Learning Novelty-Aware Ranking of Answers to Complex Questions

Shahar Harel
Technion - Israel Institute of Technology
sshahar@cs.technion.ac.il

Sefi Albo
Technion - Israel Institute of Technology
sefialbo@campus.technion.ac.il

Eugene Agichtein
Emory University and Amazon Research
eugene.agichtein@emory.edu

Kira Radinsky
Technion - Israel Institute of Technology
kirar@cs.technion.ac.il

ABSTRACT

Result ranking diversification has become an important issue for web search, summarization, and question answering. For more complex questions with multiple aspects, such as those in community-based question answering (CQA) sites, a retrieval system should provide a *diversified* set of *relevant* results, addressing the different aspects of the query, while minimizing redundancy or repetition. We present a new method, *DRN* , which learns novelty-related features from unlabeled data with minimal social signals, to emphasize diversity in ranking. Specifically, *DRN* parameterizes question-answer interactions via an LSTM representation, coupled with an extension of neural tensor network, which in turn is combined with a novelty-driven sampling approach to automatically generate training data. *DRN* provides a novel and general approach to complex question answering diversification and suggests promising directions for search improvements.

ACM Reference Format:
Shahar Harel, Sefi Albo, Eugene Agichtein, and Kira Radinsky. 2019. Learning Novelty-Aware Ranking of Answers to Complex Questions. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313457

1 INTRODUCTION

Search engines and conversational assistants are expected to answer increasingly complex questions, such as requests for advise or opinions. Perhaps the most important source are Community Question Answering (CQA) sites, which offer an opportunity for millions of online users to contribute, share, and discover knowledge. The number of answers for community questions can range from a few, to dozens, but it is not realistic to expect a user to read them all. At the same time, a user might be interested in several aspects of the answers, and would benefit from being exposed to several relevant answer aspects. For example, when looking for recommendations of a movie or a restaurant, one could be interested in several viewpoints, and not just a single answer.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313457

An ideal ranking of answers to such a question should combine relevance to the asked question, as well as novelty – by addressing different aspects of the question instead of repeating already seen information in previous answers.

So far, the best performing state-of-the-art systems addressing the CQA novelty problem have been supervised, requiring extensive training datasets [20]. For these methods, two types of training labels are required – relevance labels, and explicit aspects diversity judgments. Relevance judgments datasets are more common, and can be inferred based on implied ranking of users on social sites (e.g., through clicks or user ratings of answers) and indeed have been used in a form of a distant supervision for training learning-to-rank systems. In contrast, diversity-labeled datasets in large amounts need to be obtained for each new domain to enable supervised learning. Additionally, supervised learning-to-rank systems typically apply statistical models to features, engineered for a specific domain. The required feature engineering for each problem leads to complex and brittle systems with many steps and dependencies. In contrast, we present an approach that avoids feature engineering, yet reaches high ranking performance for multiple types of questions, by introducing a novel deep learning algorithm leveraging large unlabeled data sets with minimal distant supervision based on social signals, and does not require direct diversity judgments.

In this work, we introduce *DRN* (Diversity Ranking Network), which, to the best of our knowledge, is the first neural network approach for community-question-answering ranking that addresses both relevance of the answers, and their marginal novelty. Specifically, our model captures relationships between questions and their corresponding answers with a non-linear tensor layer, which promotes answers that are both relevant to the question, and diversified with respect to other answers. Our method combines the trained neural network scores with an iterative algorithm for efficiently ranking the community answers with respect to previously chosen answers. Our system reaches state-of-the-art results, compared to previous novelty-ranking unsupervised approaches, and shows comparable results to the supervised methods that use extensive feature engineering and labeled datasets. Specifically, our contributions are:

- A novelty-driven sampling approach (Section 3.3), for leveraging unlabeled data and minimal social signals as distant supervision to optimize a novelty-driven loss function.
- A novel end-to-end deep-learning approach, modeling interactions of question and answers for identifying novelty and relevance of answers without manual feature engineering.

