A Survey of Text summarization in Community Question Answering

Susmitha A
M.tech Student, Computer Science and Engineering
N.S.S. College of Engineering, Palakkad
E-mail: susmithasachu@gmail.com

Remya G
Assistant Professor, Computer Science and Engineering
N.S.S. College of Engineering, Palakkad
E-mail: gremyanair@gmail.com

***Abstract*—With the emergence of numerous community forums, the task associated with the same have gained significance in the recent past. With the incursion of new questions every day on these forums, the matter of recognizing methods to find answers to said questions, or even trying to discover similar questions, are of realistic importance and are challenging in their own right. The task on which we focus in this paper is an extractive summarization method, which produce summaries by choosing a subset of the sentences in the original text. Community question answering classified into two categories :Factoid and Non factoid. This survey deals with some of the aforementioned issues, and methods proposed for tackling the same in the community question answering portals.**

keywords: Community Question Answering(CQA), Extractive Summarization.

I. Introduction

With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents. This availability of documents has urged research in the area of automatic text summarization. Automatic text summarization is the task of producing a precise and fluent summary while preserving key information content and overall meaning. Recently numerous approaches have been developed for automatic text summarization and applied widely in various domains. Community-based Question Answering sites, such as Yahoo! Answers and stack overflow, provide a promising way of finding and sharing information online. This survey aims to improve the usefulness of CQA services towards better searcher satisfaction

Here discuss various methods for implementing text summarization. The automated text summarisation started in early 1950s. Sentence ranking is the major step in selecting important sentences in the document. The most traditional method in sentence ranking is using statistical features derived from the document. The statistical features used includes position of a sentence in the document, similarity with title, sentence length, presence of cue words or phrases, presence of frequent words, presence of words with high tf-idf values, presence of title words, presence of proper nouns, etc. Similar to statistical features, linguistic features are used such as Part-Of-Speech tag. The values for these features are computed for every sentence in the document. The sentences are ranked on the basis of these values. Highly scored sentences are selected from the document to include in the summary. Statistics based methods lacks fluency in the summary being generated. This is because highly scored sentences dispersed in the documents are selected without considering similarity between sentences.

In general, there are two different approaches for automatic summarization: extraction and abstraction. Extractive summarization methods work by identifying important sections of the text and generating the summary. thus, they depend only on extraction of sentences from the original text. In contrast, abstractive summarization methods aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language processing techniques and generate a new shorter text that conveys the most critical information from the original text. Even though summaries created by humans are usually not extractive, most of the summarization research today has focused on extractive summarization. Purely extractive summaries often times give better results compared to automatic abstractive summaries. Abstractive summarization methods deals with problems which are relatively harder than data-driven approaches such as sentence extraction. Some of the problems include semantic representation, inference and natural language generation . Existing abstractive summarizers often depend on an extractive preprocessing component to produce the abstract of the text.

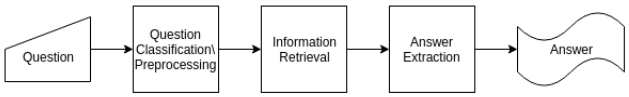
With the emergence of internet technology, people got a platform to share and discuss their ideas among them. Various forums and communities were formed through which a person may ask a question, which would then be answered by many others existing in that community. This is known as Community Question Answering (CQA). There are mainly two kinds of CQA, namely, Factoid CQA and Non-factoid CQA. In the former, facts are asked as questions thus it will have a definite answer, while in the latter, questions which can have different answers and all of it being correct are asked.

Community-based Question Answering (CQA) sites, such

as Yahoo Answers, stack overflow, and Quora , have gained substantial popularity over the recent years, providing an alternative way of online information seeking other than web search. Users dependon community for a variety of reasons, from lack of proficiency in web search to find an answer accompanied by interaction with a real human. Although some of these CQA sites allow payments in response for answering questions, answerers are usually attracted by social rewards and less tangible incentives, such as reputation or points. The CQA communities are mainly volunteer-driven, and their openness and accessibility appeal to millions of users. Accordingly, this survey focuses on extractive summarization methods and provide an overview of some of the most dominant approaches in this category.

II. Community question-answering retrieval

Research into community question-answering (CQA) is expanding due to the popularity of CQA on the web. Community question-answering retrieval can be divided into factoid CQA and non-factoid CQA. In factoid CQA a single correct answer exist . Deep neural network architectures based approaches have been proposed to select the best answer or question. After that convolutional network was introduced for re-ranking pairs of a question and candidate answers. Large question-answer pair corpus generated by applying a novel neural network architecture. Then syntactic function of the terms within CQA texts as an important factor affecting retrieval. And developed a retrieval-based automatic response model for short-text conversation.



