Journal of Information Systems and Telecommunication
Vol.11, No.1, January-March 2023, 41-47
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Customized Web Spider for Why-QA Pairs Corpus Preparation

Manvi Breja[1*]

[1].Department of Computer Engineering, Gurugram University, Gurugram, Haryana, India

Received: 9 Feb 2021/ Revised: 20 Mar 2022/ Accepted: 27 Apr 2022

Abstract

Considering the growth of researches on improving the performance of non-factoid question answering system, there is a need of an open-domain non-factoid dataset. There are some datasets available for non-factoid and even how-type questions but no appropriate dataset available which comprises only open-domain why-type questions that can cover all range of questions format. Why-questions play a significant role and are usually asked in every domain. They are more complex and difficult to get automatically answered by the system as why-questions seek reasoning for the task involved. They are prevalent and asked in curiosity by real users and thus their answering depends on the users' need, knowledge, context and their experience. The paper develops a customized web crawler for gathering a set of why-questions from five popular question answering websites viz. Answers.com, Yahoo! Answers, Suzan Verberne's open-source dataset, Quora and Ask.com available on Web irrespective of any domain. Along with the questions, their category, document title and appropriate answer candidates are also maintained in the dataset. With this, distribution of why-questions according to their type and category are illustrated. To the best of our knowledge, it is the first large enough dataset of 2000 open-domain why-questions with their relevant answers that will further help in stimulating researches focusing to improve the performance of non-factoid type why-QAS.

Keywords: Non-Factoid Questions; Web Crawler; Latent Dirichlet Allocation; Topic Modeling; Natural Language Processing.

1- Introduction

Question Answering Systems (QASs) answer the users' questions asked in natural language. With the increase in popularity of information access and providing an ease to user with an appropriate answer
to question, there is a challenging issue of automatically answering non-factoid questions. In English language, there are two broad categories of questions (1) Factoid questions of type what, where, who, when, which are simple and answered in a single phrase or sentence and (2) Non-Factoid questions comprising why and how-type are complex involving explanations and detailed reasoning in their answers. A non-factoid question possesses different answers to satisfy users' curiosity of different knowledge backgrounds [39]. Promising results are achieved for answering factoid questions; however non-factoid question answering is a challenging task. They require advanced NLP techniques to resolve open issues such as (1) ambiguity, (2) variability and (3) redundancy. Ambiguity implies difficulty faced in interpreting the context of non-factoid questions, variability implies different possible forms of answers to non-factoid questions, and redundancy refers to retrieving unnecessary texts along with possible answers of non-factoid questions [38,40]. The prime requisite for implementing non-factoid question answering system is dataset preparation depending on the requirement of the task. Without an efficient dataset of questions and their answers, it is very difficult to contribute significantly to the research community.

There are various datasets available, some of which are restricted domain [2,4,14,27,28,29,30], some correspond to only how-type questions (for ex. Yahoo! Answers Manner Questions L4 and L5 [31]) and some comprising why-type questions (for ex. only 3% of wh-questions in MSN click data, only 4.7% of 17000 questions are why-type questions in Webclopedia data collection [16, 32]). The questions in the dataset are significantly less in number which is not appropriate for research in developing why-type question answering system. Therefore, this paper has tried to develop a dataset for only why-type questions and their answers that is not only significant for our research but will contribute to research community working in the field of why-QA.

✉ **Manvi Breja**
manvibreja91@gmail.com

The paper is organized into sections where Section 1 introduces the concept of question answering with factoid and non-factoid QAS. Section 2 discusses the motivation for research in why-question answering. Section 3 describes various existing factoid and non-factoid datasets. Section 4 discusses the process of crawling and scraping applied to prepare a dataset of why-QA pairs. Section 5 analyzes the distribution of topics and different forms of why-question. Section 6 highlights applications of why-QA dataset and finally section 7 concludes with future research directions.

2- Motivation for Research in Why-QAS

Why-type questions are complex and involve variations in their answers. Their answers seek reasoning and explanations about the entity involved in the question. These questions depict the curiosity about something which are usually asked in every fields of life. There are some type of questions which have a definite reasoning and some have multiple possible answers depending on the users' knowledge, context and their experience. These questions and their answers possess cause-effect relationships where cause part is depicted in the answers and its effect part is asked in the questions. A considerable performance is already achieved in factoid type question answering, however research in non-factoid question answering is still challenging. There are different categorizations of non-factoid questions provided by [33] viz. list, confirmation, causal and hypothetical. The paper carries out an approach in why-question answering that incorporates causal relations. They are difficult to get automatically answered as it is difficult to understand the accurate users' interpretations and thus require advanced techniques. There are limited number of open-source datasets available which comprises significant number of why-type questions. This motivates us to design a dataset having open-domain why-type questions and their answers.

