I.J. Education and Management Engineering, **2020, 3, 33-41**
Published Online June 2020 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2020.03.04

Available online at ***http://www.mecs-press.net/ijem***

Text Summarization using QA Corpus for User Interaction Model QA System

K.Karpagam[1], A. Saradha[2], K.Manikandan[3], K.Madusudanan[4]

[1]*Dr. Mahalingam College of Engineering & Technology, Pollachi*
[2] *Institute of Road Transport and Technology, Erode*
[3] *Vellore Institute of Technology, Vellore*
[4]*Dr. Mahalingam College of Engineering & Technology, Pollachi*

Received: 09 March 2020; Accepted: 25 March 2020; Published: 08 June 2020

Abstract

Document summarization is capable of generating user query relevant, precise summaries from the original document for user needs. To reduce the response time summary generation, QA corpus is built for similar questions and answer with help of learning model. It has been trained and tested by Quora duplicate and Yahoo! Answer datasets. The large QA corpus has been dynamically clustered with semantic features paves a way for efficient document's retrieval. Answers are produced from datasets or generate summaries for unanswerable from the available sources. Results obtained from statistical significance test with hypothesis testing and evaluation with standard metrics proves the significant improvement in generating text summarization using QA corpus. The outcome is better in the producing close proximity of answers for the given user query.

Index Terms: Question Answering System, QA corpus, Text summarization, Machine Learning.

* Corresponding author.
E-mail address:

1. Introduction

Recent days community question answering system (CQAS) obtain a popularity in QA websites like stackoverflow, Quora, etc. The CQAS focus on complex question why question which resultant in a descriptive answer to user queries. Conventional information retrieval techniques works on displaying precise relevant answer instead of large list of documents for the complex questions [8]. Example complex questions are Why corona virus spread more faster in china?, Which is the best hospital in Chennai city? .These type of complex questions cannot be answered in a single sentence, descriptive answers will be formed from list of unique sentences. Challenges of complex questions have been addressed with machine learning, deep learning, etc. To achieve this, text summarization techniques involved. Text summarization is the process of reduction of long pieces of text document into coherent and fluent summary without changing the meaning. Machine learning models are used to train documents understanding and generating relevant summarized texts. Summarization quality is arriving by figure out semantic similarity between query and answers [11].

User interaction model is the improving user experience with the system through the well- defined user interface. Familiar types of events such as mouse clicks, touches and keyboard events support in user interaction. Human-Computer Interaction (HCI) helps in understanding more about user view in interaction with the designed system.

Question Answering System (QAS) processed through question classification, Information retrieval and information extraction for precise answer extraction. QAS systems are categorized into open-domain QA, natural language QA, Community QA and closed domain QA. Open-domain QA system supports in answering all types of queries irrespective of domain. Natural language QA focuses on queries submitted in natural language for answer production through Natural language understanding and natural langue Natural Language Generation. Closed domain QA systems is capable of answering questions related to specific domain such as commercial, education, music, weather forecasting, tourism, medical health etc.Community QA focuses on complex question which needs descriptive answers. The QA corpus is needed to reduce the search space and answer generating time by considering the stored answers in the knowledgebase.

Proposed Systems will first decide whether the question is sufficiently defined for answerable with many false assumptions. When the questions are ambiguous, it’s difficult to be answered succinctly. Answer identification task requires deeper level of language understanding in searching short answers when relevant document are known.

2. Related Works

The background study of QA corpus is deeply arises explored with existing studies with techniques. Several researchers’ works in this proposed techniques for document/ text summarization with different views in increasing output efficiency. The study is carried out in this manner for generating answers using text summarization produced by redundant information, sentences with uniqueness & limited number of words and maximizing the summary relevancy. Title, sentence weight, term weight, sentence position, inter-sentence similarity, proper noun, thematic word and numerical data are considered as features for text summarization[1].In paper[2] proposed a deep learning hybrid model for complex question with convolution and recurrent neural networks for passage-level question answer matching, representation with semantic relations and results are evaluated with datasets. In paper[3] described the solution for handling complex questions is crowd sourcing. The system gets real-time assistance from the crowd to receive answer, validate and rank them accordingly.