Question	How can I lose weight I'm 73 lbs I need to exercise but I'm fat and lazy so can you please help me does someone have tips to lose weight?
Answer 1	Exercise, Nutrient dense foods. If you lose weight very quickly you will be a lot less likely to keep it off. The first thing I would suggest is to believe in you.
Answer 2	The more you feel good about yourself the easier and faster it is to lose weight. Exercise, exercise, exercise! Please consult your doctor before beginning an exercise or weight loss program. Safest way is to lose 2-3 lbs a week.
Answer 3	diet always trumps exercise. Add or lose muscle. Too much body fat is not good for your health. Your lifestyle habits can influence . You have to diet and exercise.
Answer 4	I've gone down from 103 kg to 89 kg in three months by reducing my caloric intake and walking at least 10 kms a day and sometimes I even up to 20 km
Answer 5	trust me it wont be healthy why are people seriously giving you tips on losing weight? even if it is through healthy means, you don't need tips on losing weight.
Answer 6	If indeed some infections contribute to obesity in people , we could have a potentially very simple and effective prevention strategy vaccination...

Figure 1: An example question with candidate answers from the TREC LiveQA dataset, with relevant answers partitioned into propositions corresponding to aspects. Each aspect is color-coded according to human annotation.

- A new method for novelty-aware scoring of answers for complex questions over social media (Section 3) using a neural architecture, and an iterative algorithm for ranking answers, reaching state-of-the-art results.

We report results of an extensive experimental evaluation of different variations of our *DRN* method, showing significant improvements over state-of-the-art approaches (Section 5) on two benchmark datasets, one of which we created and will share with the research community.

Our *DRN* model lays the foundation for deep learning for answer diversification, that is aware of both relevance and novelty. Making it a promising approach for many other ranking tasks.

2 RELATED WORK

Ranking social media content, and specifically CQA content, has been an active area of research [18, 19]. CQA content in particular has a number of distinct characteristics, such as the duality of questions and answers, and social signals, that have been explored for improving retrieval performance. For example, Xue et al. [35] experimented with variations of language models for retrieving question-answer pairs from CQA archives, while Wang et al. [31] demonstrated improvements by modeling question-answer relationships. It has also been shown that additional features for CQA ranking can be more naturally incorporated using a supervised Learning-to-Rank (LTR) approach (e.g., see Bian et al. [3], Cao et al. [5], Surdeanu et al. [27]), with the added benefit that LTR systems can be optimized for the measure of interest, such as result diversity, in addition to relevance. However, supervised LTR approaches require extensive training data, a shortcoming that we address in the next section by introducing an deep learning method trained in a distant supervised manner. While numerous studies on document novelty and diversification for Web Search and IR have been done in the past [17, 23, 33] the CQA setting exhibits significant differences. First, the answers provided to a specific question are mostly relevant, as opposed to document ranking where many documents are irrelevant to the query. Therefore, common methods in IR, such as MMR perform significantly worse on this task (see Section 4.1). Additionally, many of those methods [33] are supervised and require significant labeling per domain, which is impractical for CQA. Second, CQA answers are shorter than most documents and therefore pose additional challenges to diversification. Finally, the length of the queries in Web search is significantly shorter compared to the length of the CQA questions. Specifically in the CQA domain, Szpektor et al. [28] presented a recommendation algorithm for suggesting questions to potential answerers in CQA sites. They showed that ignoring diversity degrades recommendation performance. Recently, deep learning techniques have been applied to many variants of LTR and Question Answering [13, 30, 32, 34] and specifically for Learning To Rank in CQA, e.g., [6, 21, 33, 36, 37]. While building on these previous works, our effort is distinct in that we develop an end-to-end method for automatically learning to diversify CQA ranking with only minimal distant supervision based on social signals.

3 *DRN*: DEEP NETWORK ANSWER DIVERSIFICATION MODEL

Our goal is to train a ranking function for CQA that emphasizes both relevance and marginal novelty. Consider an example CQA question in Figure 1, which shows the first six out of eleven answers to the question. The first three answers contain a wide range of aspects, with some shared across answers, and others specific to a certain answer. Answer 4 is relevant to the question, but does not add novelty with respect to former answers. Answer 5 is not highly relevant to the question, and does not address many aspects. Finally, Answer 6 is irrelevant and should be ranked last.

