Using Sparse Coding for Answer Summarization in Non-Factoid Community Question-Answering

Zhaochun Ren[†*]
z.ren@uva.nl

Hongya Song[‡*]
hongya.song.sdu@gmail.com

Piji Li[§]
pjli@se.cuhk.edu.hk

Shangsong Liang[II]
shangsong.liang@ucl.ac.uk

Jun Ma[‡]
majun@sdu.edu.cn

Maarten de Rijke[†]
derijke@uva.nl

[†]University of Amsterdam, Amsterdam, The Netherlands
[§]The Chinese University of Hong Kong, Hong Kong, China
[II]University College London, London, United Kingdom
[‡]Shandong University, Jinan, China

ABSTRACT

We focus on the task of summarizing answers in community question-answering (CQA). While most previous work on answer summarization focuses on factoid question-answering, we focus on non-factoid question-answering. In contrast to factoid CQA with a short and accurate answer, non-factoid question-answering usually requires passages as answers. The *diversity*, *shortness* and *sparseness* of answers form interesting challenges for summarization. To tackle these challenges, we propose a sparse coding-based summarization strategy, in which we can effectively capture the saliency of diverse, short and sparse units. Specifically, after transferring all candidate answer sentences into vectors, we present a coordinate descent learning method to optimize a loss function to reconstruct the input vectors as a linear combination of basis vectors. Experimental results on a benchmark data collection confirm the effectiveness of our proposed method in non-factoid CQA summarization. Our method is shown to significantly outperform the state-of-the-art in terms of ROUGE metrics.

Keywords

Community question-answering; Sparse coding; Short text processing; Document summarization

1. INTRODUCTION

In recent years, we have witnessed a rapid growth in the number of users of community question-answering (CQA). In the wake of this development, more and more approaches to CQA retrieval have been proposed, addressing a wide range of tasks, including answer ranking [13–15], answer extraction [16], multimedia QA [9], and question classification [2, 3]. There has been a very strong focus on factoid question-answering, in which there is typically just a single correct answer for a given question, e.g., "Where was X born?" In contrast, in non-factoid question-answering, multiple sparse and diverse sentences may together make up the answers. However, their sparseness and diversity make it difficult to identify all of the information that together covers all aspects of the question.

Multi-document summarization is a task that has been widely used to extract or generate salient sentences to represent a set of input documents [1]. Intuitively, document summarization can be applied to extract sentences and generate a relevant and diverse answer for a given input question, in particular in the context of non-factoid question-answering [4]. However, traditional document summarization methods face a number of challenges when used for summarizing non-factoid answers in CQA. Compared to summarizing news articles, summarizing answers in non-factoid CQA faces specific challenges: (1) Summarization in non-factoid CQA is a recall-oriented problem, in which we need to search as much relevant information as possible. However, the *diverse* topic distribution of answers in non-factoid CQA makes it difficult to generate a summary with high *recall*. (2) The *shortness* and *sparseness* of answers in non-factoid CQA is an obstacle for redundancy-based summarization methods.

The task on which we focus here is *summarizing answers in non-factoid community question-answering* [12]. We propose a sparse-coding strategy to address this summarization problem. Recently, sparse coding strategies have been proved to be effective and efficient in summarizing sparse and diverse semantic units [6]. We apply a sparse coding-based summarization strategy to find a set of sentences that can be used to reconstruct all the input sentences given the input question. In our sparse-coding framework, we directly regard all the answer sentences as basis vectors and utilize the coordinate descent method to optimize our proposed loss function. We evaluate our proposed method on a benchmark dataset released by Tomasoni and Huang [12]. In terms of ROUGE metrics, our proposed sparse-coding based method is found to be very effective in summarizing answers in non-factoid CQA. Moreover, our proposed method significantly outperforms the state-of-the-art baselines.

Our contributions in this paper can be summarized as follows:

[*] These two authors contributed equally to the paper.

ACM ISBN 123-4567-24-567/08/06…$15.00

DOI: 10.475/123_4

Table 1: Glossary.

