such, simply choosing only the questions that many workers agree on would be undesirable because the dataset would be small, and the number of answers in almost every gold answer set would be one.

Therefore, we loosened the concept of "matching" and devised an algorithm that allows multiple answers to be selected as much as possible. Specifically, the algorithm first chooses only those instances for which there is some degree of agreement in the annotation results. The algorithm searches in order from the answer set chosen by most workers. If the answer set includes a current gold answer set that has already been chosen, the algorithm adds it to the current gold answer set. We refer to Table 2, which shows a typical annotation result, to illustrate the behavior of the algorithm. The most popular choice is {A1}, so the algorithm first adds it to the current gold answer set. {A1, A2} chosen by three workers includes the already chosen current gold answer set {A1}, so the algorithm adds A2 to the current gold answer set. Since {A1, A3} chosen by two workers does not include the already chosen current gold answer set {A1, A2}, the algorithm stops. The algorithm outputs {A1, A2} as the gold answer set.

The detailed algorithm is shown in Figure 2. V is an aggregated list of votes from workers for each selected answer set. For each aggregate, S represents the answer set and N represents the number of workers who chose S. V is sorted in descending order of N, and if N is the same for two or more items on the list, it is sorted in ascending order of the number of answers in S. The i-th element of V is (S_i, N_i). In the case of the example shown in Table 2, $S_1 = \{A1\}$, $N_1 = 4$, $S_2 = \{A1, A2\}$, and $N_2 = 3$.

If the answer set chosen by k workers does not include the current gold answer set, the algorithm should look at all the answer sets chosen by the other k workers before terminating. For this reason, we use a variable end_num indicating the number of workers who chose the last answer set that the algorithm would see. To consider only the answer sets selected by two or more workers, the constant min_num is set to be 2.

Table 4 shows an example of a non-selected question. Both A3 and A5 contain information that it is important to get into the habit of training every day and not to overdo it. Although there are minor differences, the answers are similar. In this case, three workers chose A3 and three other workers chose A5, which caused a split in opinion. Such questions were not included in the dataset because of the difficulty in determining an appropriate answer set.

The number of questions in the dataset is 3,680. Table 3 shows the distribution of the number of answers in the gold answer sets. The average number of the answers is 1.38. While 69% of the questions

Q	I've heard that abdominal exercises, lying on a back and lifting and lowering legs, are effective in reducing the amount of fat in the lower abdomen. I've done too much and my muscles are sore, should I take a break from this?					
	answer	# of workers who chose the same answer set				
		3	3	2	1	1
A1 A2	In my case, when I have muscle pain, I'll keep doing it... Don't overdo it, but don't take a break.					✓
A3	It's important to continue every day. The key is to continue in moderation and without strain...If it hurts too much, you can take a break.	✓		✓	✓	
A4	I'm not sure which muscle is sore, but if it's not your stomach, it doesn't have much to do with your lower abdomen, does it?...				✓	
A5	**How about working your muscles other than your abdominal muscles? ...It's important to get into the habit of exercising every day rather than pushing yourself in the beginning...**		✓	✓	✓	

Table 4: Example of the instance that was not chosen in our dataset.

have one answer in the gold answer sets, 6% of the questions have three or more answers. Our dataset was randomly split into a training set (2,576 instances), development set (368 instances), and testing set (736 instances).

4.2 Experimental Settings

Accuracy was used to assess the overall agreement, while precision, recall, and f1 were used to assess partial agreement. Our method used $\text{BERT}_{\text{BASE}}$, which was pre-trained in the Whole Word Masking setting from Japanese Wikipedia.[3] The number of dimensions of the hidden layers of MLP^{imp} and MLP^{sim} was set to 100. Following the BERT paper, AdamW was used as the optimizer. The learning rates (2e-5, 3e-5, and 5e-5) used in the BERT paper were found to be rather large for our task, so thus a smaller one, 2e-6, was used instead. We trained and evaluated each model three times for 5 epochs and computed the average scores of the models with the highest accuracy on the development set.