We model the problem as a supervised-learning problem, optimizing to rank a triplet consisting a question and two possible answers, where a positive example is one consisting of both relevant and diverse answers. We learn a fixed-size vector representation of question and candidate answers using a LSTM layer. Then, triplets of question and two answers are combined with a neural tensor layer, and combined score for each triplet is computed. Figure 2 presents *DRN* architecture at high level. The key innovation of our

work is a novel, novelty-driven sampling process, which emphasizes relevance and novelty simultaneously for training the model. A greedy algorithm selects at each round an answer that is both relevant to the question and introduces most novelty with respect to former ranked answers.

3.1 Low-Dimensional Embedding of Question and Answer Text

To be able to perform the neural network matrix operations over question and answer, we need to represent them as a fixed-length vector. In this work, we use a deep recurrent neural network architecture (RNN) to process sequentially each word in the question and the answers, and to produce a low dimensional vector at each step. Intuitively, the output vector at each step represents the information given in the sentence, until the current word. We use an LSTM model [11, 12] and obtain the last word in the sentence embedding as the semantic vector representation for the sentence.

3.2 Modeling Question and Answer Context

An answer to a question can be a good candidate or a bad candidate depending on the context where it is given, in our case, within the context of the question asked. Modeling the interactions between the answer and the question is a key aspect in being able to judge its relevance and usefulness. One could try to model these interactions by designing hand-crafted features, which is not an easy task. Our approach is to discover such features automatically. In the past, the use of tensors to model interactions was shown to be successful [21, 26]. While those models were required to model a single interaction, to capture *novelty*, it must capture the interaction of the question with several answers. To address *novelty*, we argue that one should consider interactions between *all entries* in a triplet ($question, answerA, answerB$). With these interactions, we can model both relevance of an answer to a question, and the novelty that the answer introduces with respect to previous chosen answer. Intuitively, we want this score to be high if the first answer is relevant to the question, and diverse with respect to the second answer. Modeling such a score function can be done by linearly combining two score functions, each trained for different desired property, e.g., relevance and novelty. Using a more general non-linear function, allows the algorithm to learn to optimize both relevance and diversity simultaneously. To accomplish this, we present an extension to [26] for modeling triplets, denoted $TNTN$, the parameters of the network will be shared across the triplet, thus able to encode features representing semantic interactions between *all entries* in the triplet. This is done by firstly introducing the question into the parametrization, $W = M^{[1:n\cdot r]}v_q$, and then multiplying the answers with a bilinear tensor product. $TNTN$ can be described as follows:

$$s(q,a,a') = u^T f\left(v_a^T W^{[1:r]} v_{a'} + V\begin{bmatrix} v_a \\ v_q \\ v_{a'} \end{bmatrix} + b\right) \quad (1)$$

where $M^{[1:n\cdot r]} \in R^{n\times n\times n\times r}$, $V \in R^{r\times 3n}$,$b \in R^r$ and $u \in R^r$ $v_q, v_a \in R^n$.

v_q, v_a and v'_a are the question and two answer vector representations from LSTM layer, f is a nonlinearity applied element-wise,

