INCORPORATE USER REPRESENTATION FOR PERSONAL QUESTION ANSWER SELECTION USING SIAMESE NETWORK

Zihao Qi, Dario Bertero, Ian Wood, Pascale Fung

Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
zqiab@connect.ust.hk, dbertero@connect.ust.hk, eeianwood@ust.hk, pascale@ece.ust.hk

ABSTRACT

Many natural language questions are inherently subjective. They can not be answered properly if we do not know the personal preferences of the answerer. For example, "Do you like cats?" There is no "the only correct answer" to this question. To answer it, the model has to be able to capture the persona of the answerers. However, the users usually do not answer different questions with equal chance. Instead, while some are answered with a high frequency, others are hardly answered by anyone. To deal with this imbalanced sparsity in data, we first introduce a Siamese Network to capture the preferences patterns of the users. Then the model is ensembled with an additional dense layer to predict the answers of the users. Applying to an online dating dataset, our approach achieves a high accuracy of 78.7%.

Index Terms— Personal Question Answer Selection, Siamese Network, User Representation, Preference Learning

1. INTRODUCTION

Machines have already achieved human or near-human results in multiple semantic natural language task, including question answering [1]. However, in such tasks, a possible subject who answers the question is missing. They implicitly do not take into account the possibly subjective nature of answers. Most questions we meet in every daily life are inherently subjective and personal. They could not be solved properly if without understanding of the personality of the answerer.

Generally, this problem would be approached as recommender system or matrix completion problem by taking the user-question data as a matrix. A number of approaches to this sort of problem have been proposed, such as item-based/user-based top-N recommendations [2, 3], SVD [4, 5]. Those methods focus only on the data matrix. They do not care about what content the questions really ask about, resulting in a waste of the natural language information.

This work is partially funded by ITS/319/16FP of the Innovation Technology Commission, HKUST 16214415 & 16248016 of Hong Kong Research Grants Council, and RDC 1718050-0 of EMOS.AI.

Another related area is Question Answering (QA). QA is an incredibly broad topic. Almost every natural language processing task could be somehow fitted into this setup. Generally, QA tasks can be divided into factoid QA and non-factoid QA [6]. Usually they are solved by either information retrieval approaches or first building a semantic represenation and then mapping it into a query of structured data. These approaches, however, only take into account the content of the question and fixed external knowledge, and do not incorporate personal information about the answerer of the question and how that may effect the appropriateness of an answer.

A common challenge that will be faced in this task is the un-uniform distribution of data samples among categories. Some questions are answered with a very high frequency, while others are hardly answered by anyone. This leads to poor performance for categories that have too few training samples.

In this work, we propose a framework and a synthetic task for personal question answer selection combining semantic encoding of question texts with a user embedding designed to capture how user similarities map to similar preferences. We address the challenge of un-uniformly distributed data by proposing to pre-train a Siamese structure. We will introduce what Siamese is in Section 2 and explain our model structure in Section 3. The performance is then tested on a novel dataset in Section 4. The dataset contains about 12K users and 2.4K personal questions that focus on topics varying from daily habits to political views. Finally, the results are discussed and analyzed in Section 5.

2. METHODOLOGY

A Siamese Network is a twin network consisting of two identical halves that share the same parameters [7, 8]. It takes a pair of inputs then computes the distance between the pair of outputs. If they are labeled similarly, their distance in the output vector space will be made a bit closer. They separated further if the labels are dissimlar. Siamese Networks are usually used to learn a similarity metric. It is also useful a

tool for classification where the number of data categories is huge, and where the number of training samples for a single category is rather small [9]. In this work, we construct our Siamese Network with bi-LSTM. The general structure of the network could be viewed in Figure 1.

Contrastive loss [10] is used to distinguish between similar and dissimlar pairs of samples. Let the two samples be denoted as $\mathbf{X_1}$ and $\mathbf{X_2}$, our network denoted as a function $G_W(\mathbf{X})$. Here $\mathbf{W}$ is the network parameters. The distance between the output embedding of the samples, written as:

$$D_W(\mathbf{X_1}, \mathbf{X_2}) = ||G_W(\mathbf{X_1}) - G_W(\mathbf{X_2})||_2$$

Once the distance between the embeddings is obtained, it is given as inputs to the loss objective function, which can be formally written as a function of the input samples and network parameters:

$$\begin{aligned} L(\mathbf{W}) = &\sum_{i=1}^{P}(1 - Y^i)\frac{1}{2}(D_W(\mathbf{X_1}, \mathbf{X_2})^i)^2 \\ &+ \sum_{i=1}^{P} Y^i \frac{1}{2}(max(0, m - D_W(\mathbf{X_1}, \mathbf{X_2})^i))^2 \end{aligned} \tag{1}$$

Where $m > 0$ denotes a manually set margin which acts as a boundary. It keeps dissimilar pairs separated from each other by a distance defined by m. $(\mathbf{X_1}, \mathbf{X_2}, Y)^i$ denotes the i^{th} training sample, where $Y^i = 0$ if $\mathbf{X_1}$ and $\mathbf{X_2}$ are deemed similar and $Y^i = 1$ if they are deemed dissimilar.

