Inferring Frequently Asked Questions from Student Question Answering Forums

Renuka Sindhgatta
IBM Research - India
Bangalore, India
renuka.sr@in.ibm.com

Smit Marvaniya
IBM Research - India
Bangalore, India
smarvani@in.ibm.com

Tejas I Dhamecha
IBM Research - India
Bangalore, India
tidhamecha@in.ibm.com

Bikram Sengupta
IBM Research - India
Bangalore, India
bsengupt@in.ibm.com

ABSTRACT

Question answering forums in online learning environments provide a valuable opportunity to gain insights as to what students are asking. Understanding frequently asked questions and topics on which questions are asked can help instructors in focusing on specific areas in the course content and correct students' confusions or misconceptions. An underlying task in inferring frequently asked questions is to identify similar questions based on their content. In this work, we use hierarchical agglomerative clustering that exploits similarities between words and their distributed representations, reflecting both lexical and semantic similarity of questions. We empirically evaluate our results on real world labeled dataset to demonstrate the effectiveness of the method. In addition, we report the results of inferring frequently asked questions from discussion forums of online learning environment providing lectures to middle school and high school students.

Keywords

frequently asked questions, agglomerative clustering, question similarity, community question answering.

1. INTRODUCTION

Self-paced online learning environments provide valuable learning resources to a large number of students. A primary mechanism of interactions between the students are the discussion forums. These forums enable students to ask questions, answer questions and collaboratively learn. *Question answering forums*, are discussions forums where every thread is a question posted by a student - much like the community question answering (CQA) platforms such as StackOverflow[1], Quora[2]. Over time, a large number of students may post similar questions that could indicate topics susceptible to confusions, misconceptions or course content requiring further explanations. Most question answering forums allow a student or user to search similar questions present in the archives, using information retrieval technique. While searching similar questions is useful for a student, it provides limited view to an instructor on frequently asked questions. A potential way to aid manual identification of common or frequently asked questions, in such forums is to employ clustering, so that semantically related questions are grouped together.

Motivating Example: Table 1 lists examples of sample groups of similar questions posed by middle and high school students on Khan Academy[3]. These groupings or question clusters can help an instructor identify key concerns or confusions among students. The instructor could address confusions by providing additional content on the specific topic. For example, many students are asking questions on the slope of vertical or horizontal line. Having a view of question clusters, can be valuable to the instructor and help in refining course content.

Partition-based clustering methods such as k-means, k-mediods, k-means++ [9] need prior information about the number of clusters required. Providing number of clusters as input can be very hard for the instructors. Hence, in this work we use hierarchical clustering [9] that does not have an input requirement. Dendrograms (a tree of clusters), that capture results of hierarchical clustering, can allow instructors to extract clusters of different granularities without having to re-run the clustering algorithm. Further, most algorithms of hierarchical clustering, provide the flexibility to choose a distance metric that we utilize in this work.

Existing work on processing CQA archives, identify or rank similar questions given a new question [12]. While the problem of estimating relevance of questions to address a new question is a related to estimating similarities between questions to identify clusters, much of the work done to address

[1] www.stackoverflow.com
[2] www.quora.com
[3] www.khanacademy.org

Table 1: Examples of frequently asked questions.

C#	Video Lecture	Student Questions
C1	Graphing a line in slope intercept form	What would the line look like if the slope was a zero?
		What is the slope of a horizontal line?
		what about vertical lines? do they have slope?
		Would a vertical line imply an undefined slope, and would a horizontal line imply a zero slope?
C2	Proof of Limit $sin(x)/x$	Why not use L'hopital's rule?
		can you use l'hopital 's rule to prove this limit ?
		you can also use l'hopital 's rule to turn $sinx/x$ turn into into $cosx/1$
		Can you also prove this limit using L'Hopital's rule?
		Just use l'hopital's rule for that ...$sin(x)/x == cos(x)/1$ and $cos(x)$ for $x- > 0 = 1$

the former problem, uses supervised learning approaches that require labeled datasets for training and building models.