Since our task is one of selecting an appropriate answer set, our method cannot be directly compared to the previous studies where the appropriate number of answers in an answer set is unknown. Therefore, the following baseline methods determine the importance of an answer, and select the fixed number of answers for all the questions. Two variants for each baseline method were considered: one outputting the most important answer, the other outputting the two most important answers, since the average number of gold answer sets is 1.38. Hereafter, the method that chooses one in a method M is called M-1, and the method that chooses two is called M-2. The following baselines were compared with our proposed method:

random As a simple baseline, we experimented with a model to select answers randomly.

wordnum Since a longer answer has a lot of information, it is likely included in a gold answer set. This baseline selects an answer in descending order of its word number.

lexrank LexRank (Erkan, 1997) is a graph-based method for text summarization, which calculates the importance of a sentence by taking account of the similarity between sentences. This method was applied to our task. This baseline selects an answer in descending order of LexRank score.

ranking To compare with the conventional ranking method (Harel et al., 2019), we ran a model to select answers ranked by that method. This model requires that one appropriate answer be chosen in advance. Therefore, we ran a model to select the best answer using BERT. We used approximately 55k questions to learn the model to choose the best answer. For training on the dataset, we constructed five input sequences, each containing the question and the answer[4].

The model calculated the importance of each answer using the hidden state vectors corresponding to `[CLS]` and selected the answer with the highest score.

To verify the effectiveness of BERT, we experimented with a model that used bidirectional long short term memory (bi-LSTM) (Hochreiter and Schmidhuber, 1997) for the encoder (**LSTM**). The LSTM

[3]https://github.com/cl-tohoku/bert-japanese

[4]If a question had more than five answers, five answers were randomly sampled, including the best answer.

Model	Acc.	Prec.	Rec.	F1
random-1	0.141	0.262	0.193	0.222
random-2	0.022	0.257	0.378	0.306
wordnum-1	0.395	0.643	0.473	0.545
wordnum-2	0.129	0.495	**0.728**	0.589
lexrank-1	0.337	0.565	0.416	0.479
lexrank-2	0.098	0.450	0.662	0.536
ranking-1	0.325	0.561	0.413	0.476
ranking-2	0.056	0.385	0.567	0.459
LSTM	0.418	0.672	0.495	0.570
Proposed	**0.477**	**0.719**	0.582	**0.643**

Table 5: Evaluation results for the test set.

encoder outputs the concatenation of the last hidden states of forward and backward LSTM. We replaced the BERT-based encoders used by BERT-imp and BERT-sim with bi-LSTM. The word embeddings were initialized using pre-trained word2vec embeddings (Mikolov et al., 2013a; Mikolov et al., 2013b). The word2vec embeddings were pre-trained using Japanese Wikipedia. The optimizer was AdamW, and a learning rate of 5e-4 was chosen from among {1e-3, 5e-4, 1e-4} by using the development set.

4.3 Results

Table 5 shows the results for the test set. Among the baseline methods, the models that chose one answer achieved higher accuracy and precision, and the models that chose two answers achieved higher recall and F1. In terms of accuracy, **wordnum-1**, which selects the answer with the most words, performed the best among the baseline methods. The result is considered reasonable, as longer answers are more likely to contain more information.

Proposed outperformed the simple baselines and **ranking**, and its results demonstrated that the approach of creating a dataset for selecting a diverse and non-redundant answer set and learning on DPPs is effective. **Proposed** also outperformed **LSTM**, which is consistent with recent studies showing that pre-trained models are effective.

4.4 Analysis

We performed several analyses to better understand how our method works.

Adding a question to BERT-imp In BERT-imp introduced in Sec 3.2.1, the input is an answer. Not only the answer but also the question likely contributes to the prediction of the answer's importance. For this reason, a method that adds a question to the input in BERT-imp was tested (**+Q_BERT-imp**).

The upper-half of Table 6 shows the results. **+Q_BERT-imp** performed almost as well as **Proposed**, which demonstrates that the question did not contribute to the prediction of the answer importance. Since our method cannot capture the relevance of questions and answers with a small corpus, it is considered sufficient for BERT-imp to focus only on answers.