In paper [5] authors described to avoid the redundancy in auto text summarisation of text in answering complex questions by measuring Maximal Marginal Relevance (MMR) metric. It increases the diversity of documents compared to traditional feature-based summarisation approach. In paper [6] authors has the goal of compare human generated and machine generated summaries by different text summarization methods. The human produced summaries are obtained from English teachers and automatic summaries are obtained using Fuzzy method and Vector approach of machines.

In paper [12] discuss on several optimization methods had been adopted to reducing the time consuming process and produces approximate the Global best solution artificial bee colony [13], Genetic algorithms [14] and cuckoo search optimization [15].Most of the researchers discuss on single-text summarization system that produces patent summaries using extractive summarization.

The study of related works states that machine generated summaries uses QA corpus for consolidating unique sentence. The research gap found that the state of art of reducing time, false negatives and increases the true positive related sentences. To bridge the research gap, proposed system first focuses on generating automatic summaries from QA corpus. Next focus QA corpus development of similar questions with their respective answers to save users time. The proposed architecture with description and results are discussed in following sections.

3. Proposed System Architecture

Main objective of the proposed system is to generate the answers for factoidal questions from QA corpus and knowledgebase. The proposed system consists of various phases involved in Question Answering System development are question type processing and construction of QA corpus for answer generation. User query is input as natural language through user interface in Question processing phase. The question type is analyzed for its type in two manners. First it checks for factoidal questions using grammatical structure and keywords ‘what’, where’, ‘who’, ’when etc. If the condition is true for factoidal question the answer is searched in QA corpus. If answer found in QA corpus, then it is displayed. When the answer is not found in the QA corpus, answer is generated from knowledgebase. The following Architecture for QA corpus is shown in Figure 1.

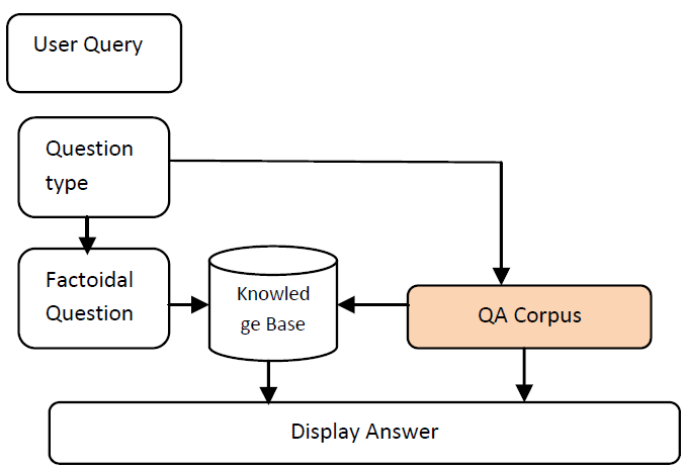


Fig.1 Architecture for QA corpus

In paper [18], [19] & [20] discuss on query is pre-processed using tokenization, stop words removal and stemming to extract keywords. Question type considered are what, when, why, which are trained using question classifier. The proposed system identifies the question types using question pattern template. Knowledge base is developed from the benchmark dataset Quora duplicate and Yahoo! Answer datasets. Document clusters are dynamically generated based on the syntactic and semantic similarities with the query keywords for storing QA pairs. The query from the user is compared with the past similar historical Questions. Multi-similar questions with answer are used to build QA pair corpus to reduce response time for candidate summaries withdrawal.

BUILDING QA CORPUS CREATION

This phase of system is dealing to enrich the knowledgebase with a QA corpus which consists of questions and answer pairs. The same questions can be enquired in different methods, but answers will be similar for those questions. Response time is reduced by identifying the similarly asked questions and stored with their related answers which improve efficiency of retrieval system.

Similar questions asked in different manner have the isolated answer. Examples of similar questions is how can I be a good musician?. & What should I do to be a great musician?. Both the question has the same answer. After identifying similar questions, QA corpus is built with similar questions and their answer. The answer generated from large data corpus is dynamically clustered with semantic features paved a way to effective retrieval of answer. If relation is not detected among the query and answer sentences, context analysis is ensured and generated the answers. If answer generation desires to use external resources such as users, subject experts and net sources in cases of answer is not found for fulfilment [21].