3. MODEL ARCHITECTURE

Siamese network contains two identical halves sharing the same training parameters. Each half we set it as a combination of two modules: User Module responsible for learning user preference, and Question Module responsible for understanding the question. User Module takes user ID token as input, which is a integer number. It passes an embedding layer to obtain user embedding, whose dimension is 64. Question Module takes question sentence vector as input. Each slot is a word token. It first passes through a pre-trained word embedding layer [11] with the dimension as 100. Then a bi-LSTM and max-pooling are employed in order to get understanding of the question. After this, the output vectors of Question Module and User Module each pass through a fully connected layer with output dimension as 64, and then are concatenated together. The output passes a final fully connected layer with the output dimension as 100. We regard this final output vector as "user-question representation". It captures the opinion of this user toward the question. Contrastive loss is used to train the Siamese to distinguish between similar and dissimilar pairs.

After finely training the Siamese Network, we then build the QA model by adding an additional Choice Module to it. It takes choice sentence vector as input. Pass through a similar network as Question Module. Then get concatenated with "user-question representation". An extra fully connected layer with sigmoid activation is then employed to get a "score" of how likely this user is going to pick this choice. The choice with the highest "score" among all candidates is picked as the prediction of the QA model.

4. EXPERIMENTS

4.1. Datasets

We test and compare our proposed model on a novel dataset collected from an online dating website [1]. This dataset consists of the answers of 14,493 users to 7,706 multiple-choice questions. The IDs are anonymized for privacy concern. On average, each user approximately answers a few hundred of the questions, leaving many others blank. To keep the results of the experiments more stable, we discard the users who have answered less than 10 questions and the questions that are answered by less than 2 users. After preprocessing, 13,628 users and 2,421 questions are left. The statistics of the frequency of being answered for each question is shown in Figure2.

Question topic.	ratio
Dating preference	21.4%
Living style	16.5%
Personalities	15.7%
World view	13.2%
Sex openness	10.7%
Habits	6.6%
Politics	5.8%
Drug attitude	4.1%
Alcohol & smoke	3.3%
Religion	3.3%
Racial	1.7%
Education	1.7%
Others (hard to define)	**19.8%**

Table 1. Statistics of the question topics, made from LDA and human analysis. Topics may overlap each other.

The statistics of the question topics are also provided in Table 2 for reference. To be noted that it is only a rough estimation made from LDA topic detection (Latent Dirichlet Allocation) [12] results and human analysis. Besides, the ratios do not necessarily have to sum up to 1 since the topics may overlap each other.