Symbol	Description
$\mathcal{D}$	candidate answers
$\mathcal{V}$	vocabulary in answers $\mathcal{D}$
$\mathcal{S}$	candidate sentences
$\mathcal{R}$	a summary of answers
$\mathcal{A}$	saliency vector
D	number of answers
S	number of sentences
L	length limit of a summary of answers
s_i	a candidate sentence $s_i \in \mathcal{D}$
q	a question
x	basis vectors corresponding to sentences
w_i	similarity between sentence s_i and q
a_i	saliency score for sentence s_i, $a_i \in \mathcal{A}$
α, λ	parameters in sparse-coding framework

- We address the task of summarizing answers to non-factoid questions in community question-answering by tackling the *diversity*, *shortness* and *sparseness* challenges.
- We regard all answer sentences as basis vectors, and apply the coordinate descent method to optimize a new loss function based on a sparse-coding framework.
- Using a benchmark dataset, our proposed method is shown to be effective and efficient. We also find that our method significantly outperforms state-of-the-art baselines, in terms of ROUGE metrics.

In §2 we formulate our research problem. We describe our approach in §3; §4 details our experimental setup and presents the results; §5 concludes the paper and lists our future directions.

2. PROBLEM FORMULATION

Before introducing the details of our method, we first formulate our research problem. Table 1 lists the notation we use in this paper.

For each non-factoid CQA thread, we suppose there exists a question q and a set of candidate answers $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$, where each candidate answer $d \in \mathcal{D}$ can be represented as a set of sentences, i.e., $d = \{s_{d,1}, s_{d,2}, \ldots, s_{d,S_d}\}$. We assume that, in total, there are S sentences in the CQA thread, i.e., $\mathcal{S} = \{s_1, s_2, \ldots, s_S\}$.

A sparse-coding-based method is proposed to reconstruct the semantic space of a topic, revealed by the answer sentences $\mathcal{S}$. A saliency score $a_i \in [0, 1]$ is determined for each sentence s_i so as to define its contribution in constructing the semantic space of the topic from the answer content. For all sentences, we determine a saliency vector $\mathcal{A} = [a_1, a_2, \ldots, a_S]$. Given a question q, a sentence set $\mathcal{S}$, and a target summary length L, the goal of answer summarization in CQA is to select a subset of sentences $\mathcal{R} \subset \mathcal{S}$ such that the total number of words in $\mathcal{R}$ is no more than L, to maximize the sum of their saliency scores, i.e., $\sum_{s_i \in R} a_i$.

3. METHOD

We propose an unsupervised compressive summarization framework to tackle the answer summarization problem in CQA. An overview of our framework is depicted in Figure 1, in which boxes indicate the question or answer sentences. The grey boxes indicate sentences that are selected in the summary.

The aim of sparse-coding is to find a set of basis vectors x_i that can be used to reconstruct M input vectors $\{x_j\}_{j\in M}$ as a linear combination of basis vectors so as to minimize a loss function. In our summarization task, each topic contains a set of answers. After stemming and stop-word removal, we build a dictionary for the topic by using unigrams and bigrams from the answers. Then, each sentence s_i in answers is represented as a weighted term-frequency vector, i.e., x_i. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_S\}$ denote the term frequency basis vectors of all sentences in the candidate answers. The basis sentence vector space can be constructed from a subset of them, i.e., the sentences included in the summary. To utilize the information contained in the question, we compute the cosine similarity w_i between vectors representing each sentence s_i and a vector representing the question q. Here, these two vectors are generated by Doc2Vec [5]. Because the summary sentences are sparse, we impose a sparsity constraint, λ, on the saliency score vector $\mathcal{A}$ using the L1-norm, as a scaling constant to determine its relative importance.

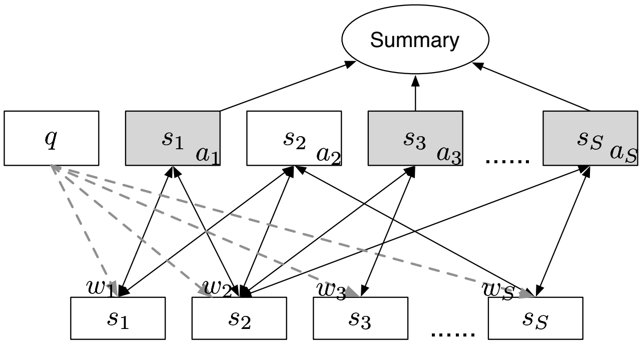


Figure 1: Overview of our sparse-coding approach to non-factoid answer summarization. Boxes indicate the question or answer sentences; each dash arrow indicates the correlations between the question and a answer sentence; each arrow reflects the correlations between two sentences.