Table 7 shows an example of different outputs between **Proposed** and **+Q_BERT-imp**. While A3 and A4 are the same in that they both recommend manila grasses as the easiest lawn to maintain, they have different contents about planting methods, so it is appropriate to choose both. The reason why **+Q_BERT-imp** only chose A3 as the correct answer set to this question is that the model rated the importance of A3, which is similar to the question, higher, while it rated the importance of A4, which included a lot of content not directly related to the question, lower. The dataset contains answers that are not appropriate answers to the question, such as A2 in this case; answers such as these can be judged by looking only at the answers.

Pre-training for BERT-imp The task of choosing the best answer might be relevant to our task. To compensate for the paucity of annotated data, we can pre-train BERT-imp in the task of choosing the best answer. This pre-training setting is the same as the model for choosing one answer in ad-

Model	Acc.	Prec.	Rec.	F1
Proposed	**0.441**	0.723	**0.579**	**0.643**
+Q_BERT-imp	0.440	**0.734**	0.569	0.641
+BA_pretrain	0.411	0.697	0.550	0.615
BA	0.408	0.704	0.522	0.600

Table 6: Evaluation results used in the analysis. These scores were measured with the development data.

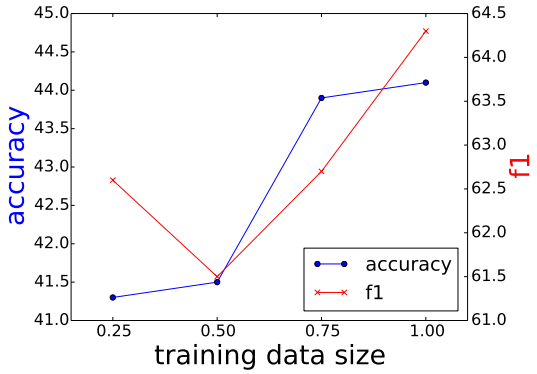

Figure 3: Correlation between the training data size and the metrics.

Q	I would like to plant a lawn. What is the easiest lawn to care for and maintain? What is the best way to plant lawns?			
	answer	gold	Proposed	+Q_BERT-imp
A1	The easiest lawn to maintain and care for is manila grass that is easily purchasable. The key to planting is to level the ground so that you can use an electric lawnmower.			
A2	This isn't the answer, but you should know that managing your lawn can be a daunting task.			
A3	...An easy one is manila grass, which is often sold at home improvement stores...To plant it, level the substrate and add fertilizer to the subsoil beforehand...	✓	✓	✓
A4	**Two types of turf are usually available on the market...The most important part of planting is to loosen the soil to some degree....Fertilizer is not necessary if the grass grows there...Lightly sprinkling soil over the surface of the turf at planting will help it to root better. Water should be sprayed thoroughly...Manila grass is the easiest to care for. Zoysia japonica is the best looking.**	✓	✓	
A5	Take care of any lawn to prevent it from being invaded by white clover.			

Table 7: Example of different outputs between **Proposed** and **+Q_BERT-imp**.

vance used in **ranking** in Sec 4.2. We ran the model which uses an additionally pre-trained BERT-imp (**+BA_pretrain**).

The bottom-half of Table 6 shows the results. **+BA_pretrain** had lower accuracy and F1 than **Proposed**, which demonstrates the pre-training in the task of choosing the best answer prevented a prediction of answer importance.

To test the correlation between the best answer and the gold answer set, we measured the scores in the setting of choosing only the best answer (**BA**). The percentage of the best answer in the gold answer set is 0.704, so the best answer might not be included in the gold answer set. Since the best answer is chosen by the questioner alone, it is prone to biases and does not correlate strongly with the gold answer set. The pre-training makes the model choose the answers that are most likely to be considered the best answers, and this may have led to a decrease in performance.

