Sentence Ranking and Answer Pinpointing in Online Discussion Forums Utilising User-generated Metrics and Highlights

Sushant Gautam*, Saloni Shikha †, Alina Devkota ‡ and Spandan Pyakurel§
Department of Electronics and Computer Engineering, Pulchowk Campus
Institute of Engineering, Tribhuwan University
Lalitpur, Nepal
Email: *072bct544@ioe.edu.np, †072bct531@pcampus.edu.np, ‡072bct504@pcampus.edu.np, §072bct539@pcampus.edu.np

***Abstract*—One of the major challenges in searching on the internet has been that search engines and online forums have not been able to extract and pinpoint exact answer to people's query despite information being available on the internet. Extraction of to-the-point answers from articles, posts and blogs tend to improve search accuracy. Sentence Ranking helps to rank answers according to a score that represents positive remark for the relevance of sentence. User-generated metrics can be used to improve sentence ranking. Also, the text selected and saved as highlights by users can be used to extract the most important parts of the content. Answer pinpointing in simple forums can be achieved by allowing users to highlight parts of the text, store it in a database and analyse such highlights using sentence ranking engine followed by answer extraction to find the best chunk of texts. It can prove to be a milestone in providing exact and relevant answers as per the searchers' intent and can also facilitate improvement of question answering in discussion forums.**

***Index Terms*—sentence ranking, user-generated content, question answering, user-generated metric, user highlights, answer pinpointing, online discussion forum, engagement metric.**

I. Introduction

Active research is being conducted in the field of question answering (QA) and information retrieval. Intelligent agents and bots are already showing their presence in the global market and are being smarter each day. The results, however, have shown that the progress in this field is yet too far from fulfilling the expectations. The Internet has a massive amount of data but, diverse and unlabelled. That is why the search engines and assistants, in spite of having access to a massive amount of data, have not been able to give users the exact answers to the questions they have been searching for. Also, to address users' immediate information need, it is necessary to have a good information retrieval system. This can be done through the creation of an ideal question-answer system.

In search engines, widely searched questions such as "What is the height of the Everest?" are provided with exact answers. This, however, is not the case with other questions. Even when the answers to the questions searched for are available on the internet, they are not pinpointed. Identifying the precise answer within a long text has thus been a challenge in online discussion forums. Pinpointing the answer requires ranking of the sentences which may possibly contain the answer and extracting it. Different techniques and algorithms aid the process and are in use. Recent researches have shown that neural networks can be used to enhance question-answering systems thus providing users with better search experience.

II. Background

Answer Sentence Ranking and Answer Extraction are the two major challenges in question-answering required for the purpose.

A. Sentence Ranking

Answer sentence selection has always been a topic of interest to researchers in the field of question-answering systems. Answer sentence ranking involves assigning different answers to a question with a rank according to the relevance of the answers. The one that is ranked higher is the one that is more likely to have the answer contained in it (see Fig. 1).

A tag is a label attached to a post for purpose of identification or categorisation that can be several words long and reflects key points of the post. Tags can be either automatically generated from a passage or inserted by users themselves. Tags help to increase search efficiency by finding exact match rather than conventional techniques where strings are searched by matching sub-strings. Characteristics of tags often have a direct relationship with the users' answers. Sometimes hierarchies of tags can be used by nesting related tags into a collapsible list. Tags can also be helpful to answer sentence ranking.

Likewise, one of the most popular meta-data tags used in social platforms, such as Instagram, Facebook, Pinterest and Google+ is the hashtag that allows users to apply dynamic tagging for the purpose of the ease in the finding of posts with specific contents. Hashtags are focused more by viewers but they also serve as links to search queries.

According to Dwivedi and Singh[1], possible approaches for answer ranking are Linguistic Approach, Statistical Approach and Pattern Matching Approach.