Putting things together, we arrive at the following loss function:

$$J = \min \frac{1}{2S} \sum\nolimits_{i=1}^{S} w_i \left\| x_i - \sum\nolimits_{j=1}^{|\mathcal{R}|} a_j \cdot x_j \right\|_2^2 + \lambda \|\mathcal{A}\|_1 \quad (1)$$

subject to:

1. $\forall a_j \in \mathcal{A}, a_j \geq 0$;
2. $\lambda > 0$;
3. $\sum_{s_j \in \mathcal{R}} |s_j| \leq L$;

Here, $|s_j|$ is the number of words in the sentence $s_j \in \mathcal{R}$. Based on our loss function, we formulate the task of summarizing answers for non-factoid CQA as an optimization problem in sparse coding. To learn the saliency vector $\mathcal{A}$, we utilize the coordinate descent method to iteratively optimize the target function about the saliency vector $\mathcal{A}$ until it converges. The details of the coordinate descent method is shown in Algorithm 1. Given a saliency score a_i for each sentence $s_i, \in \mathcal{S}$, we apply a greedy algorithm to select sentences according to their saliency score.

4. EXPERIMENTS

4.1 Dataset

We use a benchmark dataset released by Tomasoni and Huang [12]. Based on a Yahoo! Answers data collection with 89,814 question-answering threads, Tomasoni and Huang [12] removed factoid questions by applying a series of patterns for the identification of complex questions, and only leave non-factoid question-answering threads in the following patterns:

Algorithm 1: Coordinate descent algorithm for answer summarization

```
Input:
Answer sentences $\mathcal{S} = \{s_1, s_2, ..., s_S\}$, question $q$, correlation weight $w_i$ between a sentence $s_i$ and $q$, penalty parameter $\lambda$, and stopping criterion $T$ and $\gamma$
Output: Saliency vector $\mathcal{A} \in \mathbb{R}^S$;
Initialize $\mathcal{A} \to 0$; $k \to 0$;
Transfer sentences to basis vectors $\mathcal{X} = \{x_1, x_2, ..., x_S\}$;
$z = \sum_{i \in S} x_i^2$;
while $k < T$ do
    Reconstructing $\overline{x} = \sum_{i \in S} a_i^k x_i$;
    Take partial derivatives: $\frac{\partial J}{\partial a_i} = \frac{1}{S} \sum_{j \in S} w_j (x_j - \overline{x})^T \overrightarrow{x_i}$;
    Select the coordinate with maximum partial derivative:
        $i' = \arg\max_{i \in S} \left| \frac{\partial J}{\partial a_i} \right|$;
    Update the coordinate by soft-thresholding:
        $a_{i'}^{k+1} = S_\lambda(a_{i'}^k - \eta \frac{\partial J}{\partial a_{i'}})$;
    where $S_\lambda : a \to sign(a) \max(a - \lambda, 0)$;
    if $J_{\mathcal{A}^{k+1}} - J_{\mathcal{A}^k} < \gamma$ then
        break;
    end
    $k \to k + 1$;
end
```

- Why, What is the reason [. . .]
- How to, do, does, did [. . .]
- How is, are, were, was, will [. . .]
- How could, can, would, should [. . .].

The ground truth of all these QA summaries is manually generated by human experts. In total, the dataset in our experiments includes 361 answers, 2,793 answer sentences, 59,321 words and 275 manually generated summaries.

4.2 Baselines and evaluation metrics

We write **SPQAS** for our sparse-coding based method as described in Section 2. To assess the contribution of our proposed method, we perform comparisons between our proposed method and state-of-the-art baselines in our experiments.

- We use the metadata-aware question-answering summarization method (**MaQAS**, [12]) as the baseline for CQA answer summarization.
- A widely-used multi-document summarization model, **LexRank** [1], is also considered in our experiments.
- Finally, we also use **BestAns**, a baseline that uses the top-ranked answer of the QA thread,
- and **Random**, which extracts sentences randomly.

Following [12], we set the length limit of the CQA answer summary to 250 words. We remove stop words and apply Porter stemming.

