FAQ Retrieval Using Attentive Matching

Sparsh Gupta
spg005@ucsd.edu
Intuit AI
San Diego, California

Vitor Carvalho
vitor_carvalho@intuit.com
Intuit AI
San Diego, California

ABSTRACT

The task of ranking question-answer pairs in response to an input query, aka FAQ (Frequently Asked Question) Retrieval, has traditionally been focused mostly on extracting relevance signals between query and questions based on extensive manual feature engineering. In this paper we propose multiple deep learning architectures designed for FAQ Retrieval that eliminate the need for feature engineering and are able to elegantly combine both query-question and query-answer similarities. We present experimental results showing that models that effectively combine both query-question and query-answer representations using *attention mechanisms* in a hierarchical manner yield the best results from all proposed models. We further verify the effectiveness of attention mechanisms for FAQ Retrieval by conducting experiments on a completely different attention-based architecture, originally designed for question duplicate detection tasks, and observing equally impressive experimental ranking results.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

attention mechanism, neural networks, learning to rank

ACM Reference Format:
Sparsh Gupta and Vitor Carvalho. 2019. FAQ Retrieval Using Attentive Matching. In *SIGIR '19: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The ability to effectively rank question-answer pairs after an input of a query, aka FAQ retrieval, is a fundamental feature to all FAQ forums. Search is particularly important to large FAQ forums, where a good user experience can be often linked to good search capabilities.

Most of the early work on FAQ retrieval has relied on traditional feature engineering for surfacing similarities between query and questions. Many of these ideas used language parsing to discover semantic and/or syntactic structures that could better represent query-question pairs, or to convert the pairs into a common edit distance, template or structured representation [5, 7, 9]. For instance, Kothari *et al* [6] described FAQ retrieval using features such as Longest Common Subsequence or Edit Distance to compute similarities between query and questions. Overall, the vast majority of these approaches focused on finding similarity signals between queries and questions only, disregarding possible matching signals between query and answer [5, 7–10].

To the best of our knowledge, only a few traditional feature-engineered proposals actually utilized query-answer similarity signals for FAQ retrieval. One of these rare examples is the work of Jijkoun *et al* [4], that developed a system that first crawled the web for all FAQs that matched with the input query and then ranked those FAQs. The ranking model used features like vector space similarities between the query and the question, the answer and title of the FAQ page using vector space model in Lucene [1]. In this paper we draw inspiration from Jijkoun *et al* [4] for using both query-question and query-answer similarities in FAQ retrieval, but we do so not via expensive feature engineering but by learning query, question and answer representations directly from data through deep learning architectures.

Deep Learning modeling has enjoyed significant success recently, including on some types of Question Answering (QA) tasks where it has shown to outperform traditional feature engineering solutions [11, 12, 15]. For instance, Deep Matching Networks [14] and Multihop Attention Networks [12] have outperformed traditional baselines in various QA ranking datasets, with Multihop Attention Networks exhibiting state-of-the-art results. While these deep architectures have no need for feature engineering and present excellent predictive performance, to the best of our knowledge none of the proposed models have attempted to directly incorporate both query-question and query-answer signals in the same architecture.

In this paper we extend the aforementioned ideas by proposing deep learning architectures specifically designed for FAQ retrieval. These new model variants contrast with previous work by explicitly modeling both query-question and query-answer signals. Our experimental results show that models incorporating both signals tend to outperform models that use only query-question or query-answer information to rank QA pairs. Furthermore, we also found that models that aggregate question and answer information using attention mechanisms in a hierarchical manner outperform their counterparts utilizing other aggregation methods.

We further verify the effectiveness of attention mechanisms in FAQ retrieval by modifying the Bilateral Multi-Perspective Matching model (originally proposed by Wang *et al* [13] for matching sentences) to aggregate both question and answer representations with hierarchical attention blocks, and also obtaining impressive FAQ retrieval results.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Briefly stated, our main contributions are as follows. We propose different deep learning architectures specifically designed for FAQ Retrieval that are able to effectively combine both query-question and query-answer similarities, and present excellent retrieval results. We also present evidence that attention mechanisms are extremely powerful in aggregating query-question and query-answer similarities for FAQ retrieval, clearly outperforming other aggregation methods.