Table 8 shows an example of different outputs between **Proposed** and **+BA_pretrain**. A3 is the only answer in the gold answer set to the question. In the answers to a question about which doctor to see if the questioner has cystitis, A3 answers that it doesn't matter whether questioner sees a physician or a gynecologist. Although the gold answer set to this question only contains A3, **+BA_pretrain** chose A4, which contains only information that it is OK to see a gynecologist and does not include information that it is no problem to see an internist. Thus, the best answer tends to be sympathetic or positive to the questioner, so it is not always the most informative one.

Varying the size of training data We investigated how the performance changes when the training data size varies. Figure 3 shows the performance of **Proposed** when the training data size was set to 0.25, 0.5, or 0.75 in the development set. The performance tended to be improved as the corpus size increased, and did not saturate. Therefore, increasing the training data size is expected to improve performance.

Q	I seem to have cystitis. I need to go to the hospital; should I see a urologist? Or should I see a gynecologist?			
	answer	gold	Proposed	+BA_pretrain
A1	Go to an internal medicine clinic. It's better to see a specialist afterwards.			
A2	...It doesn't matter which one you go to.			
A3	You can see a physician as well. If you have a family gynecologist, that's fine too. A gynecologist treats all kinds of women's diseases...	✓	✓	
A4	**I recently had cystitis too. That was painful. The gynecologist's questionnaire had a "cystitis" section on it, and that's when I first realized that gynecologists generally treat cystitis...**			✓
A5	Talk to the hospital...			

Table 8: Example of different outputs between **Proposed** and **+BA_pretrain**.

Q	Many black bugs stuck to the underside of the leaves of a potted lvy. If possible, I would like to get rid of them without using insecticides. If there is a good method, please let me know.		
	answer	gold	output
A1	If it's an aphid, spray it with milk. If it's a spider mite, pour water on the back of the leaves and you'll drown it.		
A2	I think it would be a good idea to light some mosquito coils...		
A3	I would use wood vinegar...	✓	✓
A4	Why do you not want to use pesticides? Pesticides pose no danger when used correctly!...		
A5	**...1) Spray with milk. 2) Strain and spray crushed garlic and pepper powder boiled down with soju. 3) Soak cigarette butts in water to extract the nicotine content, and spray. It's not a plant to eat, and wouldn't it be better to spray a pesticide to get rid of them in one shot?...**	✓	✓

Table 9: Example of correct output.

Q	It is often said that as we age, our skin stops repelling water, but is it true? Also, if possible, please tell me the cause.		
	answer	gold	output
A1	In my case, my towel got wet a lot more often...		
A2	Because you lose its elasticity, you can't get polka dots on your skin...		
A3	I think it's a loss of elasticity, or in other words, a loss of facial muscles.		
A4	**It's true....The oil that my body produces on the surface of my skin becomes less and less secreted as I get older, so it stops repelling water...In the case of the face, it becomes less oily, loses its elasticity...**	✓	✓
A5	...I've heard that on skin that has lost tension etc., the water droplets are not spherical, but instead is spread out in gooey droplets.·····		✓

Table 10: Example of incorrect output.

4.5 Case Study

We analyze some of the outputs of our method. Table 9 shows an example of a correct output. The method chose A3 and A5. Both A3 and A5 contain the information about how to remove black bugs from the leaves, and both are included in the gold answer set. A5 does not contain the information in A3 about using a wood vinegar, and A3 does not contain the information in A5 about spraying milk. Therefore, the chosen answer set is non-redundant. The contents of A1 and A4 were included in A5, and our method was able to extract the various information from the answers.

Table 10 shows an example of an incorrect output. A4 is the most informative answer and includes the content of A5. However, our method has chosen both A4 and A5. As shown in this example, our method sometimes chose redundant answers, which may indicate insufficient learning of the similarity. In the future, its performance could be improved by creating a dataset with similar answers and pre-training BERT-sim to estimate the similarity between answers.