The received user query is matched with Question in QA corpus for its similarity such as keyword, length, irrespective of the word order, semantic similarity, etc. Uniqueness of the question is indicated with Boolean identifier values 0 or 1. Proposed system find similar question with user query and display the answers from QA-Corpus. If similar question is not found in QA corpus, the candidate answer for query is generated from the knowledgebase by using scoring methods. The user interface has been provided to get the answers for the user query. The top rated candidate answers are evaluated, stored in and stored in QA corpus for future use as QA pairs.

4. Results And Discussion

Datasets

Answers are populated for questions from the benchmark datasets Yahoo! Answers and Quora. Performance of QA corpus is evaluated with Yahoo! Answer dataset consists of 189,467 questions and answer pairs from 20 top-level categories with about 10,000 question/answer pairs per category. All level categories are refined with the labels of question/answer pairs.

To accomplish this, a learning model has been developed to find similar questions. It was trained and tested using benchmark Quora duplicate QA pair dataset (http://qim.fs.quoracdn.net/quora). The training for finding similar question is given by Quora duplicate dataset with totally 5, 37,933 questions with 4, 04, 289 QA pairs.

User Interaction Model for Similar Questions

User interaction model is used to implement the language modelling to map the user and answer keywords to recommend related questions for future usage. It is for similar question is experimented with 100 user input questions which analyze question in QA corpus for similarity. Given user set U = {u1, u2, u3,….un} on question set Q = {q1, q2, q3,…, qn} with a score gives the degree of interest in user u to the question q.

Experiment for finding similar question are converging in 5 iterations and observed that on are average from 95 user questions 57 questions are found to be similar with the QA pair.

A qualitative analysis for each user question is conducted concerning most similar questions for better results. By examining the proposed system, when the questions are direct and clear it is found to be more appealing and appropriate[21] . It is examined that the efficiency is reduced when the query is too long, clamped with lack of data. The histogram depicts the efficiency results about the number of similar questions DB along with user asked question as shown in Figure 2.

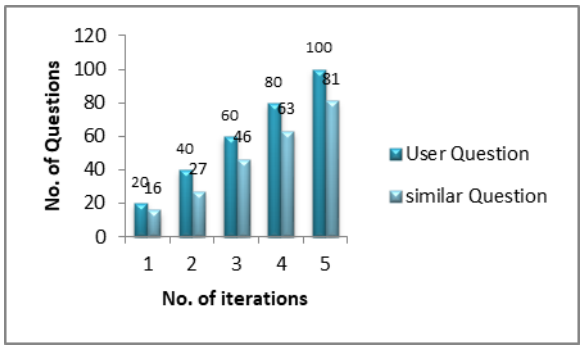
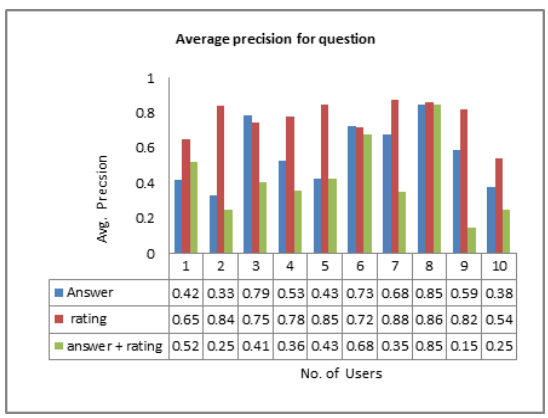


Fig 2 User Interaction Model for Similar Questions

The experimental results on answer rating by crowdsourcing are carried out with 100 sample question and answers. The proposed system works on average of answer rating with 10 crowd users taken for consideration. It is also tested for accuracy with user rating on two scale rating (rank 1 & rank2) for best answer selection. The significance test for answer aptness reaches the efficiency of 81% with error tolerance level of 0.05%.