2 MODEL SPECIFICATIONS

First we present two recently proposed models for question answer ranking that we used for our experiments, followed by the modifications we made to account for FAQ retrieval task.

2.1 Deep Matching Network (DMN)

Yang *et al* introduced an interaction based model called Deep Matching Net [14]. This model was originally meant for multi-turn conversations, but we adapted a version of it for single turn conversation which can also work as scorer for ranking tasks. This model takes a 2 channel image as input to convolutional layers, followed by fully connected layers that generate a matching score. This 2 channel image is made of 2 matrices M_1 and M_2. M_1 is formed by taking the dot product of embeddings of every word of question with every word of answer. M_2 is also formed in a similar way by taking the dot products of hidden representations words of question and answer after passing them through a bi-directional GRU.

2.2 Multihop Attention Network (MAN)

Tran *et al* [12] proposed a model that has a bi-directional LSTM layer that takes the question and answer as input to generate representations for all words of the question and answer. Multiple layers or multiple "hops" of attention are applied on these to get attended representations of question and answer at each hop. Each layer of attention places focus on different parts of the answer and question to get a matching score. The model is thus able to compare question and answer from different multiple perspectives. At each hop, cosine similarity between the question and answer is computed.

The cosine similarities between question and answer at each hop are summed to compute the final matching score (Equation 1). Here $o_q^{(k)}$ and $o_a^{(k)}$ refer to the question and answer representations after the kth hop in the network.

$$sim(q, a) = \sum_k cos\left(o_q^{(k)}, o_a^{(k)}\right) \tag{1}$$

Hinge loss (Equation 2) with L2 regularization is used to train the network.

$$L = max\{0, M - sim(q, a_+) + sim(q, a_-)\} \tag{2}$$

where M is the margin.

2.3 Baseline Aggregation Strategies

Concatenate Question-Answer Text - The idea is to concatenate the question and answer text from FAQ and use a QA ranking model to rank the FAQs. In this case, the Q for QA ranking is the input query and A is the concatenated text.

Query-Answer Matching - This baseline also uses the QA ranking model to rank FAQs by comparing only the query and the answer.

Query-Question Matching - This baseline compares query only to the question using a QA ranking model.

2.4 Aggregation Strategies

For DMN, we get the features from the convolutional block for query-question interaction and query-answer interaction by passing the query-question pair and query-answer pair through it in Siamese fashion. Let these representations be o_Q and o_A. We aggregate these representations to get the combined representation o_{comb} which is further passed to fully connected layers to generate a matching score (see Figure 1).

For MAN, given that o_q, o_Q and o_A are the representations of query, question and answer respectively during any stage in forward pass, we use o_q as it is and combine o_Q and o_A to generate the representation o_{comb}. We then generate matching scores at each hop by computing cosine similarity between o_q and o_{comb} and summing them (see Figure 2).

We combine o_Q and o_A using two different methods to get o_{comb}.

Interpolation: We add o_Q and o_A in a weighted manner characterized by the weight λ.

$$o_{comb} = \lambda o_Q + (1 - \lambda) o_A \tag{3}$$

Attention: We use an attention mechanism to combine o_Q and o_A. This allows the network to dynamically decide whether the question is more important or the answer, separately in every dimension of representation.

$$o_{comb} = tanh(W_q o_Q + W_a o_A) \tag{4}$$

Figures 1 and 2 show the use of these aggregation strategies in Deep Matching Network (DMN) and Multihop Attention Network (MAN), respectively.