CQA answers selected not only by their relevance to the question but also by diversity and novelty of answers. After that a set function is designed to re-rank questions given a user review. Deploy users social networks for inferring a user model, and thus improve the performance of expert finding in CQA. Multimedia question answering has also received attention recently . Several approaches have already been proposed for the non-factoid question answering. Most recent work on non-factoid question answering retrieval focuses on the task of answer passage retrieval. Design of semantic and context features for answer sentence retrieval in non-factoid CQA. Unlike previous research on non-factoid CQA, the aim of answer summarization is to explore all the useful information from candidate answers given a question. very little work on this task has been reported in the recent years.

III. Different approaches for answer summarization

In[1] Li, P., Bing et al. Introduced a sparse-coding-based method that is able to calculate the salience of the text units by jointly considering news reports and reader comments. Another reader-aware characteristic of the framework is to improve linguistic quality via entity rewriting. For the optimization problem of sparse coding, Coordinate Descent method is used in this paper. Phrase Extraction and Salience Calculation is done using a Stanford parser to obtain a constituency tree for each input sentence. After that, noun phrases(NPs) and verb phrases(VPs) are extracted from the tree. The salience of a phrase depends on expressiveness score and concept score. During the preparation of entity mentions for rewriting stage, the coreference resolution for each document using Stanford coreference resolution package is conducted.The actual rewriting follows some operations. The operations will be carried out on the selected phrases output from the optimization component in the post-processing stage by considering entity rewriting, phrase co-occurrence, shortness avoidance, pronoun avoidance and length constraint.

In[2], the problem of comments-oriented document summarization and aim to summarize a Web document (e.g., a blog post) by considering its content and also the comments left by its readers. Comments left by readers on Web documents, such as tags, comments, ratings and others, contain valuable information that can be utilized in different information retrieval tasks including document search, visualization, and summarization. The proposed system identifies three relations by which comments can be linked to one another, namely Topic, Quotation and Mention. Based on these three relations the system derives three graphs : topic graph, quotation graph and mention graph. The importance of each comment is then scored by either Graph-based method or Tensor-based method. Finally, the proposed system extract sentences from the given web document by Feature-biased approach or Uniform-document approach to generate the summary.

In[3] suggest a new method for maximizing the number of informative content-words for best reported results in multi-document summarization. This model has two parts: one uses machine learning techniques to compute scores for each word in the set of documents to be summarized (which is called the document cluster), the second part uses a search algorithm to find the best set of sentences from the document cluster for maximizing the scores. Each sentence is scored as the average of the probabilities of the words in it. The summary is then generated through a simple greedy search algorithm: which select the sentences with the highest-scoring content-word, and average score of the sentences. This process continue until the maximum summary length has been reached. In order to not select the same or similar sentence multiple times, SumBasic find probabilities of each

word in the selected sentence and squaring them, modeling the likelihood of a word occurring twice in a summary.

Chin-Yew Lin[4] ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes various measures which automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. Here introduced four different ROUGE measures: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S which is included in the ROUGE summarization evaluation package. It also includes several automatic evaluation methods that measure the similarity between summaries like ROUGE-1, ROUGE -2, ROUGE -S4, ROUGE -S9, ROUGE -SU4, and ROUGE-SU9, which worked reasonably well when stop words were excluded from matching, exclusion of stopwords usually improved correlation, and correlations to human judgments were increased by using multiple references. It showed that the ROUGE package could be used effectively in automatic evaluation of summaries

In[5], An extract summarization, emphasizing diversity, coverage and balance. An extract-based document summarization system extracts sentences from the original document and uses it in summary. In this work, the system considers three things for an effective summarization : diversity, coverage and balance. The system utilizes structural Support Vector Machine with three types of constraints to enforce diversity, coverage and balance in the generated summaries. Based on the relations between sentences, the system proposes an independence graph to model the structure of a given document so as to represent the dissimilarity between pairs of sentences and reducing the search space in the structure learning process. The system uses cutting plane algorithm to solve the resulting optimization problem and then use the trained model for the summary generation. Several features are considered by the system for summarization. If the corresponding feature holds true, it is set to 1 and if not, it is set to 0.

In[6] Tang, J et al. describe a query-oriented summarization method. Here extract an informative summary from a document collection for a given query. It is very useful for users to grasp the main information related to a query. This method suggests a new setup of the problem of multi-topic based query-oriented summarization. This paper, aim to conduct a thorough investigation on the problem of multi-topic based query- oriented summarization which identify the major tasks of the problem and propose a probabilistic approach to solve the tasks. A statistical topic model is presented to discover multiple topics in a document collection. Two strategies for simultaneously modeling document contents and the query information are discussed. The first strategy directly integrates the query information into the generative process of the topic model. Thus the model estimates a mixture of a document-specific topic distribution and a query-specific topic distribution. The other strategy is to use a regularization form to constrain the topic model by the query information.