1) Linguistic Approach for Answer Ranking: The linguistic approach relies on the use of Artificial Intelligence

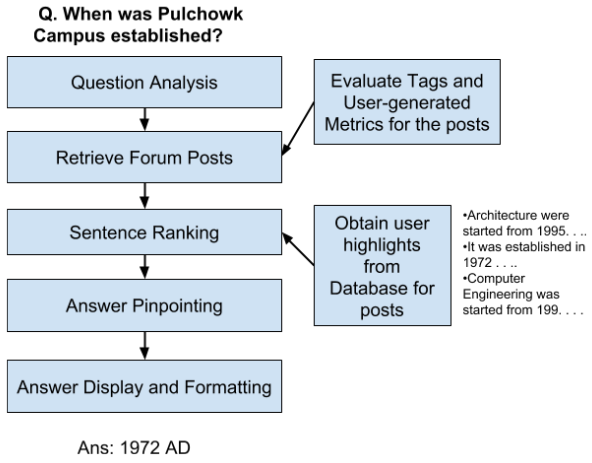

Fig. 1: QA based System Implementation Model

techniques integrating with Natural Language Processing techniques and knowledge base to form question-answering logic. Information organised in the form of production rules, logic, frames, templates, ontology and semantic networks are utilised during analysis of question-answer pair. Sometimes knowledge-based QA systems rely on a rule-based mechanism to identify question classification features.

2) Statistical Approach: This approach deals with a large amount of data and their heterogeneity and is independent of a query language. Support vector machine (SVM) classifiers, Bayesian classifiers, Maximum entropy models are some techniques that have been used for question classification purpose. Pattern matching approach uses text patterns or templates to identify answers.
3) Pattern Matching Approach: This approach uses text patterns or templates to identify answers. For example, the question, "When did world war II end?", follows the pattern "When did <event name> end?" and its answer pattern will be like "<event name> ended on <date/time>". Systems can be made to learn such text patterns from text passages rather than employing complicated linguistic knowledge or tools to text for retrieving answers.

B. Answer Extraction

The answers to questions posted in forums may or may not contain the exact answers to the questions in the thread. Answer extraction deals with extracting smaller parts (which may be in the form of words, phrases or sentences) from long posts for providing the readers with the precise answer to the question. Sentence ranking is followed by answer extraction where the answers are extracted. Sultan[2], in his paper, has explained the generic framework that is followed by most of the extraction algorithms. For any question, there are candidate answer sentences from each of which chunks of texts are identified. These chunks are, then, evaluated according to some criterion. The criterion depends on the method used. The best chunk is then identified. After we have located the best chunks from different sentences, equivalent chunks are grouped together and the quality of each group is computed. Finally, a chunk is extracted from the best group supposed to be the most precise answer to the given question.

C. Web 2.0 and Internet Revolution

The term 'Web 2.0' was invented by Darcy DiNucci in 1999 and got popularised at the O'Reilly Media Web 2.0 Conferences in late 2004[3]. Websites in Web 2.0 allowed user interactions and collaborations in a virtual community as creators of user-generated content. The idea behind Web 2.0 was very distinct at that time before which the web only allowed visitors from viewing the static content without significant interactions. The idea of Web 2.0 can be decomposed into three components: Rich Internet Application (RIA), Web-Oriented Architecture (WOA) and Social Web. Web 2.0 sites included various features and techniques including search, extensions and signal which Andrew McAfee referred by the acronym SLATES[4].

Like all other things, internet sites have also undergone both a revolution and an evolution. As the global push towards online presence and information sharing continues, websites and forum platforms have also emerged and bloomed. Currently, we have access to a diverse range of contents than ever and the trend continues. Only in 2016, around 96,000 petabytes of information was transferred which was double than that in 2012[5]. On the other hand, there are already over a billion websites all over the internet full of information over diverse range[6].

III. Related Works

As the web and the virtual digital assistant technologies are enhancing, various works have been done on almost all major aspects of answer extraction, sentence ranking and answer pinpointing.

A. Answer sentence selection

Echihabi and Marcu, 2003[7], have explained question-answering system as a pipeline of only two high level modules: An Information retrieval engine that obtains information system resources R relevant to an information that may contain answers to a given question Q1 and an answer identifier module which ranks each information resource for its relevancy with question Q1. For example, if a whole sentence S from resource R is accepted as the most likely answer, cosine similarity between S and Q1 can be used to calculate the likelihood of an answer. Researches have shown that such word-overlap method is practically not a good enough metric for answer selection. Enhanced Models of lexical semantic resources have improved the performance over systems which focuses only on syntactic analysis through dependency tree matching [8].