Our Contributions: We address the problem of inferring frequently asked questions (FAQ) by harnessing a distance metric that that uses the similarity of the words in the question using a lexical database (such as WordNet[4]) and the word embedding space representation that depicts contextual similarity of words. We further provide a flexible way of cutting the output of the clustering algorithm, *dendrogram*, allowing the end user to identify clusters of questions. A range, specifying the number of points needed to define a cluster is taken as input. The generated clusters are sorted by the distance metric, thus enabling instructors to filter and identify relevant question clusters.

2. RELATED WORK

In this section we position our work in the context of existing literature along two directions: (1) Analyzing textual content available in student discussion forums, (2) Processing questions in community-based question answering (CQA) systems.

2.1 Student Discussion Forums

There has been a growing body of research on analyzing the textual discussion forum data in Massively Open Online Courses (MOOCs).

A precursor to analyzing questions is determining the utterance of students or classifying the dialog act of the students (such as asking questions, giving feedback or agreeing and disagreeing). Ezen-can et al. [4], apply k-medoids clustering algorithm and qualitatively evaluate the clusters to group dialog acts and topics. In our work, we analyze posts that are categorized as questions. Topic analysis of MOOC discussion content using Structural Topic Model (STM) has been explored by Reich et al. [15]. While topic labels are useful in providing a broad overview of the themes that are attracting student discussions, they do not help the instructor in analyzing finer details of what students are asking or answering. In one of the recent work Thushari et al. [2], present a 'topic-wise organization' of discussion posts by using Latent Dirichlet allocation (LDA) on the discussion data. The authors present a topic visualization dashboard that would assist MOOCs staff in understanding emergent discussion themes or identifying popular topics [1]. Our work uses questions in the student question answering forums and evaluates the semantic similarity between pairs of questions to identify similar question clusters. The work presented here can be used on the subset of discussion posts that have been tagged or organized into a topic.

In addition, discussion forum data has been utilized for a wide variety of purposes, recent among these is the analysis of information seeking behavior of students (that includes querying, refining the query, reading and browsing), while they learn programming [8]. Sentiment analysis in discussion forums [18], examining relationship between students' discussion behaviors and their learning [17] [6], explore various possibilities of using the forum as a rich source of data.

2.2 Community Question Answering (CQA)

The popularity of CQA indicates that users find them useful in finding answers to their questions. However, there are several issues related to CQA that has led to a large body of research: 1) Identifying good and relevant answers to questions can help users filter noise in the responses. 2) Identifying questions that may be repeated or closely related to previously asked questions can help eliminate redundancy. The latter issue, relates very closely to the problem we address in our work.

One of the recent tasks in SemEval 2016 [12] dealt with identifying and ranking a set of 10 related questions given a new question. The participating teams in the task, built supervised machine learning models that used distributed representation of words, knowledge graphs to define lexical and semantic features [5], neural network approaches including convolution neural nets (CNN) or Long short term memory (LSTM) networks [11], [16], [13]. The focus of their work is to rank the questions in a relevant manner considering semantic similarity. A prerequisite to using these approaches in practice, is the need of a labeled dataset. In our work, we use an unsupervised method that circumvents the need for labeled data.

Clustering questions answers (QA) from the CQA systems to ease tasks such as tagging has been less explored. In one of the recent works [14], the authors identify clusters of related QA. The approach is based on classical k-means clustering algorithm, but mixes the similarities of the questions and answers to define an objective function that is optimized over