A new term weighting method[7] is introduced that makes use of the syntactic information available for each query term occurrence in the document, on top of term occurrence statistics. i.e, weight in the relative importance of the part-of-speech tag and the syntactic role of all occurrences of the matched query terms in the document, effectively summing a syntactic weight for the matched terms. The scoring formula integrates the syntactic information associated with the occurring query terms together with statistical-based measures of the similarity between the document and the query. These syntactic ranking features are Incorporated into the final ranking score of the document, which also includes a rich set of frequency-based scoring features.

Convolutional neural network model[8] can be used for sentence level classification tasks which can be used for both static and task-specific vectors. The basic working principle of CNN is pooling. The proposed system trains a simple CNN with one layer of convolution on the top of word vectors obtained from an unsupervised neural language model. The system initially keeps the word vectors static and learn only the other parameters of the model. Despite little tuning hyperparameters, this simple model achieves excellent results suggesting that the pre-trained vectors can be used for classification tasks. A filter is applied for a window of h words to produce a new feature resulting in a feature map. Max-over time pooling is applied and takes the maximum value as feature corresponding to that filter. These features from penultimate layer are passed over a softmax layer whose output is a probability distribution over labels.

In[9] An unsupervised sentence salience framework for Multi-Document Summarization is used. It can be divided into two components: latent semantic modeling and salience estimation. A neural generative model called Variational Auto-Encoders (VAEs) is used to describe the observed sentences and the corresponding latent semantic representations. The Neural variational inference is used for the posterior inference of the latent variables. For salience estimation, an unsupervised data reconstruction framework, which jointly considers the reconstruction of latent semantic space and observed term vector space, is used.Thereafter, the VAEs-based latent semantic model is integrated into the sentence salience estimation component in a unified fashion, and the whole framework is trained jointly by back-propagation via multi-task learning.

Another work[10] suggest syntactic structure of the original sentence with a language model are trained on the headline data in order to produce a compressed output. The method utilizes a local attention-based model that generates each word of the summary conditioned on the

input sentence. The model can easily be trained end-to-end and scales to a large amount of training data. Potentially the model can also learn to combine words; although it is inherently limited in representing contiguous phrases. This system focuses on neural machine translation and directly parameterize the original distribution as a neural network. Neural Language model is the language model used for estimating the contextual probability of the next word. The system uses a neural network language model with an additional encoder element. Beam-search decoder generates hypothesis at each point in the summary generation and recombines the hypothesis with the existing summary and filters the hypothesis so generated.

In[11] ranking answers by the amount of new aspects they introduce with respect to higher ranked answers, on top of their relevance estimation. Answers are ranked in a greedy manner based on the amount of diverse propositions they contain, taking into account each proposition relevance to the question as well as its dissimilarity to propositions in higher ranked answers. For each answer the relevant propositions are found and extracted. Novelty is assigned to those extracted propositions. Compare the novelty of propositions contained in the answer to rank the answers. The model uses a graph ranking model for retrieving a diverse set of answers from a large answer archive given a complex input question. Redundancy relations among answers are modeled using the maximal marginal relevance algorithm by assigning a negative sign to the weight of edges between the answer nodes in the graph of candidate answers.

Daume et al.[12] proposed BayeSum, a Bayesian summarization model for query-focused summarization. Bayesian sentence-based uses both term-document and term-sentence associations. This system achieved significance performance and outperformed many other summarization methods. The model describe multi-document summarization as a prediction problem based on a two-phase hybrid model. First, present a hierarchical topic model to discover the topic structures of all sentences. Then the similarities of candidate sentences with human-provided summaries are computed using a novel tree-based sentence scoring function. In the second step it make use of these scores and train a regression model according the lexical and structural characteristics of the sentences, and employ the model to score sentences of new documents (unseen documents) to form a summary.

Mattia Tomasoni et al[13]. Describe a metadata-aware measures for answer summarization in CQA. Here introduces a new framework of automatic processing information coming from community Question Answering (CQA) portals with the purpose of generating complete, relevant and trustful summary in response to a question. It utilizes the metadata available in the User Generated Content (UGC) to bias automatic multi-document summarization techniques toward high quality information. The problem is cast as an instance of the query-biased multi-document summarization task, where the question was seen as a query and the available answers as documents to be summarized. Here mapped each characteristic that an ideal answer should present to a measurable property that the final summary could exhibit:

- Quality to assess trustfulness in the source
- Coverage to ensure completeness of the information presented
- Relevance to keep focused on the users information need
- Novelty to avoid redundancy

Quality of the information was achieved through machine earning (ML) techniques under best answer supervision in a vector space consisting of linguistic and statistical features about the answers and their authors . Coverage was estimated by semantic comparison between the knowledge space of a corpus of answers to similar questions which had been retrieved through answer API . Relevance was computed as information overlap between an answer and its question , and Novelty of the result was computed as inverse overlap with all other answers to the same question.