3- Available Factoid and Non-Factoid Datasets

This section discusses the brief characteristics of the datasets available on web. It also highlights their merits and demerits which significantly motivates us to prepare a dataset for why-QA pairs.

Suzan Verberne in 2006 [1] released a first open-domain why-QA dataset on Web. It is an open-source dataset comprising 395 why-questions and 769 corresponding answers, out of which 166 questions were further paraphrased. In addition to the question, dataset also comprises its source document with relevant user-formulated answers. It is effective to be utilized by the researchers working on QAS but not large enough and needs to be expanded further. InsuranceQA in 2015 [2] comprises restricted questions on insurance domain. The questions are collected from insurance library website, comprising 16889 questions and 27413 answers. The questions are usually asked from real users and answers are constructed by domain experts. The dataset can be utilized for applying deep learning models on answer selection task. In 2018, a large dataset named WikiPassageQA [3] is developed which comprises 4165 open-domain non-factoid questions split into 0.8/0.1/0.1 resulting training set comprising 3332 questions, development set comprising 417 questions and testing set comprising 416 questions. With the questions, documentID, document name and answer passage are also stored in the dataset. It is effective to be utilized for answer passage retrieval from relevant documents by applying deep learning models. Further in 2018, another restricted domain why-QA dataset; PhotoshopQuiA [4] is released which comprises 2854 why-questions on Adobe Photoshop collected from five CQA websites, viz. Adobe Forums [6] Stack Overflow [5], Graphics Design [7], Super User [8] and Feedback Photoshop [9]. With each why-QA pair, dataset also includes question id, URL, title, date, questioner, his level, open or resolved state, full question text, HTML and for each answer: answer date, answerer, his level, answer votes, text, full answer HTML and a value indicating if answer is best or not. The dataset is effective to be utilized for understanding different characteristics of why-questions and further instigating them for recommendation systems and chatbots. Freebase QA in 2019 [10] is an open-domain QA developed by complementing trivia type QA pairs with Freebase triples, comprising 54K matches from 28,348 unique questions with 20,358, 3994, and 3996 training, development and evaluation sets respectively. The dataset comprises following entries version of dataset, set of unique questions in dataset comprising Question-ID, original question, processed question, semantic parse-id, topic entity in question, name of topic entity, Freebase MID of topic entity, path from topic entity to answer node in Freebase, Freebase MID of answer, answer string from original QA pair . It is effective enough to train machine learning models as QA pairs are matched with (subject, predicate, object) triples that also helps to understand the meaning of questions and search for correct answers in Freebase. It can be utilized for several applications like reading comprehension [11], natural language understanding and search. It is a complex knowledge-base dataset which makes invaluable for more advanced ML methods. ANTIQUE in 2020 [12] is a collection of 34,011 non-factoid QA pairs usually asked by users on Yahoo!Answers [13] where 2426 questions & 27.4k

judged answers are training set and 200 questions & 6.5k judged answers are testing sets. Out of 34,011 QA pairs, 36% of questions are why-questions majorly covering intent of questions asked on community QA sites. TutorialVisualQA in 2020 [14] comprises 6195 non-factoid QA pairs based on 76 tutorial videos. Dataset includes varied fields like questions, video_id, manually annotated answer fragments, answer_start and answer_end denoting indexes of beginning and ending answer sentences. This dataset can be helpful for implementing a model to understand answer boundary sentences which can further be utilized for other educational, instructional, cooking videos and many more.

The brief description and features of the available datasets comprising why-questions is discussed in Table 1.