Average precision for questions and average precision for answers are calculated to analyze active user participation on outcome on query answering, rating on answers generated and users who response on both answering and feedback rating. Totally sample 10 crowd users are considered for taking this survey report.

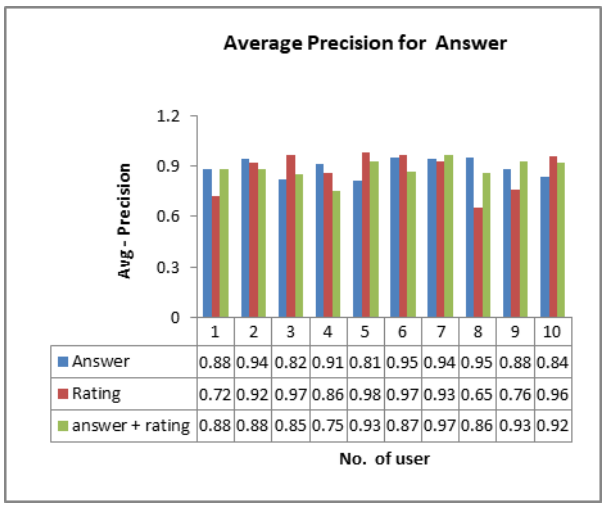
Avg-Precision for questions which is used to calculate the average score of all question asked by the user. Avg-Precision for questions is calculated for 10 crowd users with their related answer, rating and answers cum rating as shown in Figure 3.



	1	2	3	4	5	6	7	8	9	10
Answer	0.42	0.33	0.79	0.53	0.43	0.73	0.68	0.85	0.59	0.38
rating	0.65	0.84	0.75	0.78	0.85	0.72	0.88	0.86	0.82	0.54
answer + rating	0.52	0.25	0.41	0.36	0.43	0.68	0.35	0.85	0.15	0.25

Fig 3 Avg-Precision for questions

The Avg-precision for answers is used to calculate the average score for all answered question to the user by 10 crowd users along with their answer, rating and answers cum rating. Figure 4 shows the Avg-precision of answers from 10 users with their answer, rating and answers cum rating.



	1	2	3	4	5	6	7	8	9	10
Answer	0.88	0.94	0.82	0.91	0.81	0.95	0.94	0.95	0.88	0.84
Rating	0.72	0.92	0.97	0.86	0.98	0.97	0.93	0.65	0.76	0.96
answer + rating	0.88	0.88	0.85	0.75	0.93	0.87	0.97	0.86	0.93	0.92

Fig 4 Average precision for answer

5. Performance Evaluation

The effectiveness of algorithm/methods is measured based on heuristics how much close to the right answer. Mean Average Precision (MAP) is used to evaluate the rank of retrieved relevant documents with assumption that user is interested in finding many relevant documents for each query [23].

$$MAP = \frac{1}{n} + \sum_{Q_i} \frac{1}{R_i} \sum_{Dj \subseteq Ri} \frac{j}{r_{ij}} \quad (1)$$

Where n is the number of test questions, r is the rank of the jth relevant document in Qi and R is the relevant document for Qi .

Mean Reciprocal Rank (MRR) is a statistic quality measure for evaluating single highest-ranked relevant candidate answers to a set of sample queries.

$$MRR = 1/|Q| \sum_{i=1}^{Q} \frac{1}{rank_i} \quad (2)$$

Where $rank_i$ denotes rank position of the first relevant document for the ith query.

The quality of answer pool is determined by MMR (Maximal Marginal Relevance) an iterative method for content selection from single documents which maximize relevance and less redundancy in automatic summarization (Dorota Gowacka et al., 2013; Jan Frederik Forst et al 2009).The related information for query is taken from different sources but the documents/paragraph occurrence is redundant. Highly relevant documents are more useful than marginally relevant document, so it is used only based on the need and relativity. The low cosine similarity is calculated for non redundant sentences. It evaluates the similarity

between sentences under consideration to append in candidate answer list. It is calculated by empirical formula as

$$MMR = \arg\max_{D_i \in R \setminus S} [\lambda(sim_1(D_i, Q) - (1-\lambda)(\max_{D_j \in S} sim_2(D_i, D_j))] \quad (3)$$

Where D is the document, Q is Query, S is already retrieved sentences, and R is the reference sentences.