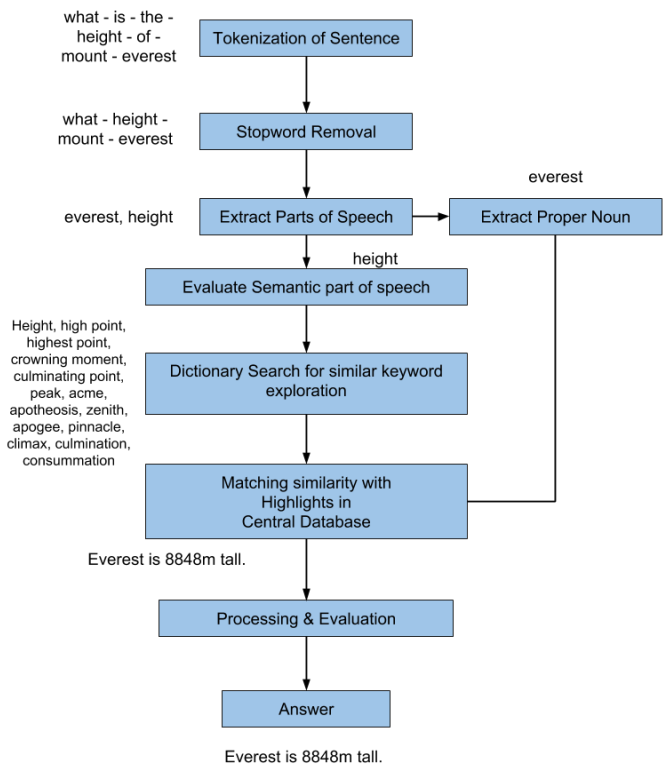

Fig. 2: Question-answering using Highlights from Central Database

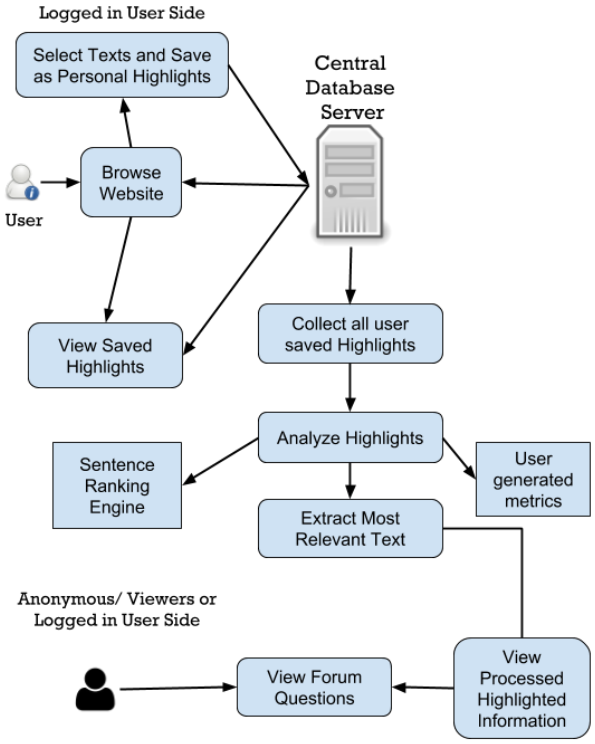

Fig. 3: Implementation Model for Information Extraction from User Highlights

B. DNN for answering questions

Researchers have been using semantic-parser constructed using Inductive Logic Programming from the inception of question-answering systems[9].

Semantic similarity model using convolutional neural networks have been used in question-answering to decompose questions into entities (Eq) and relation patterns. The similarity of question entities (Eq) with entities in the knowledge base (Ekb) and the similarity of relation patterns and relations between them have been evaluated using convolutional neural network models[8].

Recently, researches have been done to enhance intelligent recommendation systems using user-generated contents to have a significant effect on decisions in providing rich and customised user experiences through neural networks and tensor factorisation models[10].

According to Lai, Bui and Li[11], existing deep learning methods for answer selection can be examined along two dimensions: (i) learning approaches and (ii) neural network architectures where learning approaches use point-wise, pair-wise and list-wise approaches to learn the ranking function hθ. Siamese Architecture, Attentive Architecture, Compare-Aggregate Architecture are three main types of general architectures for measuring the relevance of a candidate sentence to a question.

IV. Methodology

Various methods like linguistic, statistical and pattern matching methods can be used for the ranking process. The possible answer sentence is segmented into words i.e tokenized. Then the stop-words are removed from the list of words. The proper noun is extracted and the semantic part of speech is analysed. The similar keyword for the semantics is matched with the highlights from the central database to generate more relevant results. Finally, after processing and evaluation, the answer is deducted (see Fig. 2).

While the answer extraction improves the search efficiency for answers, it can also be helpful in the validation of the information provided in answers. This is because in online discussion forums or any other question answering, where answers are provided by the people, the higher is the number of highlights for a particular answer (or a part of it), the more trustworthy the answer is.