Table 1: Brief of some existing datasets on why-questions and their answers

Datasets	Description
Suzan Verberne dataset (2006)	395 why-questions and 769 answers formulated by annotators, collecting text documents from Textline Global News (1989) [34] and The Guardian on CD-ROM (1992) [35]. Dataset is not large enough to be utilized for further research
InsuranceQA	Comprises 15867 non-factoid questions and 24981 unique answers related to insurance domain. Accurate number of why-questions in dataset is not mentioned. The dataset is restricted to domain thus can't be utilized for all research purposes
WikiPassageQA	Comprises total 4165 open-domain non-factoid questions and 244136 candidate passages as answers to these questions. Out of these only 14% are why-type questions which is definitely not cover all range of why-type questions and thus not significant for use in why-QAS.
PhotoshopQuiA	2854 why-QA pairs collected from 5 community QA websites. The dataset is restricted to Photoshop domain and mostly cover QA pairs asked on community sites
ANTIQUE	Comprises 2626 non-factoid questions collected from Yahoo! Answers community question answering websites and their 34011 corresponding annotated answers. Out of these non-factoid questions, approximately 900 are of why-type which are not enough for research in only why-QAS

4- Open-Domain Why-QA Dataset Preparation

This section discusses the process of collecting why-QA pairs and processing them to create a final dataset.

4-1-Customized Crawling and Scraping

Five web sites viz. Answers.com [15], Yahoo! Answers [13], Suzan Verberne's open-source dataset [16], Quora [17] and Ask.com[18] are visited by web crawler and Scrapy [19] which is an open source framework written in Python is utilized to scrap or extract required data from these websites. A customized spider is implemented which extracts all questions beginning with 'Why'. BeautifulSoup [20] is used to parse the HTML source of web page and questions are extracted from the 'href' tag of HTML source. The process resulted in a collection of Why-questions having different forms like 'Why is/was/were', 'Why do/does/did', 'Why should/would' and 'Why NP/PN' with their negative variants.

The functioning of web crawling and scraping is depicted in the form of Flowchart illustrated in Figure 1 and discussed further.

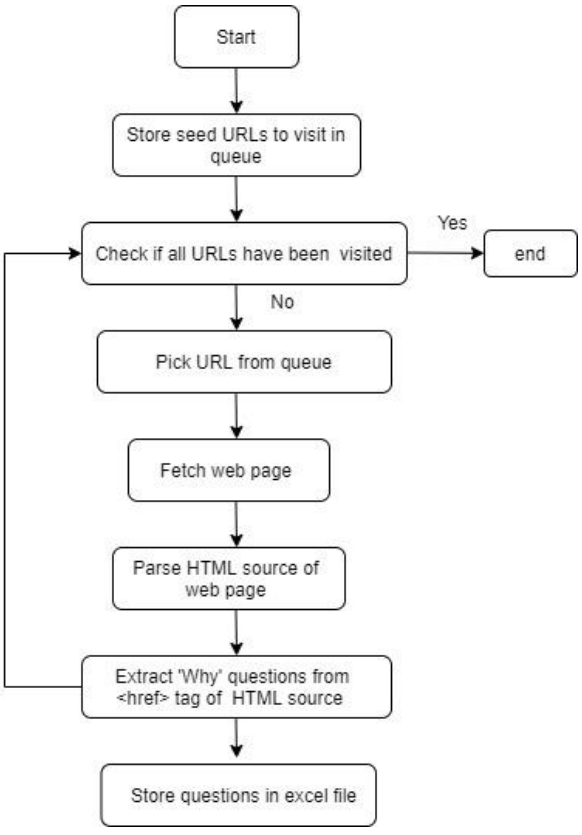

Fig. 1. Flowchart describing functioning of web crawler

Web crawler also known as spider or search engine bot is a part of search engines for example, Google, Bing, Yahoo etc. It is a software program that accesses a website, downloads and indexes information content available on the web page. To make automated browsing, the crawler is designed to repeat accessing and downloading content depending on the programming instructions included in crawler.

Since content on Internet is expanding and updating daily, it is important to index the webpages so that it can be easily accessible and downloadable based on users' query. Web crawler starts from a seed URL and crawl the webpages, follow hyperlinks to those URLs, thus periodically visit pages in order to index the updated content on webpages. Sometimes the webpages which are required to crawl depends on number of hyperlinks to the page, amount of visitors on page and other factors determining the importance of page corresponding to the required content.

Web crawling and scraping are two distinguished terms. Web crawlers continuously visit web links and download contents on web pages. Web crawlers obey robots.txt (robots exclusion protocol) file provided by web server of particular web page. This file specifies the rules for which pages and links to pages can be crawled. On the other hand, Web scraping is more targeted than web crawling and visit only specific webpages to download their content without any permission from the web server.

Figure 1 describes the process of crawling where crawler starts with seed URLs which are question answering sites. If all sites are visited, the process is stopped otherwise crawler visits each webpage and retrieves its HTML source. Here the crawler is designed to download questions beginning with 'why', thus it is extracted by scraping <href> tag of HTML source of web page. The process is repeated for all QA websites available and the questions are stored in excel file.