[1] available upper research request

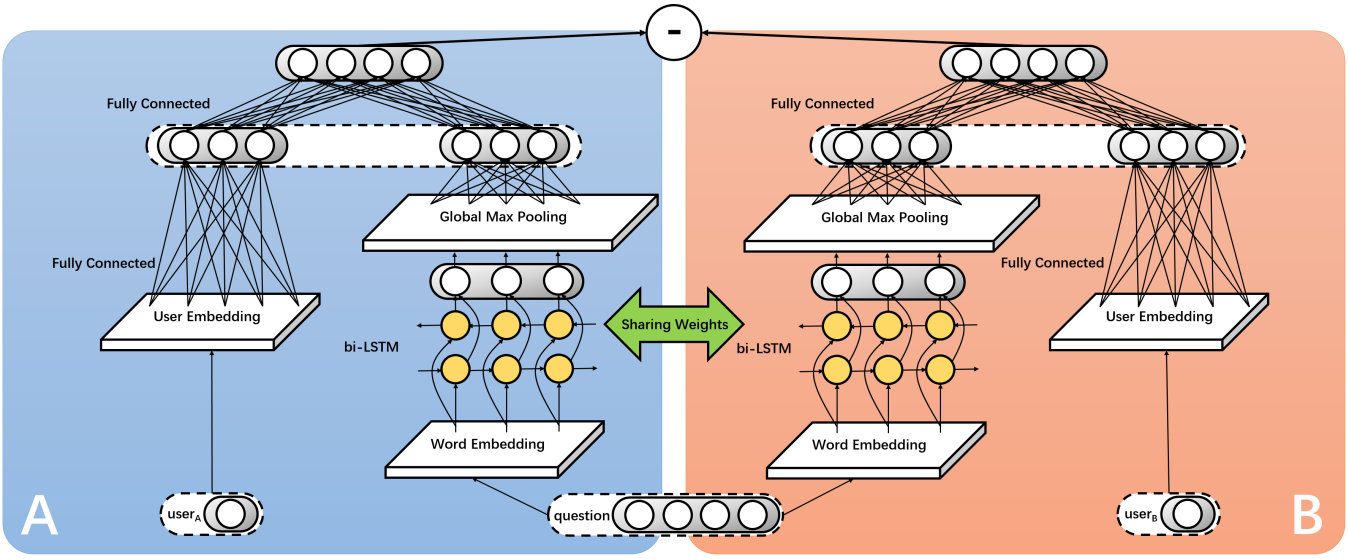

Fig. 1. Overall picture of the Siamese Network. The subtraction is then passed to contrastive loss function.

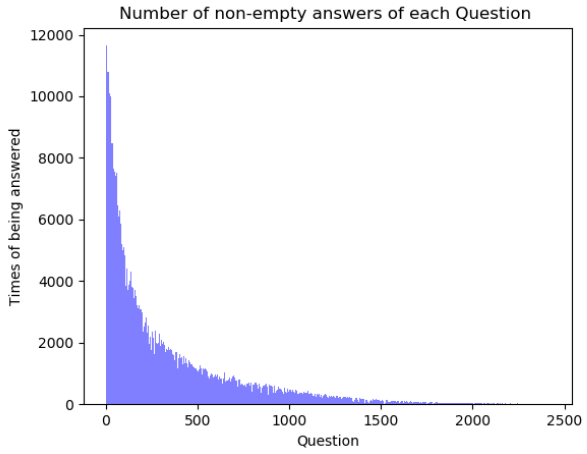

Fig. 2. The frequency of being answered for each of the 2,421 questions, in descending order. It can be seen that the distribution is heavily nonuniform.

4.2. Setup

We rewrite the dataset into the format of separated $(user, question, choice)$ tuples, which totalled 742,599 tuples. They are then randomly partitioned into training set (70%, 519,818 tuples), validation set (15%, 111,391 tuples), and testing set (15%, 111,390 tuples). The performance is evaluated by prediction accuracy and F1-micro score.

4.3. Results

We compared our method with several baselines. The setup and configuration of the baselines are briefly introduced as following.

1. ***Most Common*** In this case, we will always pick the most-commonly picked choice as the prediction. It could be an interesting baseline because many of the questions are heavily biased. For example, 87.2% users choose **"No"** in the question **"Is astrological sign at all important in a match?"**.

2. ***Singular Value Thresholding*** To test this baseline [13], the dataset needs to be rewritten into 0-1 user-choice matrix. Each slot in the matrix represents whether a choice of a question has been picked by a user.

3. ***Collaborative Filtering*** It is straight forward to apply CF once we realize that answers to questions are just like ratings to movies. Each time given a user, we find the closest $K = 32$ users to this given user by Pearson correlation similarity [14]. Their corresponding answer histories are aggregated to give the prediction.

   Neural Matrix Facotorization (NeuMF) [15] proposes a deep learning structure for collaborative filtering. We follow a similar structure with 3 hidden layers.

4. ***Memory Network*** We follow the similar setting as the factoid question answering research [16]. The transformation from their task to ours is straight forward by taking all answer history of a user as the "story". The output layer includes all choices for all questions, but the prediction is picked only among the possible candidates.

We discover that our method yields better results than all other baselines. We also try to prove the utility of the Siamese structure and pre-trained word embeddings by adding two

	Acc.
Most-Common	61.8%
Singular Value Thresholding	42.3%
Collaborative Filtering	71.3%
NeuMF	72.9%
Memory Networks	62.3%
ours (Siamese with pre-trained wordEmb)	**78.7%**
ours (no Siamese)	73.1%
ours (no pre-trained wordEmb)	75.5%

Table 2. Test accuracy on our novel dataset for various methods.

comparison experiments. It can be see that the classifier performs about 5% worse if without Siamese network learning the embeddings of users. Also, it performs 3% worse if without plugged in with pre-trained word embeddings. This fact supports our hypothesis that natural language knowledge and preference patterns knowledge could help the prediction on personal question answering task.