Answer summarization in non-factoid CQA focuses on collecting all relevant and meaningful sentences to the input question from candidate answers. The challenges we face in summarizing non-factoid CQA are :

- Shortness of answers : Candidate answers given by a user can be a single word or phrase, which must then be made into a sentence so as to include it in the summary.
- Sparsity of syntactic and context information : In some cases, users may not be able to provide a soundful and complete answer. Such answers must be made to be sound and complete to include it into the summary.
- Diverse topic distribution : In non-factoid CQA the candidate answers can be from a range of different topics. Thus the summary should be made only after considering all these topics.

In [14] Hongya Songy et al. introduced a new method called sparse coding based summarization strategy. Sparse coding is used to find a linear combination of basis vectors to minimize the reconstruction error function. Cosine similarity between vectors representing each candidate sentence and the question vector is found. Because the summary sentences are sparse, a sparsity constraint is imposed on the saliency score vector. Putting all this together a loss function is generated. Coordinate descent method is utilized to iteratively optimize the target function about the saliency vector until it converges. Given a saliency score for each candidate sentence, maximal marginal relevance (MMR) is applied which incrementally computes saliency - ranked list, and computes a maximal

diversity ranking among candidate sentences. Finally sentences are selected according to their saliency score.

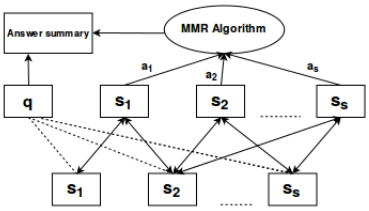


Fig. 1. Sparse coding approach

IV. Conclusion

In this survey, we discussed recent advances in automatic text summarization. In which major categories of CQA, facoid and non-factoid methods are discussed. We received a large body of the existing work, highlighting common characteristics and difference between existing researches. Compared to factoid based summarization non-factoid based methods involves more effort. Traditional document based summarization methods are inefficient for non-factoid based question answering. Sparsity and diversity of answers form a challenge for summarization methods. In this field a sparse coding based automatic summarization method will improve the quality of answers.

References

[1] **Li, P., Bing, L., Lam, W., Li, H. and Liao, Y.**, (2015)." Reader-Aware Multi-Document Summarization via Sparse Coding". *In IJCAI*, pp. 1270-1276.

[2] **Hu, M., Sun, A. and Lim, E.P.**, (2008)," Comments-oriented document summarization: understanding documents with readers' feedback". *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 291-298.

[3] **Yih, W.T., Goodman, J., Vanderwende, L. and Suzuki, H.**, (2007), "Multi-Document Summarization by Maximizing Informative Content-Words". *In IJCAI* , pp. 1776-1782.

[4] **Lin, C.Y.**, (2004), "Rouge: A package for automatic evaluation of summaries. In Text summarization branches out:" ,*Proceedings of the ACL-04 workshop*.

[5] **Li, L., Zhou, K., Xue, G.R., Zha, H. and Yu, Y.**, (2009),." Enhancing diversity, coverage and balance for summarization through structure learning". *In Proceedings of the 18th international conference on World wide web* ,pp. 71-80.

[6] **Tang, J., Yao, L. and Chen, D.**, (2009), . Multi-topic based query-oriented summarization. *In Proceedings of the 2009 SIAM International Conference on Data Mining* ,pp. 1148-1159.

[7] **Carmel, D., Mejer, A., Pinter, Y. and Szpektor, I.**, (2014)," Improving term weighting for community question answering search using syntactic analysis". *In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* ,pp. 351-360.

[8] **Chen, Y.**, (2015). "Convolutional neural network for sentence classification", Master's thesis, University of Waterloo.

[9] **Li, P., Wang, Z., Lam, W., Ren, Z. and Bing, L.**, (2017), "Salience Estimation via Variational Auto-Encoders for Multi-Document Summarization". *In AAAI*, pp. 3497-3503.

[10] **Rush, A.M., Chopra, S. and Weston, J.**,(2015). "A neural attention model for abstractive sentence summarization". *In EMNLP*.

[11] **Omari, A., Carmel, D., Rokhlenko, O. and Szpektor, I.**, (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*, pp. 215-224.

[12] **Daum III, H. and Marcu, D.**, (2006)," Bayesian query-focused summarization". *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* ,pp. 305-312.

[13] **Tomasoni, M. and uang, M.**, (2010)," Metadata-aware measures for answer summarization in community question answering". *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 760-769.

[14] **Song, H., Ren, Z., Liang, S., Li, P., Ma, J. and de Rijke, M.**, (2017)." Summarizing answers in non-factoid community question-answering". *In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* pp. 405-414.