We adopt the ROUGE evaluation metrics [8], a widely-used recall-oriented metric for document summarization that evaluates the overlap between a gold standard and candidate selections. We use ROUGE-1 (*unigram based method*), ROUGE-2 (*bigram based method*) and ROUGE-L (*longest common subsequence*) in our experiments. Statistical significance of observed differences between the performance of two runs is tested using a two-tailed paired t-test and is denoted using $^\blacktriangle$ (or $^\blacktriangledown$) for strong significance for $\alpha = 0.01$; or $^\vartriangle$ (or $^\triangledown$) for weak significance for $\alpha = 0.05$.

4.3 Results

Table 2 lists the ROUGE performance of all the methods that we consider in terms of ROUGE-1, ROUGE-2 and ROGUE-L.

Table 2: Overview of performance comparisons of all methods in answer summarization. Statistically significant differences between SPQAS and MaQAS, and between MaQAS and LexRank, are marked in the upper right hand corner of the ROUGE score, respectively.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Random	0.425	0.345	0.420
BestAns	0.420	0.373	0.418
LexRank	0.584	0.438	0.565
MaQAS	0.674$^\blacktriangle$	0.588$^\blacktriangle$	0.663$^\blacktriangle$
SPQAS	**0.753**$^\blacktriangle$	**0.678**$^\blacktriangle$	**0.750**$^\blacktriangle$

We find that SPQAS outperforms the other four baselines, and significantly outperforms MaQAS in terms all three ROUGE metrics. Random and BestAns perform the worst. Since genetic summarization methods neglect the correlation between a question and answers, the LexRank method does not perform well in the CQA answer summarization task. We also find that the difference between MaQAS and LexRank is always significant.

We further compare SPQAS with MaQAS: SPQAS offers relative performance improvements of 11.7%, 15.3% and 13.1%, respectively, for the ROUGE-1, ROUGE-2 and ROUGE-L metrics. We also find that SPQAS outperforms the MaQAS baseline with a statistical significance difference at level $\alpha < 0.01$ in terms of all ROUGE metrics. Figure 2 shows the ROUGE-1 performance of SPQAS and MaQAS with varying average length of answers per thread. We can find that most of answers' length is between 100 and 200 words. Moreover, we find that SPQAS has a similar ROUGE-1 performance as that of MaQAS for most threads that the average length of answers is more than 200 words. We also find that for most of the threads, with the increase of the average length, the ROUGE-1 performance of both SPQAS and MaQAS decreases monotonically.

4.4 Case study

To illustrate our method, to answer a question about "how to cure indigestion," we generate a summary with our sparse-coding model. The question, candidate sentences and our summary are given in Figure 3. We can find that the summary extract sentences from candidate answers. By reviewing the answer summary of this QA thread, we find that the answer summary generated by our model can find important and different aspects of answers given the question, which, intuitively, verifies the effectiveness of our sparse-coding method in searching salient and diverse results.

5. CONCLUSION AND FUTURE WORK

We have considered the task of answer summarization for non-factoid community question-answering. We have identified the main challenges: the diverse topic distribution, and the shortness and sparseness of answers. We have proposed a sparse-coding strategy to predict the saliency vector of each candidate sentence, in which