5. DISCUSSION

We check which question is doing the best at distinguishing the users. We do it in three steps: Firstly, for each question, we separate the users into groups according to their answers. Secondly, we compute their inter-group distance. Lastly, we sort the questions by their averaged inter-group distance. The question with larger averaged inter-group distance are considered to be better at distinguishing users. For a question with group $A, B, C...$ and u denoting the users, we define inter-group distance for any two groups A, B by:

$$D_{A,B} = \frac{1}{|A||B|} \sum_{u_i \in A, u_j \in B} D(u_i, u_j) \tag{2}$$

Where $D(u_i, u_j)$ is the distance between user u_i and u_j. Here we use cosine distance. It is found that the question "Is interracial marriage a bad idea?" separates the user embeddings the best in our model. To validate whether our network really helps to extract the preferences of the users, we use t-SNE to reduce the dimensions of the user embeddings and plot them out in 3D coordinates. Every user is labeled by their answer to the question. We then randomly pick a user, check its neighbors. If most of the neighbors pick the same choice as their central does, our model is convinced to be able to extract perferences.

In Figure 3, we randomly pick a user who answers "Yes" to the question "Is interracial marriage a bad idea?", and another user who answers "No" to this question. We highlight their closest 50 users found by cosine distance. It can be seen that most of the neighbors pick the same choice as their central user. It not only shows that our model is capable of capturing this preference but also shows that people's opinion

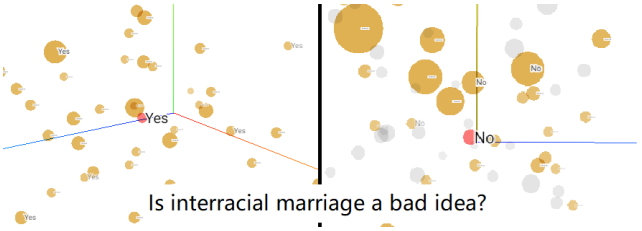

Fig. 3. The closest 50 users to the given user. Most of the neighbors pick the same choice as their central.

towards interracial marriage does make a distinction between users in our dataset.

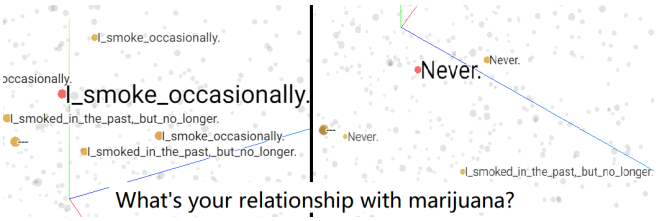

Fig. 4. User's opinion toward drugs like marijuana also make a distinction in our dataset.

In Figure 4, we check whether the opinion toward light drugs like marijuana will make a separation between the users. This question is picked carefully and manually. By prior knowledge, we know that people's tolerance of marijuana varies a lot with country, region and culture differences. Thus we expect to see strong distinguishing in this question. From our observation, "marijuana" does make a separation between the users. Another interesting thing is that "I smoked in the past but no longer" turn out to be a middle state between "Never" and "I smoke occasionally" in this picture. It is showing around both "Never" and "Occasionally". This is coordinated with the literal logic. It suggests that our model is implicitly learning the meaning behind the questions.

6. CONCLUSION

Our main contribution of this paper is to propose a framework and a synthetic task for personal question answer selection. We point out that a usual challenge faced in this task is the imbalanced sparsity among questions. We address this challenge by applying Siamese networks to pre-train the user embeddings, which greatly overcomes the lack of training samples. We show in the experiments that our model yields better results comparing to straightforward baselines such as Collaborative Filtering. We also prove that our model is capable of capturing the preference patterns. It acquires user embeddings that show meaningful distributions.

7. REFERENCES

[1] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al., "Building watson: An overview of the deepqa project," *AI magazine*, vol. 31, no. 3, pp. 59–79, 2010.

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 285–295.

[3] Zhi-Dan Zhao and Ming-Sheng Shang, "User-based collaborative-filtering recommendation algorithms on hadoop," in *Knowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on*. IEEE, 2010, pp. 478–481.

[4] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391, 1990.

[5] Yehuda Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 426–434.

[6] Lynette Hirschman and Robert Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, vol. 7, no. 4, pp. 275–300, 2001.

[7] Franck Leclerc and Rejean Plamondon, "Automatic signature verification: The state of the art1989–1993," *International journal of pattern recognition and artificial intelligence*, vol. 8, no. 03, pp. 643–660, 1994.

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, 2005, vol. 1, pp. 539–546.

[9] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE, 2006, vol. 2, pp. 1735–1742.

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[13] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[14] Yehuda Koren and Robert Bell, "Advances in collaborative filtering," in *Recommender systems handbook*, pp. 77–118. Springer, 2015.

[15] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua, "Neural collaborative filtering," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 173–182.

[16] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov, "Towards ai-complete question answering: A set of prerequisite toy tasks," *arXiv preprint arXiv:1502.05698*, 2015.