A. User-generated contents (UGC)

User-generated contents involve all the contents which may be in the form of images, posts, comments, testimonials, etc. which are posted by users at online forums and social sites. Jos van Dijck, in his paper ‘Users like you? Theorising agency in

user-generated content' has stated that the meta-data harvested by Google from the UGC traffic is more valuable than the contents provided by users to its sites for advertising[12]. However, apart from advertising, the meta-data generated as a by-product of UGC can be a prime source of users' intent which can be used in the ranking of sentences for a relevant answer.

B. Engagement metrics

Engagement metrics include bounce rates for landing pages, the visit duration (i.e. the session length) of visitors, screen flow as well as the number of views, likes, shares, comments and clicks the posts have. These help in tracking the audience engagement, which in turn, provides the idea as to which posts are more accepted by the users. The visit duration gives knowledge of the time users spend on the pages (and the posts). Thus these metrics reveal a lot about user engagement which can be used in answer sentence ranking.

V. Implementation

Implementation of the described system can easily be done using some components of user engagement metrics and user-generated contents. Front-end web technologies like JavaScript and AJAX can be used to add features to forums. Browser-based plugins and add-ons can also be used to let users highlight the texts. Various methods, analytic tools and algorithms can be used for evaluating user-generated metrics which can also be used to provide rich user experience to the visitors (see Fig. 3).

A. Text Selection

Fig. 4: JavaScript Based Based Pop-up after text selection for Highlight in Forum

Whenever a logged-in user in forum/blog selects text, a pop-up is displayed (see Fig. 4). It facilitates users in saving the selected text i.e. in highlighting it. The highlight is saved by the user to be used as a private note. A user in the forum can't access another user's highlight library. However, such saved highlights can be accessed by sentence ranking engines as a heuristic for ranking purpose.

B. Saving User-Highlight to Central Database server

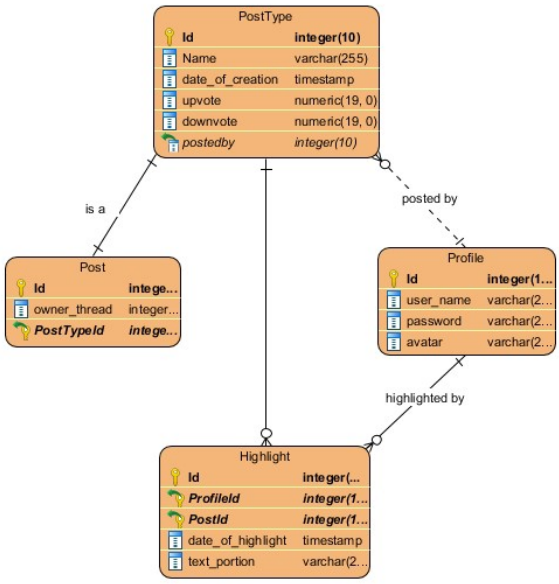

Fig. 5: Database Schema for User Highlights

In the central database server, the highlighted text along with the user who owns it is saved along with the date (see Fig. 5). The records of the database can be further used for the purpose of analysing the highlights.

C. Analysing Highlights using Sentence Ranking Engine

In different blogs and forum, there are several answers to a particular question. Out of those answers, there may be different highlights saved in a central database server. Using those highlights, each text is ranked to find the relevant answer. Tags are also useful in ranking the sentences. Some sentences are completely discarded for no relevance to the question.

D. Display relevant text for Blogs or Questions

Finally, the sentence with the highest rank is regarded as the most relevant text and is considered to be the answer to the question. So, the pinpoint answer to the question is displayed to users as the most relevant text as analysed by the sentence ranking engine (see Fig. 6).

VI. Results

The development of information extraction and sentence raking was analysed. It was found that the user-generated metrics and highlights can be used to improve sentence-ranking and answer-pinpointing. Also, the use of neural networks for developing models was explored along with various linguistic, statistical and pattern matching methods to be used in question-answering and important-part-pinpointing.

The team had also worked on a web-based project, parallel to the research, that uses JavaScript based pop-up (after text selection) in a web-page to be saved as a private note. It can be accessed by the system to find out the most highlighted part of the web-page. Such information collected is used for showing most relevant information about the page to the visitors.