[4] https://wordnet.princeton.edu/

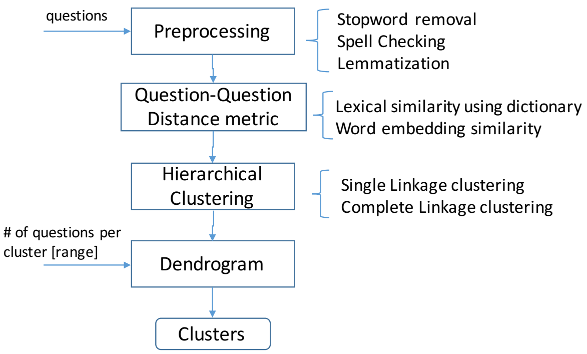

questions
Preprocessing
Stopword removal
Spell Checking
Lemmatization
Question-Question Distance metric
Lexical similarity using dictionary
Word embedding similarity
Hierarchical Clustering
Single Linkage clustering
Complete Linkage clustering
# of questions per cluster [range]
Dendrogram
Clusters

Figure 1: Identifying commonly asked questions.

multiple iterations. While our goal is to cluster questions and use an unsupervised model, we do not rely on the answer information, primarily because the answers given by peers students may contain irrelevant information, especially with students from middle school.

3. IDENTIFYING COMMON QUESTIONS

Our method to infer or identify commonly asked questions is organized into multiple steps, as shown in Figure 1. The first step deals with preprocessing the question to remove any noise. Next, we focus on the key aspect of any clustering algorithm; the choice of (dis)similarity function or distance metric between a question pair. The hierarchical clustering algorithm uses the distance metric to derive the output as a dendrogram. Finally, the dendrogram is partitioned and the clusters are identified.

3.1 Preprocessing

In the preprocessing phase, for each question we filter all URL, email addresses or other similar such patterns which may be irrelevant in the context of the data being analyzed. The misspellings are corrected using the WordNet database. Stopwords are removed and the remaining words in each question are lemmatized to their base forms using the lemmatizer provided by Stanford Core NLP parser[5]

3.2 Question-Question Distance Metric

The distance function uses the combination of both the lexical and word embedding similarity. We define the distance metric between question pairs q_i, q_j as follows:

$$dist(q_i, q_j) = ((\Omega \cdot D_{bow}(q_i, q_j))^x + ((1-\Omega) \cdot D_{vec}(q_i, q_j)^x)^{1/x} \quad (1)$$

where, $D_{bow}(q_i, q_j)$ is the distance computed based on the lexical similarity and D_{vec} is the distance computed based on word embeddings for question pair (q_i, q_j). The following section describes the distance metrics in detail. The distance function Ω is the weight associated with lexical or word embedding based distance. As stated by the authors in [14], the metric represented as $(a^x + b^x)^{1/x}$ approximates to $max\{a, b\}$ for high positive values of x and to $min\{a, b\}$ for high negative values of x.

3.2.1 Lexical Similarity

Each question is represented as a bag of words vector. The dimension of the vector being the vocabulary size of the question corpus W. Each word w_i in the question and its associated synonyms are identified from the WordNet lexical database. The words are weighted by their *idf* measure. The *idf* measure is given by

$$idf(w_i) = log\left(\frac{|D|}{df(w_i)}\right) \quad (2)$$

where, D is the corpus size and $df(w_i)$ is the number of documents containing w_i. Similarity between two question $Sim_{bow}(q_i, q_j)$ is computed using the cosine similarity of the question vectors. The distance is defined as:

$$D_{bow}(q_i, q_j) \equiv 1 - Sim_{bow}(q_i, q_j) \quad (3)$$

3.2.2 Word Embedding Similarity

Each question is represented as a weighted combination of embeddings of words in the question. The word vector v_w for each word w in the question is identified using the distributed representation of words generated by the word2vec tool [10]. Each question q is represented as:

$$V_q = \frac{1}{|q|} \sum_{w \in q} log(\frac{|D|}{df(w)}) \cdot v_w \quad (4)$$

Similarity between two question $Sim_{vec}(q_i, q_j)$ is computed using the cosine similarity of the question vectors. The distance between question pairs q_i, q_j is defined as:

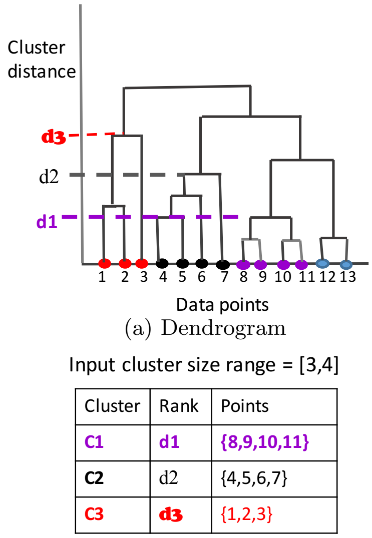
$$D_{vec}(q_i, q_j) \equiv 1 - Sim_{vec}(q_i, q_j) \quad (5)$$

3.3 Hierarchical Clustering

We use agglomerative hierarchical clustering. Initially, each question is in its own cluster. The nearest clusters are merged until there is only one cluster left. The end result is a cluster tree or dendrogram. The tree can be cut at any level to produce different clusters. There are two types of clustering methods. The *Single Linkage* approach, merges two clusters by considering the minimum distance between the points in clusters to be merged. In *Complete Linkage* approach, two clusters are merged by considering the maximum distance between the points in the clusters. Complete linkage clustering results in more compact clusters as the merge criterion considers all points in the cluster. We use complete linkage clustering. The worst case run time complexity of agglomerative clustering is $\mathcal{O}(n^2 \log n)$ which makes it too slow for large datasets. The primary advantage of the clustering approach is that it does not require any prior input to generate the cluster tree.

We evaluated another clustering algorithm Density-based spatial clustering of applications with noise (DBSCAN) [3], which has a worst case run time complexity of $\mathcal{O}(n^2)$. The inputs to the DBSCAN, are the minimum number of points to form a cluster and the distance threshold *eps* such that, for every point in the cluster, there exists another point in the same cluster whose distance is less than the *eps*. Selecting distance threshold as an input can be a challenge. The resulting clusters can vary significantly with *eps*.

[5] http://stanfordnlp.github.io/CoreNLP/



(a) Dendrogram

Input cluster size range = [3,4]

Cluster	Rank	Points
C1	d1	{8,9,10,11}
C2	d2	{4,5,6,7}
C3	d3	{1,2,3}

(b) Resulting clusters

Figure 2: (a) Dendrogram (b) Clusters identified for input range of number of points.

3.4 Dendrogram

The output of the hierarchical clustering is a dendrogram as shown in Figure 2(a). A typical approach is to cut the dendrogram at a specific distance and identify the resultant clusters. However, a dendrogram can be cut at different distances based on the domain or application specific information. In our scenario, an important input from the instructor, is the minimum number of points or questions in cluster, for it to be considered as a FAQ. An instructor may decide, that she would like to address groups of at least 4 similar questions, or provide a range of question sizes as input. Figure 2(b) depicts such a scenario of wanting a range of [3, 4] questions in each cluster. We use number of questions as the input and provide a list of question clusters sorted by the cluster distance. Hence, clusters that are linked with lower distance values form good quality clusters. As the distance function increases, the quality of the resulting cluster would be poor.

4. EXPERIMENTAL EVALUATION

In this section, we evaluate our method for identifying FAQ. We use a labeled data set from a CQA archive and create reference clusters.

4.1 Data

To evaluate the suitability of our approach, we use SemEval 2016 Task 3 dataset that contains questions and answers from Qatar Living forum [12]. The data relevant for our evaluation contain questions categorized as *Original question*. For each original question, a set of 10 related questions are annotated as *PerfectMatch*, *Relevant* and *Irrelevant*. Using the labeled information, we build a set of reference clusters or ground truth, which contain the original question and the related questions that are either *PerfectMatch* or *Relevant*. Table 2 contains the details of the data set. The test dataset contained of 770 questions.

Table 2: SemEval 2016 Task3 dataset used.