Table 1 shows the efficient working of the summarizer with MAP, MMR and MRR values on various datasets.

Table 1 MAP, MRR and MMR for Sentence Selection

No. of cluster	No. of Questions	MAP		MRR		MMR	
		DUC (2001)	20News group	DUC (2001)	20News Group	DUC (2001)	20News group
10	50	0.378	0.3692	0.307	0.357	0.348	0.329
20	50	0.389	0.375	0.309	0.325	0.312	0.347
30	50	0.381	0.335	0.331	0.352	0.318	0.328
40	50	0.42	0.354	0.327	0.343	0.34	0.349
50	50	0.405	0.365	0.395	0.347	0.39	0.353

6. Conclusion

This proposed work is developed a framework for answering question through machine generated summaries using QA corpus. Learning model was trained by human generated summaries and focused on reducing the response time by developing a QA corpus with similar questions and answer. User interaction modelling is used to analyze the performance for user question and similar question with outcome on average precision per question and average precision per answer. Proposed system efficiency uses benchmark datasets and refined QA corpus and produces precise summary which were evaluated by standard metrics such MRR, MAP, MMR. Obtained result shows that proposed system outer performs than existing system in reducing time and space by initially searching and displaying the answers from QA corpus. Future improvement is to apply deep learning for natural language understanding to the answer dynamically abbreviated questions with set of human annotations.

References

[1].Rasmita Rautray, Rakesh Chandra Balabantaray&Anisha Bhardwaj ., Document Summarization using Sentence Features, International Journal of Information Retrieval Res(IJIRR), Volume 5 Issue 1, pp.36-47, 2015.

[2].Ming Tan, Cicero dos Santos, Bing Xiang & Bowen Zhou, Improved Representation Learning for Question Answer Matching', Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, August, pp. 7-12,2016.

[3].Denis Savenkov & Eugene Agichtein, 'CRQA: Crowd-Powered Real-Time Automatic Question Answering System', Proceedings, The Fourth AAAI Conference on Human Computation & Crowdsourcing , Association for the Advancement of Artificial Intelligence, pp.189 -198, 2016.

[4].Dorota Gowacka, Tuukka Ruotsalo , Ksenia Konyushkova , Kumaripaba Athukorala , Samuel Kaski&

Giulio Jacucci , Directing Exploratory Search: Reinforcement Learning from User Interactions with Keywords, IUI'13, Copyright ACM, pp.117-127, 2013.

[5].Jan Frederik Forst, Anastasios Tombros& Thomas Roelleke, 'Less Is More: Maximal Marginal Relevance as a Summarization Feature', Advances in Information Retrieval Theory, ICTIR, pp. 350-353, 2009.

[6].Farshad Kiyoumarsi, , Evaluation of Automatic Text Summarizations Based on Human Summaries, 2nd GLOBAL CONFERENCE on LINGUISTICS &
FOREIGN LANGUAGE TEACHING, LINELT-2014, Procedia - Social & Behavioral Sciences, ELSEVIER, pp. 83 – 91, 2014.

[7].Gunnar Schröder, Maik Thiele &Wolfgang Lehner , , 'Setting Goals & Choosing Metrics for Recommender System Evaluations ', 5th ACM Conference on Dresden University of Technology Recommender Systems Chicago, October 23th, 2011.

[8].Youzheng Wu, Chiori Hori, Hideki Kashioka & Hisashi Kawai, , 'Leveraging social QA collections for improving complex question answering', Computer Speech & Language, Volume 29 Issue 1, pp.1-19,2015.

[9].Hiroyuki Sakai & Shigeru Masuyama, 'A Multiple-Document Summarization System with User Interaction', In Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, C04-1144, ACL Anthology, pp.1001- 007,2004.

[10].Gang Liu & Tianyong Hao, 'User-based Question Recommendation for Question Answering System', International Journal of Information & Education Technology volume 2, Issue 3, pp. 243-246, 2012.