4-2-Processing of Why-Questions

A collection of 2000 why-questions are downloaded and stored in excel file. The questions are further processed to rectify grammatical and spelling mistakes. The why-questions of the form starting with 'Why' and ending with '?' are retained in the dataset. Question comprising more than one questions are separated into individual questions, coreference resolution [21] is applied to filter out significant meaningful why-type questions. Coreference resolution is NLP (Natural Language Processing) task which finds linguistic expressions such as pronouns in the question and replace them with real-world entity such as noun phrases which helps to understand the appropriate meaning of the question and help determine the main focus of question.

4-3- Different Fields in the Dataset

With each question, different attributes are also maintained in the dataset which are:

a) *Data Source of question:* There are five web sites from which web spider has extracted why-type questions. The data source from which the why-question is collected is stored as an attribute 'source_ques' in dataset.

b) *Category of the question:* With each why-type question, its category is stored in dataset. The questions collected from Answers.com and Yahoo have their category assigned like Science, Math, History, Technology and so on whereas questions collected from other web sites are assigned category by applying LDA process [22,23] which is one of the topic modelling approach. Latent Dirichlet Allocations abbreviated as LDA models Dirichlet distributions to classify questions to a particular topic. This category is stored with field name 'categ_ques' in the dataset.

c) *Relevant answer candidates:* For each why-type question, five answer candidates are maintained in the dataset. Only if the question is answered by an expert on the considered five question answering websites, its corresponding answer is saved as an answer candidate otherwise the answer candidates are retrieved by posing the question on Google search engine. The question is queried on the Google which extracts relevant web pages from which first five are studied manually to retrieve appropriate answer passages treated as answer candidate.

Since the answer to why-question involves explanation to the causation asked in the question, appropriate answer candidate is retrieved by determining the answer boundary around causal cue phrases [24] such as 'because', 'since', 'due to', 'in order to', 'therefore', 'as a result' etc. The causal phrases determine explicit causality involved in text. The why-type questions reflect the effect part and its cause part is reflected from the answers to why-type questions.

d) *Document title and link:* Each why-question is accompanied with the document title and its link which is the title and the link of web page from which an answer candidate is retrieved. It is saved as an attribute 'doc_title_link' in the dataset.

The snapshots of the generated dataset are illustrated in Figure 2 and Figure 3 below:

Data source of question	category of question	Relevant Answer Candidates	Document Title	Document Link
Quora	Travel	The Taj Mahal is located on the right bank of the Yamuna River in a vast Mughal garden that encompasses nearly 17 hectares, in the Agra District in Uttar Pradesh. It was built by Mughal Emperor Shah Jahan in memory of his wife Mumtaz Mahal with construction starting in 1632 AD and completed in 1648 AD, with the mosque, the guest house and the main gateway on the south, the outer courtyard and its cloisters were added subsequently and completed in 1653 AD. The existence of several historical and Quaranic inscriptions in Arabic script have facilitated setting the chronology of Taj Mahal.	Taj Mahal	https://whc.unesco.org/en/list/252/
Quora	Travel	"The Taj Mahal, meaning "Crown of the Palace", is an ivory-white marble mausoleum on the south bank of the Yamuna river in the Indian city of Agra. It was commissioned in 1632 by the Mughal emperor, Shah Jahan (reigned from 1628 to 1658), to house the tomb of his favourite wife, Mumtaz Mahal. The tomb is the centerpiece of a 17-hectare (42-acre) complex, which includes a mosque and a guest house, and is set in formal gardens bounded on three sides by a crenellated wall."	Why was the Taj Mahal built?	https://www.quora.com/What-is-the-real-reason-The-Taj-Ma
Quora	Travel	Often described as one of the wonders of the world, the stunning 17th Century white marble Taj Mahal was built by Mughal emperor Shah Jahan as a mausoleum for his beloved wife Mumtaz Mahal, who died in childbirth.	Taj Mahal: Was India's 'monument to love' built out of guilt?	https://www.bbc.com/news/world-asia-india-27970691
Quora	Travel	The Taj Mahal was built by the Mughal emperor Shah Jahan (reigned 1628–58) to immortalize his wife Mumtaz Mahal ("Chosen One of the Palace"). She died in childbirth in 1631, after having been the emperor's inseparable companion since their marriage in 1612	Taj Mahal	https://www.britannica.com/art/Shah-Jahan-period-architect
Quora	Travel	The Taj Mahal is an ivory-white marble mausoleum on the southern bank of the river Yamuna in the Indian city of Agra. It was commissioned in 1632 by the Mughal emperor Shah Jahan (reigned from 1628 to 1658) to house the tomb of his favourite wife, Mumtaz Mahal; it also houses the tomb of Shah Jahan himself.	Taj Mahal	https://en.wikipedia.org/wiki/Taj_Mahal