Questions		Training	Test
Original Questions		200	70
Related Questions	*Total*	1,999	700
	Relevant	606	152
	PerfectMatch	181	81
	Irrelevant	1,212	467
Total		2,199	770

4.2 Evaluation Metrics

The quality of clustering is measured using F-Measure, combining the precision and recall scores used in information retrieval [7]. Each generated cluster C_{gen} is treated as a result of the query and each reference cluster C_{ref} is considered as the desired set of documents or points:

$$precision(C_{gen}, C_{ref}) = \frac{C_{gen} \cap C_{ref}}{C_{gen}} \tag{6}$$

$$recall(C_{gen}, C_{ref}) = \frac{C_{gen} \cap C_{ref}}{C_{ref}} \tag{7}$$

$$F - Measure(C_{gen}, C_{ref}) = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{8}$$

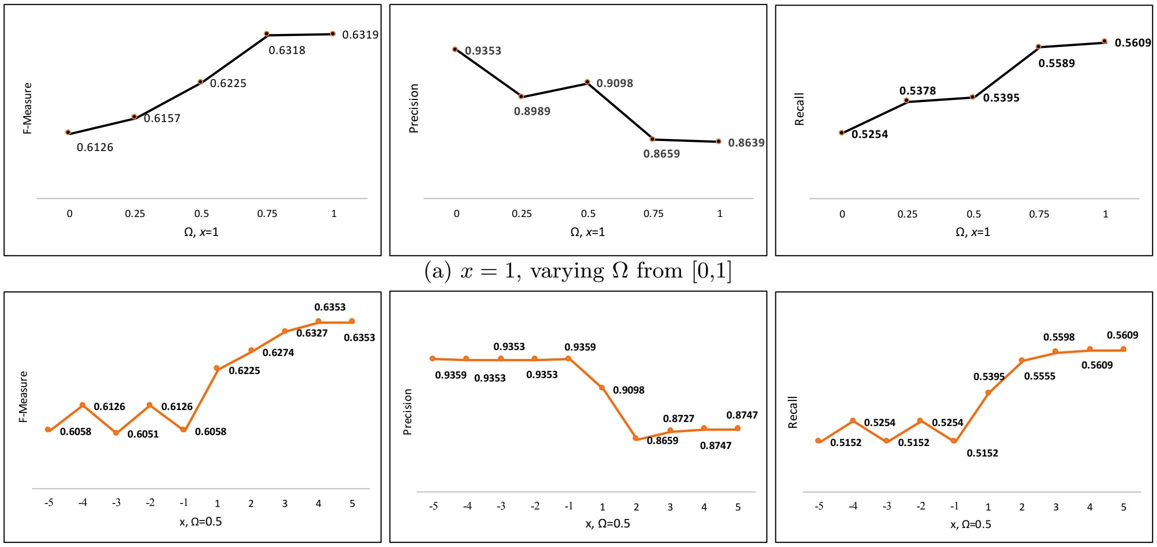
The average precision, recall and F-Measure values are computed for each cluster containing the “original question”. For the purpose of evaluation, we use the test data set and identify the partition or the distance threshold at which the maximum average F-Measure is obtained.

4.3 Results

The results of our approach are presented in Figure 3. We evaluate the cluster measures by considering the question-question distance metric using various values of Ω and x. High F-Measure and recall is achieved when we use lexical similarity as the primary distance metric. Using word embedding as a primary similarity metric results in higher precision, which could be suitable in scenarios where the data is noisy or contains large number of irrelevant questions. Figure 3(a) has varying weights associated to lexical and word embedding based similarity. When $x = 0.5$, a balance between high precision and high recall is achieved. Further, Figure 3(b), shows the metrics achieved by varying Ω. Here, the best results are achieved with $\Omega = 4$, with an F-measure of 0.653, a precision of 0.874 and recall of 0.5609. The SemEval 2016 Task 3 participants reported unofficial precision, recall and F-Measure values. Here, for each original question, *Relevant'* and *PerfectMatch* questions are categorized as true pairs and *Irrelevant* questions are categorized as false pairs. The precision values reported by the top 4 participants ranged from 0.636 to 0.763. The recall values were higher and ranged from 0.553 to 0.759. The F-Measure was between 0.64 and 0.71. The results of our method are comparable and encouraging as we have used an unsupervised model.