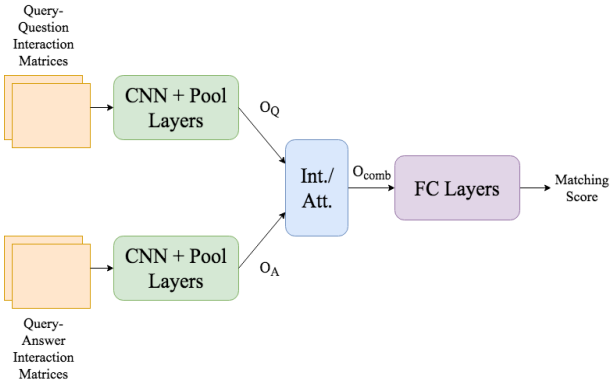

Figure 1: Aggregation of question and answer representations used in DMN; Int./Att. block shows aggregation of question and answer representation using interpolation/attention mechanisms

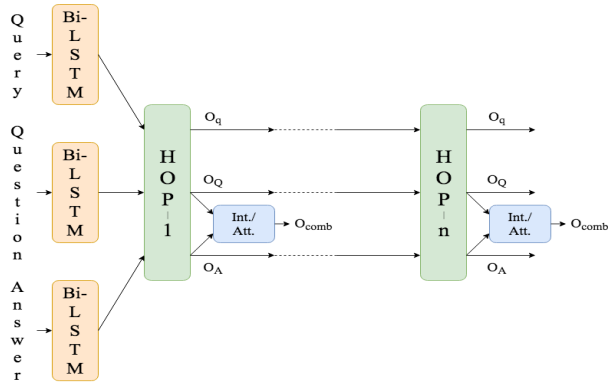

Figure 2: Aggregation of question and answer representations used in MAN; Int./Att. block shows aggregation of question and answer representation using interpolation/attention mechanisms

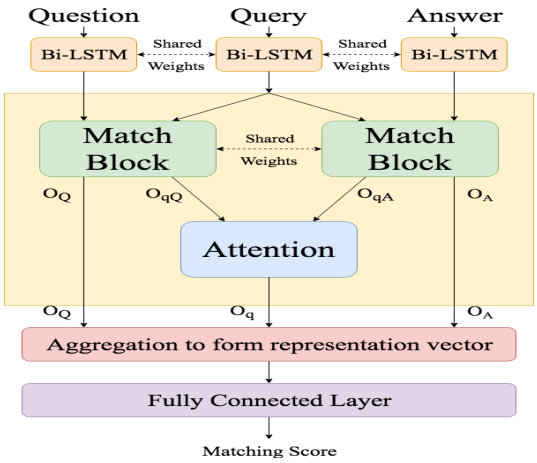

Figure 3: Symmetric Bilateral Multi-Perspective Matching Block (SymBiMPM Block); o_Q: question repr., o_A: answer repr., o_{qQ}: query repr. after attending to question, o_{qA}: query repr. after attending to answer, o_q: final query repr.

2.5 SymBiMPM

The **Sym**metric **Bi**lateral **M**ulti-**P**erspective **M**atching (SymBiMPM) model is inspired from the Bilateral Multi-Perspective Matching model [13] for question answer ranking. This model uses a **multi-perspective matching block** to compare two sequences and generate the matched representations for both these sequences. This block has four different matching mechanisms that are used on the input sequences. Matching is applied in both the directions, i.e. if P and Q are the two inputs, then the output is a matched representation of P obtained by attending to Q, and a matched representation of Q obtained by attending to P.

We use the multi perspective matching block in a symmetric fashion for query-question and query-answer matching followed by attention layer and fully connected layers to get the final matching score. The architecture for Symmetric Bilateral Multi-Perspective Matching model has been shown in Figure 3. A multi-perspective match block is first used to generate attended representations of query-question. The same match block is used to generate matched representations of query-answer. This step results in one representation each for question and answer, and two representations for the query. We combine these two representations using an attention mechanism given by Equation 5.

$$o_q = tanh(W_q o_{qQ} + W_a o_{qA}) \tag{5}$$

where W_q and W_a are attention matrices used. We use the final time step for each sequence in each direction to form a representation vector having matching and attended information from all the three sequences, similar to the way Wang *et al* [13] aggregate the outputs of the matching block. This representation vector is passed though a multilayer perceptron to get matching score between query and the FAQ.