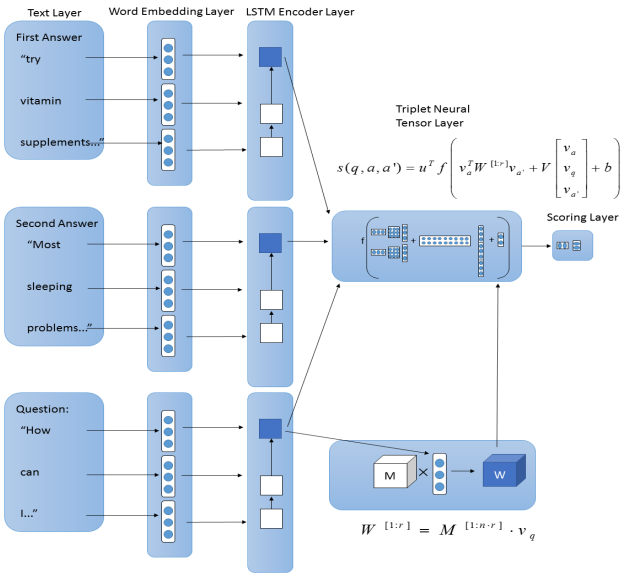

Figure 2: Visualization of the *DRN* network

The tensor product $M^{[1:n\cdot r]}v_q$ is done by multiplying $n \cdot r$ matrices of $n \times n$ by v_q, thus results at r matrices of size $n \times n$, or in other words a tensor $W^{[1:r]} \in R^{n\times n\times r}$. The bilinear tensor product $v_a^T W^{[1:r]} v'_a$ results in a vector $h \in R^r$, where each entry is computed by one slice i = $1, \ldots, r$ of the tensor. The other parameters to be learned are $V \in R^{r\times 3n}$, $b \in R^r$ and $u \in R^r$.

3.3 Novelty Driven Sampling

Intuitively, our scoring function should learn to address both relevance and novelty. To tackle this, we introduce a novel sampling approach, which presents three types of examples to our network, one positive and two types of negative examples. To guide the sampling, we introduce a *Novelty* function that is able to measure the amount of diversity between two given answers.

To construct the positive example for a given question – $(q, Ba, BaDiv)$ for both high relevance and novelty, we select the best answer Ba in a distant supervised manner to be the most voted answer by Yahoo! users, which is commonly assumed to be highly relevant to the question. To address novelty with respect to Ba, we construct $BaDiv = argmax_{a\in A(q)} Novelty(Ba, a)$ via the *Novelty* function.

We construct two types of negative examples – one expressing low novelty of a relevant answer, and the second representing low relevance. The first negative example is constructed by selecting an answer Ba as before, and $BaSim = argmin_{a\in A(q)} Novelty(Ba, a)$ is an answer the *Novelty* function suggests to be similar to Ba and therefore not Novel.

The second type of a negative example, $(q, Ra, RaDiv)$ is constructed by selecting a random answer Ra to a random question, and therefore assumed to have poor relevance to the question q of the triplet. Such a triplet is used as a negative example, even though the answer $RaDiv = argmax_{a\in A(Random(q))} Novelty(Ra, a)$ is *novel* with respect to Ra by the *Novelty* function.

Several *Novelty* function families can be used for the sampling process without any significant alternations to *DRN* . We experimented with cosine distance over topic representation of the text, e.g. LDA [4], and over embedding space, e.g. doc2vec [16]. On validation dataset, we found that the LDA function performed better for the diversification task. Thus, we represent each answer with its corresponding LDA topic probabilities vectors and their level of diversity is measured with the cosine distance.

Loss function for training *DRN* : Our loss function is inspired by the MMR [7] framework. Our approach generalizes the idea of combining relevance and diversity functions by encoding them both into a loss function. We train the network with max-margin criterion, which provides an alternative to probabilistic, likelihood-based estimation methods by concentrating directly on the robustness of the decision boundary of a model [29]. Intuitively, a good example will be separated in space by some margin from the sampled corrupted examples. Our loss function for diversification is defined as:

$$\begin{aligned} L = \sum_{C,C'} & - [s(q, Ba, Ba_{Div}) - s(q, Ba, Ba_{Sim})] \\ & - [s(q, Ba, Ba_{Div}) - s(q, Ra, Ra_{Div})] \\ & + \gamma + \lambda \left\|\Theta\right\|_2^2 \end{aligned} \tag{2}$$

where $\gamma > 0$ is the margin hyper-parameter, C is the training collection of good examples $(q, Ba, BaDiv)$ triplets from the data and C' denotes the collection of all corrupted triplet examples $(q, Ba, BaSim)$ and $(q, Ra, RaDiv)$. To minimize our objective function we used the Adam Optimizer [14].

Algorithm 1 Ranking for Diversification Algorithm

1: NN Inference: Compute s(q,a,a') - $n \times n$ scores matrix for question q
2: R $\leftarrow$ $SelectFirstAnswer$
3: for $j = 1$ to $|A(q)| - 1$ do
4: R $\leftarrow$ $argmax_a \Sigma_{a' \in R} s(q, a, a')$
5: end