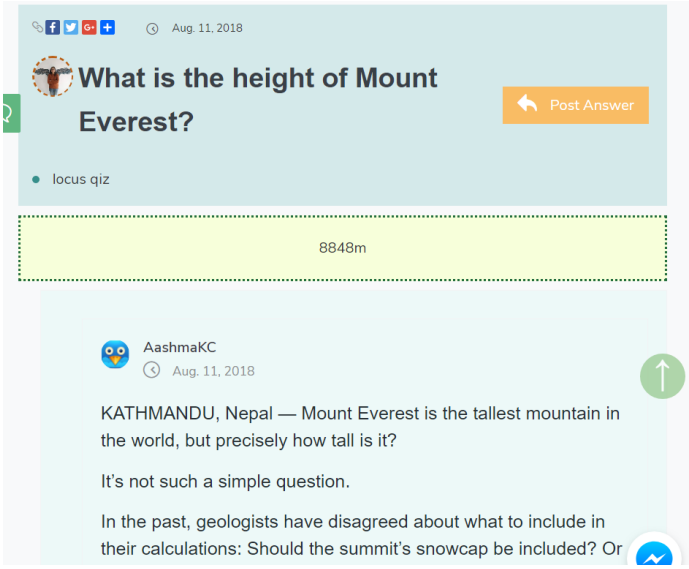

Fig. 6: Forum showing the relevant information about the post extracted from saved user Highlights.

VII. Conclusion

This article describes the possible uses of user-generated contents in sentence-ranking and answer-pinpointing in online websites to extract information. It explains different approaches that can be used for answer sentence-ranking and answer-extraction.

Despite the advantages of highlighting text, it has not been adopted by forums and websites for a long time. Although it has been commenced by a few websites such as the Medium, its use is not as ample as it needs to be. The question naturally arises as to why the feature of highlighting texts has not come into practice for a long time. This is because only after the advent of Web 2.0, the industry started focusing on client-side technologies including AJAX and JavaScript framework allowing for a rapid and interactive user experience. This made highlighting texts in web applications possible thus allowing websites to enable their users to enable rich user experience to highlight the part of text they want.

With the advent in technology, the intelligent systems/algorithms will be more intelligent and efficient in finding the user-demanded information from within the contents. We believe that user-generated metrics and data can be of great help for information-extraction.

VIII. Acknowledgement

The authors are highly indebted to faculty members of Department of Electronics and Computer Engineering, Pulchowk Campus, mainly Dr Arun Timalsina, Dr Basanta Joshi, Dr Aman Shakya and Mr Anil Verma for supporting us throughout project development and research work. The authors also would like to thank Ms Mansi Karna for helping in the publication.

References

[1] S. K. Dwivedi and V. Singh, "Research and reviews in question answering system," *Procedia Technology*, vol. 10, pp. 417–424, 2013.

[2] M. A. Sultan, V. Castelli, and R. Florian, "A joint model for answer sentence ranking and answer extraction," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 113–125, 2016.

[3] T. O'reilly, *What is web 2.0.* " O'Reilly Media, Inc.", 2009.

[4] A. P. McAfee, "Enterprise 2.0: The dawn of emergent collaboration," *MIT Sloan management review*, vol. 47, no. 3, p. 21, 2006.

[5] V. N. I. Cisco, "The zettabyte era: Trends and analysis," *Updated (29/05/2013), http://www. cisco. com/c/en/us/solutions/collateral/serviceprovider/visualnetworking-index-vni/VNI_Hyperconnectivity_WP. html*, 2014.

[6] I. Stats, "Internet live stats," *Pobrano z lokalizacji Internet Live Stats: http://internetlivestats. com (20.02. 2017)*, 2017.

[7] A. Echihabi and D. Marcu, "A noisy-channel approach to question answering," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 16–23, Association for Computational Linguistics, 2003.

[8] W.-t. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 643–648, 2014.

[9] J. M. Zelle and R. J. Mooney, "Learning to parse database queries using inductive logic programming," in *Proceedings of the national conference on artificial intelligence*, pp. 1050–1055, 1996.

[10] A. Taneja and A. Arora, "Modeling user preferences using neural networks and tensor factorization model," *International Journal of Information Management*, vol. 45, pp. 132–148, 2019.

[11] T. M. Lai, T. Bui, and S. Li, "A review on deep learning techniques applied to answer selection," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2132–2144, 2018.

[12] J. Van Dijck, "Users like you? theorizing agency in user-generated content," *Media, culture & society*, vol. 31, no. 1, pp. 41–58, 2009.