5. INFERRING FAQ FROM STUDENT QA SYSTEM

In order to verify the relevance of the approach, we ran the clustering tool on a student question answering platform. The dataset for the analysis, was extracted from the Khan Academy, by permission, using screen scapping pro-



(a) $x = 1$, varying Ω from [0,1]

(b) $\Omega = 0.5$, varying x from [-5,5]

Figure 3: F-Measure, Precision and Recall values by varying Ω and x.

Table 3: Sample FAQ inferred using proposed method from Khan Academy question answering forum.

C#	Video Lecture	Student Questions
C1	Graphing a line in slope intercept form	what do the b stand for in the equation y = mx + b ?
		what do the m stand of in y = mx + b ?
		why do we use m and b in the equation y = mx + b ?
		what if the m and the b be zero in the equation y = mx + b ?
C2	Introduction to limits	isn't 0/0 indeterminate not undefined
		Sal said that 0/0 is undefined. Shouldn't it be not a number?
		At 1:18, why is 0 divided by 0 undefined? My teacher taught us it's 0...
		is 0/0 undefined, or one? and Why?
		I thought that 0/0 is called a indeterminant not undefined. Correct my logic please
		WHY is anything divided by 0 considered as undefined??
C3	Definition of function	I'm trying to understand but, I see what he is doing but what ever he is saying is in slow motion so I don't understand. And what is a piecewise function
		Do you have a video where they give you a graph of a piecewise function, but need to find the rule?
		How to find inequalities for piecewise functions?
		How do you graph piecewise functions?
		what is a piecewise function?
C4	Proof of sin x by x	i m a class 9 student and dont have 100% knowledge on trigonometry (just went through his videos once) so i dnt get what i am missing here: should he prove that for 3rd and 2nd quadrant as well?!
		Is this statement is not applicable to 2nd &3rd quadrants ? Why?
		exactly why does this only apply to 1st and 4th quadrant why not, 2nd and 3rd?
		what about the 2nd and 3rd quadrants?
		X would not be negative in the 4th quadrant.,x is only negative in 2nd and 3rd quadrant.
		why is he working in the first and fourth quadrants only? because the absolute value remains the same in all quadrants
		@14:22 Khan says that cos(x) is always the x value in the first and fourth quadrants. Doesn't he mean that cos(x) and x have the same sign in the first and fourth quadrants?
		Why do we consider x only in the first and the fourth quadrant? Does it change the result if we need to consider all the quadrants?
		I feel like I understand everything except going into the fourth quadrant. From 8:32 to the end of the video, he is discussing the fourth quadrant.
		Why go into the fourth quadrant, and why does he stay away from the second and third quadrant?
		why is he working in the first and fourth quadrants only? because the absolute value remains the same in all quadrants

tocol. We considered micro lectures of 8^{th} grade mathematics and micro lectures covering differential calculus. On the learning platform, each micro lecture video has easy access to the page where questions for that lecture, can be asked or viewed. Asking questions is voluntary. Each learner can view questions that have been previously asked by their peers. Once a question is asked, a discussion thread is initiated with peer students providing answers. The data set contains about 22000 questions from 300 video lectures. As questions are asked in the context of a given micro lecture, we infer the FAQs for each lecture. This helps us reduce the running time of our clustering algorithm.

5.1 Discussion

Table 3 presents a subset of the clusters or FAQs extracted. Four example clusters or FAQ are presented. We were able to extract 4 to 10 questions in each of the sample clusters. We observed several clusters with irrelevant questions, that resulted from poor semantic match when the question content contained numerous mathematical expressions, symbols and less text. Our results can improve with domain specific preprocessing. The current preprocessing step does not parse or process mathematical expressions. Identifying expressions and tagging them as a special tokens for computing question-question distance could provide better results. We noticed several abbreviations in the questions, that were not handled by our preprocessing step. In addition, many students had questions related to content presented at specific time periods in the video lectures. Annotating terms representing video lecture time period, as a part of preprocessing could help ascertain intervals of time within the lectures, where students are seeking more information. Such domain specific processing of content in questions could help improve the question-question distance metric and reduce noise in the generated clusters.