3.4 Ranking Algorithm

In the previous sections we described the construction of an embedding of questions and answers (section 3.1), which is then used in a neural tensor network to assign relevance and novelty scores. In this Section we leverage these scores into a novelty-driven ranking Algorithm (1). The algorithm greedily selects answers to maximize scores learned by the neural network. For each question and the set of its n corresponding answers, we perform $n \times n$ forward inference steps, thus producing a score for every combination of the question and two of its answer. This step results in an $n \times n$ scores matrix for each question and its answers. We then select the first answer and add it to the ranked answers list. We note that in this work, we focus on the diversification aspect, i.e., given a selected first answer, we wish to select relevant answers that introduce new aspects. To initialize our ranking algorithm, we must choose the first answer. Selecting the best answer to a question is a well researched task in the literature [2, 25]. Inspired by prior work, we choose the first answer using a classifier optimized to select answer with high number of aspects.

Going back to our ranking algorithm, in line 4 we iteratively choose the next answer as the one with the highest sum of scores with respect to the question and the answers chosen so far. Intuitively, the answer that maximize this step, is the one with high relevance to the question, and simultaneously a high level of novelty with respect to the answers chosen in earlier steps, where all these answers are taken into account.

4 EXPERIMENTAL SETTING

In this section we describe the state of the art baselines that we compare to our method, *DRN* , as well as the evaluation metric and the CQA datasets used in the experiments. All the methods, including baselines, first used basic pre-processing, including: tokenization, stop-words removal, and spell checking, which is an important step in CQA domain due to its noisy nature.

Tuning and Final Configuration Based on a validation set, we tuned our network hyper-parameters, with final values of r = 5, learning rate = 0.001, batch size = 100, negative examples = one of each type, regularization λ = 0.0001, LSTM state size = 50.

4.1 Evaluation Metrics

One of the most prominent metrics for diversification measurement in QA and IR systems is $\alpha NDCG$ [9]. To compute $\alpha NDCG$ we first define $r_k^i = 1$ as an indicator function for the event where the i-th answer contain proposition related to the k-th aspect. The complete $\alpha NDCG$ formulation is:

$$\alpha NDCG = \sum_k \frac{1}{\log_2(k+1)} \sum_i r_k^i (1-\alpha)^{s_{i,k-1}} \tag{3}$$

where $s_{i,k-1} = \sum_{j=1}^{k-1} r_j^i$.
Intuitively, $\alpha NDCG$ rewards answers with the highest discounted cumulative gain, where gain is measured by the novelty the answer aspects introduce with respect to previously ranked answers.

Additionally, we present results with an additional common IR metric for diversification – ERR-IA [8]. ERR-IA is defined as the expectation of ERR over the different aspects, where R_i is the probability that the user will be satisfied with the i-th document (or in our case, the i-th answer) and n is the number of answers retrieved. Specifically, *ERR* is defined as:

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r \tag{4}$$

Together, $\alpha NDCG$ and ERR-IA comprise the two main metrics recommended to use for evaluating search result diversity, as surveyed in reference [24].

4.2 Baseline Methods

BM25 [22]: Originally proposed for ad-hoc information retrieval, BM25 ranks documents according to term frequencies.
MMR [7]: The maximal marginal relevance is a traditional method promoting diversity by combining two standalone similarity and diversity functions.
LDARanker [15]: This is an unsupervised algorithm that aims to diversify ranking by selecting answers with good coverage over the set of all answers to a question.
SimRanker [20]: Serves as our state-of-the-art *supervised* baseline method. This algorithm use the notion of propositions and leverage on data with diversity labels for training parameters to

predict whether two propositions are diverse, then a greedy procedure selects iteratively answer that best combines a set of relevant propositions that are both diverse, and important.

SimRanker-ESA [20]: Serves as our state-of-the-art *unsupervised* baseline method. *SimRanker − ESA* only differs from the supervised version in the similarity function between propositions, as it uses a cosine similarity of the "Explicit Semantic Similarity", or CQA-ESA[10] representations of each proposition.

Due to lack of access to a proprietary method described in [20], our method *DRN* is compared to *SimRanker* and *SimRanker − ESA* only on the first benchmark with results given by the authors, for calibration of our method to the state of the art. On the second benchmark, we compare *DRN* with *BM*25, *MMR* and *LDARanker*.