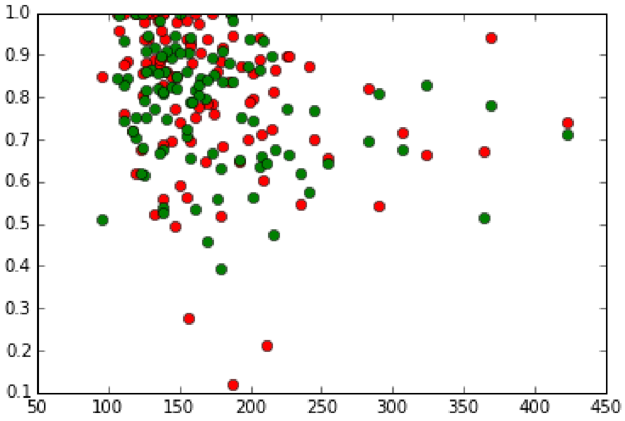
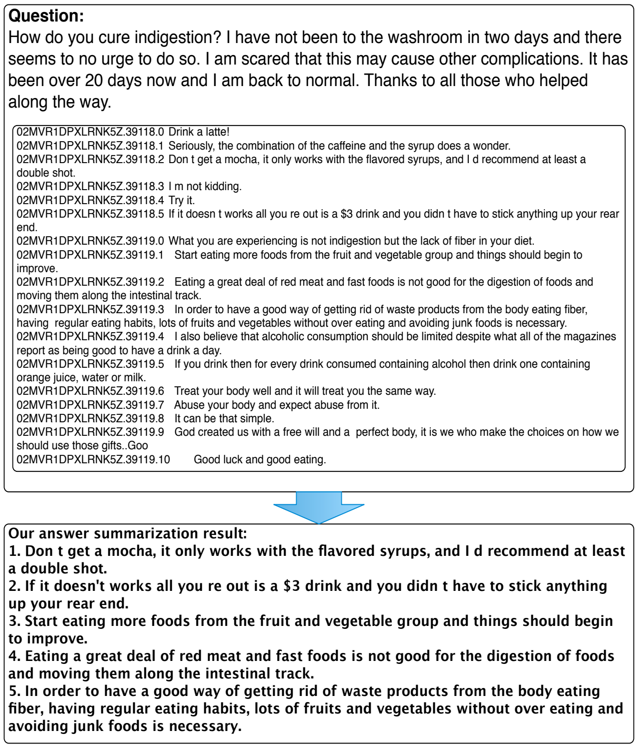


Figure 2: ROUGE-1 performance of SPQAS (red) and MaQAS (green) with different length of answers. The x-axis denotes the average number of words of answers in each QA thread, whereas the y-axis denotes the ROUGE-1 value.



Question:
How do you cure indigestion? I have not been to the washroom in two days and there seems to no urge to do so. I am scared that this may cause other complications. It has been over 20 days now and I am back to normal. Thanks to all those who helped along the way.

02MVR1DPXLRNK5Z.39118.0 Drink a latte!
02MVR1DPXLRNK5Z.39118.1 Seriously, the combination of the caffeine and the syrup does a wonder.
02MVR1DPXLRNK5Z.39118.2 Don t get a mocha, it only works with the flavored syrups, and I d recommend at least a double shot.
02MVR1DPXLRNK5Z.39118.3 I m not kidding.
02MVR1DPXLRNK5Z.39118.4 Try it.
02MVR1DPXLRNK5Z.39118.5 If it doesn t works all you re out is a $3 drink and you didn t have to stick anything up your rear end.
02MVR1DPXLRNK5Z.39119.0 What you are experiencing is not indigestion but the lack of fiber in your diet.
02MVR1DPXLRNK5Z.39119.1 Start eating more foods from the fruit and vegetable group and things should begin to improve.
02MVR1DPXLRNK5Z.39119.2 Eating a great deal of red meat and fast foods is not good for the digestion of foods and moving them along the intestinal track.
02MVR1DPXLRNK5Z.39119.3 In order to have a good way of getting rid of waste products from the body eating fiber, having regular eating habits, lots of fruits and vegetables without over eating and avoiding junk foods is necessary.
02MVR1DPXLRNK5Z.39119.4 I also believe that alcoholic consumption should be limited despite what all of the magazines report as being good to have a drink a day.
02MVR1DPXLRNK5Z.39119.5 If you drink then for every drink consumed containing alcohol then drink one containing orange juice, water or milk.
02MVR1DPXLRNK5Z.39119.6 Treat your body well and it will treat you the same way.
02MVR1DPXLRNK5Z.39119.7 Abuse your body and expect abuse from it.
02MVR1DPXLRNK5Z.39119.8 It can be that simple.
02MVR1DPXLRNK5Z.39119.9 God created us with a free will and a perfect body, it is we who make the choices on how we should use those gifts..Goo
02MVR1DPXLRNK5Z.39119.10 Good luck and good eating.

Our answer summarization result:
1. Don t get a mocha, it only works with the flavored syrups, and I d recommend at least a double shot.
2. If it doesn't works all you re out is a $3 drink and you didn t have to stick anything up your rear end.
3. Start eating more foods from the fruit and vegetable group and things should begin to improve.
4. Eating a great deal of red meat and fast foods is not good for the digestion of foods and moving them along the intestinal track.
5. In order to have a good way of getting rid of waste products from the body eating fiber, having regular eating habits, lots of fruits and vegetables without over eating and avoiding junk foods is necessary.