6. CONCLUSION

Our goal in this work was to identify FAQ from the question answering systems of online learning environments. We used agglomerative clustering, an unsupervised learning approach, to identify the FAQ as it did not require any prior inputs to identify groups of questions. A distance metric was defined to harnesses similarity based on bag of words and word embeddings. Our empirical evaluation on labeled dataset shows the effectiveness of our approach, with the precision and F-Measure values comparable to the existing methods that use supervised models. We extracted questions asked by students from Khan Academy and FAQ was extracted for each topic. In future, we would include the answers provided by students in identifying similar questions. The answers can be filtered based on the votes received, student popularity and other related answers in the posts. This would result in improving the quality of extracted FAQ.

7. REFERENCES

[1] T. Atapattu, K. Falkner, and H. Tarmazdi. Topic-wise classification of MOOC discussions: A visual analytics approach. In *International Conference on Educational Data Mining*, 2016.

[2] T. Atapattu and K. E. Falkner. A framework for topic generation and labeling from MOOC discussions. In *ACM Conference on Learning @ Scale*, 2016.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*, 1996.

[4] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth. Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In *International Conference on Learning Analytics And Knowledge*, 2015.

[5] M. Franco-Salvador, S. Kar, T. Solorio, and P. Rosso. UH-PRHLT at semeval-2016 task 3: Combining lexical and semantic-based features for community question answering. In *International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, 2016.

[6] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders. Superposter behavior in mooc forums. In *ACM Conference on Learning @ Scale*, 2014.

[7] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *International Conference on Knowledge Discovery and Data Mining*, 1999.

[8] Y. Lu and S. I. Hsiao. Seeking programming-related information from large scaled discussion forums, help or harm? In *International Conference on Educational Data*, 2016.

[9] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[10] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.

[11] M. Mohtarami, Y. Belinkov, W. Hsu, Y. Zhang, T. Lei, K. Bar, S. Cyphers, and J. Glass. SLS at semeval-2016 task 3: Neural-based approaches for ranking in community question answering. In *International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, 2016.

[12] P. Nakov, L. Màrquez, A. Moschitti, W. Magdy, H. Mubarak, A. A. Freihat, J. Glass, and B. Randeree. Semeval-2016 task 3: Community question answering. In *International Workshop on Semantic Evaluation, SemEval@NAACL-HLT*, 2016.

[13] H. Nassif, M. Mohtarami, and J. Glass. Learning semantic relatedness in community question answering using neural models. *Association for Computational Linguistics*, page 137, 2016.

[14] D. P. Mixkmeans: Clustering question-answer archives. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

[15] J. Reich, D. Tingley, J. Leder-Luis, M. E. Roberts, and B. M. Stewart. Computer assisted reading and discovery for student generated text in massive open online courses. *Journal of Learning Analytics*, 2015.

[16] S. Romeo, G. D. S. Martino, A. Barrón-Cedeño, A. Moschitti, Y. Belinkov, W. Hsu, Y. Zhang, M. Mohtarami, and J. R. Glass. Neural attention for learning to rank questions in community question answering. In *International Conference on Computational Linguistics*, 2016.

[17] X. Wang, D. Yang, M. Wen, K. R. Koedinger, and C. P. Rosé. Investigating how student's cognitive behavior in MOOC discussion forum affect learning gains. In *International Conference on Educational Data Mining*, 2015.

[18] M. Wen, D. Yang, and C. P. Rosé. Sentiment analysis in MOOC discussion forums: What does it tell us? In *International Conference on Educational Data Mining*, 2014.