4.3 Datasets

In 2009, Yahoo published a dataset with questions and their corresponding answers from Yahoo Answers web site. The corpus contains 4,483,032 questions and their answers. In addition to question and answer text, the corpus contains a small amount of metadata, (e.g. most voted answer). We use a subset of this dataset with randomly chosen 250k questions and their corresponding answers for generating training data with our novel sampling approach.

Yahoo-Novelty Dataset: Recently, [20] developed a dataset specifically for diversification, the dataset consist a random sample from Health category in Yahoo answers. The data consist of questions and answers, with their relevant propositions clustered into aspects.

LiveQA-Novelty Dataset: To be able to test our system on larger scale, and on questions from a variety of categories, we have developed an additional larger dataset designed specifically for diversification, following the methodology presented by Omari et al. [20]. The questions were sampled from three categories of the LiveQA 2015 dataset [1]: Health, Pets, and Arts & Humanities. We then selected the subset of the questions with at least 7 answers. First, each answer is partitioned into propositions, and each proposition is then manually judged for its relevance to the question. Propositions found to be relevant to a question from all of its answers are then grouped into aspects.

Specifically, annotation of the LiveQA-Novelty data was done in 3 phases: (1) Identifying relevant propositions in answers[1] - We have used Amazon Mechanical Turk for the annotations. Each HIT* (A Human Intelligence Task) contained a question and an answer partitioned automatically into propositions, and the MTurk workers were given the instruction to mark each proposition as relevant or not to the question. (2) Filtering questions [2] - Second step of relevant propositions filtering was done manually by the authors. During this phase we removed all the questions with fewer than 5 total relevant propositions in their answers. (3) Aspects clustering - the authors manually grouped relevant propositions for each of the test question into clusters representing aspects. We computed the answer/aspect-cluster agreement of two annotators using Cohen's kappa statistic. The score $kappa = 0.72$ suggests moderate to high agreement indicating the labeling is reliable. As a result of this labeling, our LiveQA-Novelty dataset contains 207 questions, with 2,488 answers in total, with at least 7 answers for each question. We make our labeled LiveQA-Novelty Dataset available for the research community [1].

[1] https://github.com/shaharharel/CQA_Diversification

Model	α NDCG (Yahoo)	α NDCG (LiveQA)	ERR-IA (Yahoo)	ERR-IA (LiveQA)
LDARanker	0.62	0.58	0.6	0.51
BM25	0.67	0.52	0.53	0.43
MMR	0.66	0.56	0.57	0.46
SimRanker-ESA	0.75	-	-	-
SimRanker	**0.80***	-	-	-
DRN	**0.81***	**0.68***	**0.65***	**0.56***

Table 1: $\alpha NDCG$ for all methods over Yahoo-Novelty dataset (Left) and LiveQA-Novelty dataset (Right); significant improvements are marked with *.

5 EXPERIMENTAL RESULTS

In this section we present our experimental results, and analyze the performance of *DRN* . Then, we take deeper look at *DRN* internal latent space behavior.

5.1 Main Results

Table 1 report the performance of *DRN* and the baseline algorithms on Yahoo-Novelty and LiveQA-Novelty datasets respectively. Statistically significant ($p < 0.05$) improvements are marked with "*"s. The results show that on both datasets, *DRN* outperforms the baselines *BM*25, *MMR*, *LDARanker* and *SimRanker − ESA*. We hypothesize that our algorithm is able to automatically learn novelty-driven features defined over interactions of questions and answers, while *BM*25 and *SimRanker − ESA* are mostly dependent on semantic similarity between questions and answers. One might consider *LDARanker* comparable to our method, as both are based on a topic models (i.e., *DRN* utlizes it during the sampling phase). However, *LDARanker* does not consider the interactions with the question when considering the novelty-driven answers ranking, thereby ignoring essential information. The only method directly comparable to ours is the *SimRanker*-Supervised. However, this is a supervised approach, which requires data with diversity labels for training. In contrast, our method *DRN* is trained in a distant supervised manner, which requires only minimal social signals and does not require any human-annotated diversity labels.