Figure 3: An example answer summary for a question-answering thread about "how to cure indigestion." The answer summary extracts sentences from candidate answers.

we directly regard all the answer sentences as basis vectors and propose a new loss function. We utilize a coordinate descent method to optimize our target function. We have demonstrated the effectiveness of our proposed method by showing a significant improvement over multiple baselines tested with a benchmark dataset.

Limitations of our work include its ignorance of syntactic information and of semantic dependencies among answers. We also find that our method does not perform so well on answers with long text. As to future work, entity-based document expansion is worth considering [7, 10, 11]. Also, transferring our method to the cross-language CQA answer summarization and online answer summarization setting should be given new insights. It is interesting to consider a personalized summarization task on question-answering communities, based on user clustering [17]. Finally, supervised and semi-supervised learning can be considered for improving the accuracy in CQA answer summarization.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant No. 61272240 and 61103151, the Big Data Institute, University College London, Ahold, Amsterdam Data Science, the Bloomberg Research Grant program, the Dutch national program COMMIT, Elsevier, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nr 312827 (VOX-Pol), the ESF Research Network Program ELIAS, the Royal Dutch Academy of Sciences (KNAW) under the Elite Network Shifts project, the Microsoft Research Ph.D. program, the Netherlands eScience Center under project number 027.012.105, the Netherlands Institute for Sound and Vision, the Netherlands Organisation for Scientific Research (NWO) under project nrs 727.011.005, 612.001.116, HOR-11-10, 640.006.013, 612.066.930, CI-14-25, SH-322-15, 652.002.001, 612.001.551, the Yahoo Faculty Research and Engagement Program, and Yandex. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

6. REFERENCES

[1] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, pages 457–479, 2004.

[2] G. Feng, K. Xiong, Y. Tang, A. Cui, J. Bai, H. Li, Q. Yang, and M. Li. Question classification by approximating semantics. In *WWW*. ACM, 2015.

[3] K. Hacioglu and W. Ward. Question classification with support vector machines and error correcting codes. In *HLT-NAACL*. ACL, 2003.

[4] M. Keikha, J. H. Park, and W. B. Croft. Evaluating answer passages using summarization measures. In *SIGIR*. ACM, 2014.

[5] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[6] P. Li, L. Bing, W. Lam, H. Li, and Y. Liao. Reader-aware multi-document summarization via sparse coding. In *IJCAI*, 2015.

[7] S. Liang, Z. Ren, and M. de Rijke. The impact of semantic document expansion on cluster-based fusion for microblog search. In *ECIR*. Springer, 2014.

[8] C. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*. ACL, 2004.

[9] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua. Beyond text qa: Multimedia answer generation by harvesting web information. *Multimedia, IEEE Transactions on*, 15(2):426–441, 2013.

[10] D. Odijk, E. Meij, and M. de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *Open research Areas in Information Retrieval (OAIR 2013)*, July 2013.

[11] Z. Ren, M.-H. Peetz, S. Liang, W. van Dolen, and M. de Rijke. Hierarchical multi-label classification of social text streams. In *SIGIR*. ACM, 2014.

[12] M. Tomasoni and M. Huang. Metadata-aware measures for answer summarization in community question answering. In *ACL*. ACL, 2010.

[13] M. Wang. A survey of answer extraction techniques in factoid question answering. *Computational Linguistics*, 1(1), 2006.

[14] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen. Cqarank: jointly model topics and expertise in community question answering. In *CIKM*. ACM, 2013.

[15] L. Yang, Q. Ai, D. Spina, R.-C. Chen, L. Pang, W. B. Croft, J. Guo, and F. Scholer. Beyond factoid QA: Effective methods for non-factoid answer sentence retrieval. In *ECIR*. Springer, 2016.

[16] X. Yao, B. Van Durme, C. Callison-Burch, and P. Clark. Answer extraction as sequence tagging with tree edit distance. In *HLT-NAACL*. ACL, 2013.

[17] Y. Zhao, S. Liang, Z. Ren, J. Ma, E. Yilmaz, and M. de Rijke. Explainable user clustering in short text streams. In *SIGIR*. ACM, 2016.