3 EXPERIMENTS

3.1 Datasets

We used two datasets for experiments - SemEval CQA Task 3 and Tax Domain QA, which we plan to make public soon. These dataset specifications are given in Table 1.

Table 1: Size specifications for all datasets.

Dataset	SemEval	Tax Domain QA
# of data points (train/dev/test)	4k / 1k / 1k	65k / 20k / 20k
Avg length of queries	6	4
Avg length of questions	49	13
Avg length of answers	275	110

3.1.1 SemEval CQA Task 3. This dataset[1] was intended for community question answering (CQA) originally, but the task 3 data had the QA pairs grouped by search query terms, which facilitated the transformation of this data into FAQ Retrieval format where FAQs are ranked for a query and are awarded ranks for Perfect Match, Relevant and Irrelevant. A pairwise labelling strategy is used to convert the labels into binary labels, and hinge loss (Equation 2) is used to train the models. The maximum token length used in all experiments for query, question and answer is 9, 70 and 350 respectively.

3.1.2 Tax Domain QA. This dataset is collected from a popular tax domain FAQ retrieval platform. The dataset ranks the FAQs in response to a query from 1 to 5, 1 being the least relevant and 5 denoting the most relevant FAQ. We use pairwise-labelling strategy to convert this data as well to binary label format. We used a maximum of 6, 20 and 180 tokens for query, question and answer respectively.

3.2 Metrics

We use Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG@5) as metrics to evaluate the baselines and all the models. For both these metrics, 95% bootstrap confidence intervals [2] have been computed by randomly sampling roughly

[1] http://alt.qcri.org/semeval2017/task3/

1/3rd of the test data results 30 times with replacement, as suggested by Hogg and Tanis [3].

4 RESULTS AND CONCLUSIONS

The results for all the models on the two datasets are tabulated in Tables 2 and 3. Clear trends can be observed from these results, which are similar for both datasets. It can be observed that most of the MAN variants perform better than DMN variants, which is not surprising since MANs have recently shown state-of-the-art results for question answering [12].

For both DMN and MAN models, comparing the query only to the question yields better results than comparing the query only to the answer, showing that question text has information that is more relevant than answer text for the FAQ ranking task. It can further be observed that comparing query text to just the question text is better than comparing query text to concatenated text of question and answer ($Concat(Q, A)$ in Tables 2 and 3). We believe this is mainly due to the fact that the concatenation of question and answer is typically very long, and adds more noise than signal to the modeling of these FAQ retrieval tasks.

On the other hand, aggregating the question and answer text representations in a non-trivial way (via $Interpolation(Q, A)$ or $Attention(Q, A)$) results in significantly better performance than the three baseline aggregation methods, thus showing that the use of both question and answer information can indeed be beneficial for FAQ retrieval. Furthermore, it can be observed for both DMN and MAN models that use of attention mechanism for aggregation ($Attention(Q, A)$) is significantly better than interpolation mechanisms ($Interpolation(Q, A)$), hence confirming our claim that attention allows the model to compare relative importance of question and answer separately for different features which in turn results in more model flexibility and better performance.

Note also that, while completely different than the other models, SymBiMPM (that uses a symmetrical attention architecture or Sym-Attention) displays results that are better than all other models and comparable to the best attention-based MAN models on both datasets[2]. This shows that effectiveness of attention-based mechanisms in learning good representations for query, questions, and answers on FAQ retrieval tasks.

Conclusions In this work we proposed multiple deep learning models for FAQ retrieval. We compared various possible aggregation methods to effectively represent query, question and answer information, and observed that answers in FAQs can provide valuable and beneficial information for retrieval models, if properly aggregated. We also observed that attention mechanisms are consistently the most effective way to aggregate FAQ inputs for ranking, with the best results in all our experiments.