We can observe a significant difference between the performances measured by αNDCG in the Tables 1. The reason for that, is the rate of relevant answers in Yahoo Answers Novelty Based Answer Ranking dataset compared to liveQA Answers Novelty Based Answer Ranking dataset. The former holds a relevant answer rate of 0.8 which means, 80% of the answers contain 1 relevant aspect or more, while the latter holds a rate of 0.45, which makes it a more realistic and difficult environment for all discussed algorithms. Also liveQA Answers contains questions from various categories and therefore complicates the ranking.

5.2 Performance Analysis

In the example in Figure 3, eight relevant aspects appear in ten answers that were provided for the question. *DRN* is able to cover three of them in the first answer, adds another four in the following one where two of them are novel aspects. The third answer the algorithm choose has no relevant aspects – presenting an example where our approach might fail. This failure can be explained by the topic-driven nature of our *Novelty* function (Section 3). The third answer contains a lot of seemingly new topics (e.g., "meditation" and "house plant"). Since our algorithm rewards diverse topics, it

Question	How can I sleep longer? Ok, so it may seem like a silly question, but how do I sleep longer? No matter how tired I am at night, it's impossible for me to sleep more than 5 or 6 hours...
Answer 1; Aspects: 2,4,7	try vitamin supplements. . At night be sure to turn off all electronics that receive radio waves! . Make sure your diet and exercise are at a standard level.
Answer 2; Aspects: 0, 1, 2, 4	Most sleeping problems are probably caused by stress, depression, anxiety and worry. To sleep better just relax and switch off. some light exercise. Common OTC sleep aids include Chamomile tea, 5-HTP, Melatonin.
Answer 3; Aspects: ∅	When it comes to sleep, if I don't get enough, my day gets all out of wack.got a house plant for my bedroom, tried meditation before bed.started drinking chamomile tea.In aggregate, all those things really just moved the needle...
Answer 4; Aspects: 3, 4, 5, 6, 7	Nice shower before bed helps. make sure no disturbance from any noise. Good nutrition and vitamins everyday helps too who knows maybe the cellular radiation has effect on your sleep rhythm. Dont watch tv too late. Make sure room is dark enough so no disruption in the morning.

Figure 3: Example question with its top four answers ranked by *DRN*. The aspects are numbered in the left column and colored in different colors across answers according to human gold-standard annotation.

BM25 Ranking
Answer 1 - Aspects: 2, 4, 7
Answer 2 - Aspects: 2
Answer 3 - Aspects: ∅
Answer 4 - Aspects: 0, 1, 2, 4

Table 2: BM25 ranking by aspects over the same example as in figure 3.

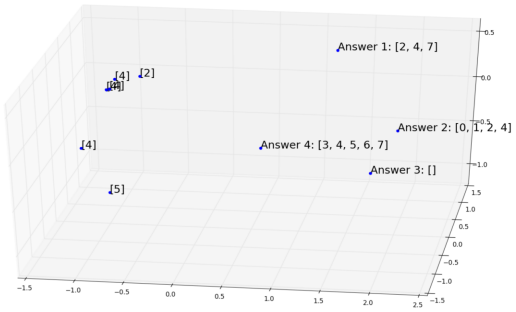

Figure 4: *DRN* LSTM answers vector space for the same question as in figure 3.

might fail if the answer topics are diverse and have a disambiguation in relevance. This issue can be addressed by introducing the network with other types of negative examples which are not based on topic modeling. Another possible explanation is that this answer is controversially annotated as having no relevant aspects to the question. Going back to the ranking, we observe that the 4th ranked answer contains five relevant aspects, of which four are novel aspects. For comparison, table 2 reports the BM25 ranking over the same example from figure 3. This example illustrates the capability of *DRN* to emphasize relevance and novelty simultaneously, by selecting answers that are both relevant to the question and novel with respect to the previously chosen answers.

As discussed earlier, our sampling process and objective function are aimed at learning novelty-driven features automatically, thus we expect the LSTM representation for the answers to express these novel features. Figure 4 gives some intuition of *DRN* behavior and latent features for the example answers from 3. Using the PCA method, we project the answer LSTM vector representations into three most informative dimensions. Each dot in space is an answer, annotated with the list of the aspects it contains. Answers presented in Figure 3 are also labeled with the answer number. We can see a large gap between answers contain low number of aspects (left side), and the answers containing many of the aspects (right side). Moreover, observing the aspects' behavior across answers, we can see that answers with similar aspects tend to be nearby in space, for example, all four answers which contain only the aspect 4 are close, while answers with only the aspects 2 or 5 are further. On the right side of the figure, the answers are far from each other because of the large number of novel aspects they introduce with respect to each other. One exception is Answer 3, analyzed previously.