5 Conclusion

We proposed a new task to select a diverse and non-redundant answer set in CQA. We created a dataset for this task, and proposed a model for it by using DPPs and BERT. Our method performed better than several baseline methods. In future work, we aim to pre-train BERT-sim to calculate the similarity between answers and apply our method to CQA platforms in other languages.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: Factoid question answering over social media. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 467–476, New York, NY, USA. Association for Computing Machinery.

Florian Boudin, Marc El-Bèze, and Juan-Manuel Torres-Moreno. 2008. A scalable MMR approach to sentence scoring for multi-document update summarization. In *Coling 2008: Companion volume: Posters*, pages 23–26, Manchester, UK, August. Coling 2008 Organizing Committee.

Xin Cao, Gao Cong, Bin Cui, and Christian S. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 201–210, New York, NY, USA. Association for Computing Machinery.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS '18, page 5627–5638, Red Hook, NY, USA. Curran Associates Inc.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019a. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy, July. Association for Computational Linguistics.

Sangwoo Cho, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2019b. Multi-document summarization with determinantal point processes and contextualized representations. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 98–103, Hong Kong, China, November. Association for Computational Linguistics.

Tanya Chowdhury and Tanmoy Chakraborty. 2019. Cqasumm: Building references for community question answering summarization corpora. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '19, page 18–26, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

H. Dong, J. Wang, H. Lin, B. Xu, and Z. Yang. 2015. Predicting best answerers for new questions: An approach leveraging distributed representations of words in community question answering. In *2015 Ninth International Conference on Frontier of Computer Science and Technology*, pages 13–18.

Gusfield Erkan. 1997. Algorithms on strings, trees and sequences.

Shengbo Guo and Scott Sanner. 2010. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, page 833–834, New York, NY, USA. Association for Computing Machinery.

Shahar Harel, Sefi Albo, Eugene Agichtein, and Kira Radinsky. 2019. Learning novelty-aware ranking of answers to complex questions. In *Proceedings of The World Wide Web Conference*, WWW '19, page 2799–2805, New York, NY, USA. Association for Computing Machinery.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November.

Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. 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, Marseille, France, May. European Language Resources Association.

Odile Macchi. 1975. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122.

Zelda Mariet, Yaniv Ovadia, and Jasper Snoek. 2019. DPPNet: Approximating determinantal point processes with deep networks. In *NeurIPS*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS '13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 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, Denver, Colorado, June. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545, San Diego, California, June. Association for Computational Linguistics.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada, August. Association for Computational Linguistics.

Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. 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*, SIGIR '16, page 215–224, New York, NY, USA. Association for Computing Machinery.

Roi Reichart and Anna Korhonen. 2013. Improved lexical acquisition through DPP-based verb clustering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 862–872, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hongya Song, Zhaochun Ren, Shangsong Liang, Piji Li, Jun Ma, and Maarten de Rijke. 2017. Summarizing answers in non-factoid community question-answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 405–414, New York, NY, USA. Association for Computing Machinery.

Idan Szpektor, Yoelle Maarek, and Dan Pelleg. 2013. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. pages 1249–1260, 05.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual lstm architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 695–704, New York, NY, USA. Association for Computing Machinery.

Qiongjie Tian, Peng Zhang, and Baoxin Li. 2013. Towards predicting the best answers in community-based question-answering services. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press.

Arseny Tolmachev and Sadao Kurohashi. 2017. Automatic extraction of high-quality example sentences for word learning using a determinantal point process. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–142, Copenhagen, Denmark, September. Association for Computational Linguistics.

Xin-Jing Wang, Xudong Tu, Dan Feng, and Lei Zhang. 2009. Ranking community answers by modeling question-answer relationships via analogical reasoning. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 179–186, New York, NY, USA. Association for Computing Machinery.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2165–2173, New York, NY, USA. Association for Computing Machinery.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 475–482, New York, NY, USA. Association for Computing Machinery.

Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. 2017. Attentive interactive neural networks for answer selection in community question answering. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3525–3531. AAAI Press.

Zhou Zhao, Hanqing Lu, Vincent W. Zheng, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Community-based question answering via asymmetric multi-faceted ranking network learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3532–3538. AAAI Press.