Table 2: SemEval Results; 95% Bootstrap Confidence Interval for NDCG@5 is $\leq \pm 0.000225$ and for MRR is $\leq \pm 0.00047$; q: query, Q: question, A: answer

Model	Input Aggregation	NDCG@5	MRR
DMN	q | A	0.5829	0.5797
	q | Concat(Q, A)	0.6579	0.6467
	q | Q	0.674	0.6552
	q | Interpolation(Q, A)	0.6861	0.6728
	q | Attention(Q, A)	**0.6905**	**0.6984**
MAN	q | A	0.7013	0.6905
	q | Concat(Q, A)	0.7201	0.7088
	q | Q	0.7386	0.7336
	q | Interpolation(Q, A)	0.7544	0.7566
	q | Attention(Q, A)	**0.7619**	**0.7583**
SymBiMPM	Sym-Attention(q, Q, A)	**0.7617**	**0.758**

Table 3: Results on Tax Domain QA; 95% Bootstrap Confidence Interval for NDCG@5 is $\leq \pm 0.000218$ and for MRR is $\leq \pm 0.00035$

Model	Input Aggregation	NDCG@5	MRR
DMN	q | A	0.9026	0.7271
	q | Concat(Q, A)	0.9064	0.7277
	q | Q	0.9075	0.7289
	q | Interpolation(Q, A)	0.9094	0.732
	q | Attention(Q, A)	**0.9107**	**0.7375**
MAN	q | A	0.9097	0.735
	q | Concat(Q, A)	0.9071	0.7278
	q | Q	0.9131	0.7399
	q | Interpolation(Q, A)	0.9136	0.7451
	q | Attention(Q, A)	**0.9152**	**0.7462**
SymBiMPM	Sym-Attention(q, Q, A)	**0.9154**	**0.7472**

REFERENCES

[1] [n. d.]. Apache Lucene: A high-performance, full-featured text search engine library. http://lucene.apache.org

[2] B. Efron and R. Tibshirani. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* 1 (1986), 54–75.

[3] Robert V. Hogg and Elliot A. Tanis. 2005. *Probability and Statistical Inference, 7th Edition.* Pearson.

[4] Valentin Jijkoun and Maarten de Rijke. 2005. Retrieving Answers from Frequently Asked Questions Pages on the Web. In *CIKM*. ACM, 76–83.

[5] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, and B. Temelkuran. 2002. Omnibase: Uniform Access to Heterogenous Data for Question Answering. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB)*. 230–234.

[6] Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy, and L. Venkata Subramaniam. 2009. SMS Basen Interface for FAQ Retrieval. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*. ACL and AFNLP, 852–860.

[7] D. Molla, Rolf Schwitter, Fabio Rinaldi, James Dowdall, and Michael Hess. 2003. NLP for Answer Extraction in Technical Domains. In *EACL - NLP for QA Workshop*. 5–12.

[8] E. Sneiders. 1999. Automated FAQ Answering: Continued Experience with Shallow Language Understanding. In *Question Answering Systems. Papers from the 1999 AAAI Fall Symposium*. AAAI Press, 97–107.

[9] E. Sneiders. 2002. Automated Question Answering using Question Templates that Cover the Conceptual Model of the Database. In *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB)*. 235–239.

[10] W. Song, M. Feng, N. Gu, and L. Wenyin. 2007. Question Similarity Calculation for FAQ Answering. In *SKG*. IEEE, 298–301.

[11] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved Representation Learning for Question Answer Matching. In *ACL*. ACL, 464–473.

[2] Differences are not statistically significant under a paired t-test, with p-values > 0.05

[12] Nam Khanh Tran and Claudia Niederee. 2018. Multihop Attention Networks for Question Answer Matching. In *SIGIR*. ACM, 325–334.
[13] Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. In *IJCAI*. 4144–4150.
[14] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. In *SIGIR*. ACM, 245–254.
[15] Lei Yu, Karl M. Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep Learning for Answer Sentence Selection. In *NIPS Deep Learning Workshop*.