6 CONCLUSIONS

Modeling both relevance and novelty for ranking answers is a key requirement for effective and user-friendly complex question answering and result presentation, especially with increasing adaption of conversational interfaces and small-screen mobile devices. Previous state of the art approaches have been supervised, requiring extensive manual annotations to obtain both relevance and diversity labels. To address this problem, we presented *DRN*, a novel method which reaches state-of-the-art results using distant supervision for relevance, and no labeled data for diversity. Our method automatically learns a novelty-aware answer scoring function from CQA archives, based on the Neural Tensor Network representations of the interactions between questions and answers, and a novel sampling approach for training, which emphasizes relevance and novelty. We present a new answer ranking algorithm utilizing our model, which iteratively ranks answers by their diversity and relevance with respect to the formerly chosen answers. Experimental results based on two data sets demonstrated that our algorithm outperformed the state of the art baselines that required significant feature engineering and labeled datasets. In future work, we plan to address multiple forms of diversity, by extending our sampling phase to emphasize multiple types of *Novelty* functions and novelty-driven negative examples. Our method could be extended to other settings of complex question answering over social media, and more broadly to complex question answering and web search tasks that require retrieving relevant, useful, and novel answers.

REFERENCES

[1] Eugene Agichtein, David Carmel, Dan Pelleg, Yuval Pinter, and Donna Harman. 2015. Overview of the TREC 2015 LiveQA Track.. In *TREC*.

[2] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 183–194.

[3] 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*. ACM.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[5] 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*. ACM, 201–210.

[6] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization.. In *AAAI*. 2153–2159.

[7] 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*. ACM, 335–336.

[8] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 621–630.

[9] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

[10] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis.. In *IJcAI*, Vol. 7. 1606–1611.

[11] Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850* (2013).

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[13] Mohit Iyyer, Jordan L Boyd-Graber, Leonardo Max Batista Claudino, Richard Socher, and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs.. In *EMNLP*. 633–644.

[14] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Ralf Krestel and Nima Dokoohaki. 2011. Diversifying product review rankings: Getting the full picture. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 138–145.

[16] Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.. In *ICML*, Vol. 14. 1188–1196.

[17] Shangsong Liang, Zhaochun Ren, and Maarten De Rijke. 2014. Fusion helps diversification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (SIGIR)*. 303–312.

[18] 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)*. 27–48.

[19] 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)*. 269–281.

[20] 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*. ACM, 215–224.

[21] Xipeng Qiu and Xuanjing Huang. 2015. Convolutional Neural Tensor Network Architecture for Community-Based Question Answering.. In *IJCAI*. 1305–1311.

[22] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[23] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*. ACM, 881–890.

[24] Rodrygo LT Santos, Craig Macdonald, Iadh Ounis, et al. 2015. Search result diversification. *Foundations and Trends® in Information Retrieval* 9, 1 (2015), 1–90.

[25] Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and predicting answer quality in community QA. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 411–418.

[26] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*. 926–934.

[27] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2011. Learning to rank answers to non-factoid questions from web collections. *Computational linguistics* 37, 2 (2011), 351–383.

[28] Idan Szpektor, Yoelle Maarek, and Dan Pelleg. 2013. When relevance is not enough: promoting diversity and freshness in personalized question recommendation. In *Proceedings of the 22nd international conference on World Wide Web (WWW)*. 1249–1260.

[29] Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 896–903.

[30] Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Vol. 2. 707–712.

[31] 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*. ACM, 179–186.

[32] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural qa as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 271–280.

[33] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling Document Novelty with Neural Tensor Network for Search Result Diversification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 395–404.

[34] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *International Conference on Machine Learning*. 2397–2406.

[35] 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*. ACM, 475–482.

[36] Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. 2017. Attentive Interactive Neural Networks for Answer Selection in Community Question Answering.. In *AAAI*. 3525–3531.

[37] 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 *AAAI*. 3532–3539.