[11].Karpagam.K & Saradha.A, "Text Summarization using Machine Learning Approaches for Question Answering System", International Journal of Advances in Computer and Electronics Engineering, ISSN: 2456 - 3935, Volume 4, Issue 2, pp.1–5, , 2019.

[12].John, A., Premjith, P.S., Wilscy, M ," Extractive multi-document summarization using population- based multi-criteria optimization", Expert System Application 86, pp.385–39, 2017.

[13].Sanchez-Gomez, J.M., Vega-Rodríguez, M.A., Pérez, C.J.,"Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approaches", Knowledge-Based System, 2017.

[14].Al-Radaideh, Q.A., Bataineh, D.Q, " A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms", Cognitive Computer, 2018.

[15].Rautray, R., Balabantaray, R.C. "An evolutionary framework for multi-document summarization using cuckoo search approach", Mdscsa. Application Computer Inform. 14 (2), pp.134–144, 2018.

[16].Litvak, M., Vanetik, N., Last, M., Churkin, E.. Museec,, A multilingual text summarization tool, pp. 73–78, 2016.

[17].Thomas, S., Beutenmüller, C., de la Puente, X., Remus, R., Bordag, S, "Extractive text Summarizer", Proceedings of the SIGDIAL 2015 Conference, pp. 260–269, 2015.

[18].Abdelkrime, A., Djamel Eddine, Z., Khaled Walid, H. ,"All summarizer system at multilingual single and multi-document summarization" Proceedings of the SIGDIAL 2015 Conference,pp.237–244, 2015.

[19].Litvak, M., Vanetik, N., Last, M., Churkin, E.. Museec," A multilingual text summarization tool", pp. 73–78,2016,.

[20].Thomas, S., Beutenmüller, C., de la Puente, X., Remus, R., Bordag, S,,., "Extractive text Summarizer", Proceedings of the SIGDIAL 2015 Conference, pp. 260–269,2015.

[21].Benjamin Timmermans, Lora Aroya& Chris Welty ,"Crowd sourcing ground truth for Question Answering using Crowd Truth", Web Science, Oxford, United Kingdom, ACM,2015.

[22].Hiroyuki Sakai & Shigeru Masuyama, ,"A Multiple-Document Summarization System with User Interaction", In Proceedings of the 20th International Conference on Computational Linguistics, COLING '04, C04-1144, ACL Anthology, pp.1001- 1007,2004.

[23].Gunnar Schröder, Maik Thiele &Wolfgang Lehner, 'Setting Goals & Choosing Metrics for Recommender System Evaluations ', 5th ACM Conference on Dresden University of Technology Recommender Systems Chicago, October 23th, 2011.

Author's Profile

Dr.K.Karpagam is working as Asst. Professor in Department of Computer Applications, Dr.Mahalingam College of Engineering and Technology, Pollachi. She has 15 years of academic experience and currently pursuing research in Anna University, Chennai. The area of research interests includes Machine learning ,text mining, Question Answering system and information retrieval..

Dr.A.Saradha is working as Professor and HOD, Department of Computer Science and Engineering, IRTT, Erode. She has 25+ years of academic experience. Her research interests are Semantic Web, Human Computer Interaction, Image processing.

Dr.K.Manikandan is working as Associate Professor in the School of Computer Science & Engineering ,VIT Vellore. He received Best Faculty Award-2014 from ICTACT,Govt of Tamilnadu. Research Area includes Mobile Network ,Cloud Computing, Big Data Analytics and Machine learning

K.Madusudanan is working as Asst. Professor in Department of Computer Applications, Dr.Mahalingam College of Engineering and Technology, Pollachi. He has 17 years of academic experience. The area of research interests includes semantic web, text mining and information retrieval.

How to cite this paper: K.Karpagam, A. Saradha, K.Manikandan, K.Madusudanan. "Text Summarization using QA Corpus for User Interaction Model QA System", International Journal of Education and Management Engineering(IJEME), Vol.10, No.3, pp.33-41, 2020.DOI: 10.5815/ijeme.2020.03.04