Fig. 2. Snapshot 1 of dataset

Why is Venus hottest planet of solar system?	Quora	Science & Mathematics	Venus is the second planet from the Sun and our closest planetary neighbor. Similar in structure and size to Earth, Venus spins slowly in the opposite direction from most planets. Its thick atmosphere traps heat in a runaway greenhouse effect, making it the hottest planet in our solar system with surface temperatures hot enough to melt lead.	Venus	https://mobile.arc.nasa.gov/public/explore/missions/pages/galileo/extent/venus.html
Why is Venus hottest planet of solar system?	Quora	Science & Mathematics	Venus is the hottest planet in the solar system. Although Venus is not the planet closest to the sun, its dense atmosphere traps heat in a runaway version of the greenhouse effect that warms Earth. As a result, temperatures on Venus reach 880 degrees Fahrenheit (471 degrees Celsius), which is more than hot enough to melt lead. Spacecraft have survived only a few hours after landing on the planet before being destroyed.	Venus: The hot, hellish & volcanic planet	https://www.space.com/64-venus-second-planet-from-the-sun-brightest-planet-in-solar-system.html
Why is Venus hottest planet of solar system?	Quora	Science & Mathematics	Even though Mercury is the closest planet to the Sun, Venus is the hottest planet in our solar system. This is because Mercury has almost no atmosphere, while Venus has a very thick atmosphere. This causes all the heat to be radiated back into space on Mercury. However, the heat is trapped on Venus, with the average temperature being 462°C.	Why is Venus the hottest planet in our solar system?	https://indianty.com/knowledge/why-is-venus-the-hottest-planet-in-our-solar-system-1688260417879
Why is Venus hottest planet of solar system?	Quora	Science & Mathematics	Venus is so hot because it is surrounded by a very thick atmosphere which is about 100 times more massive than our atmosphere here on Earth. As sunlight passes through the atmosphere, it heats up the surface of Venus. Most of this heat cannot escape back into space because it is blocked by the very thick atmosphere of Venus. The heat becomes trapped and builds up to extremely high temperatures. This trapping of heat by the atmosphere is called the greenhouse effect because it is similar to how the glass in a greenhouse traps heat. The greenhouse effect on Venus causes the temperatures at its surface to reach 864 degrees Fahrenheit (462 degrees Celsius), making Venus the hottest planet in the entire Solar System!	Why is Venus so hot? Surface Of Venus: Why Is Venus The Hottest Planet?	https://coolcosmos.ipac.caltech.edu/ask/38-Why-is-Venus-so-hot-
Why is Venus hottest planet of solar system?	Quora	Science & Mathematics	Venus is the hottest planet in our solar system because it is covered by a thick layer of clouds composed of carbon dioxide and other gases, which prevent the heat from the sun from escaping back into outer space. This is why the planet continues absorbing the heat from the sun and becomes increasingly hot.		https://www.scienceabc.com/nature/universe/surface-of-venus-why-is-venus-the-hottest-planet.html

Fig. 3. Snapshot 2 of dataset

5- Results and Discussion

The crawler designed which visits each question answering web sites, extracts HTML source and retrieve why-type questions available on the QA sites, resulted in a collection of around 2000 why-type questions. The dataset is further enhanced by incorporating data source, category, five answer candidates, document link of the respective answer candidate. The dataset is analyzed on the distribution of topics and distribution of different forms of why-questions. The analysis is performed in order to contrast the dataset with available open-source datasets and further help the researchers get an idea of the variations involved in the dataset collection.

5-1- Variation of Topics in Why-Questions of Dataset

The why-questions in dataset have variation in their topics. Various different categories are visualized like 'Education', 'Politics', 'Health', 'Science & Mathematics', 'Family & Relationships', 'Travel', 'Sports', 'Electronic Products', 'Food, Drink & Dining Out' and 'Entertainment & Music' etc. The distribution of questions belonging to these major categories and others is illustrated in Figure 4 below:

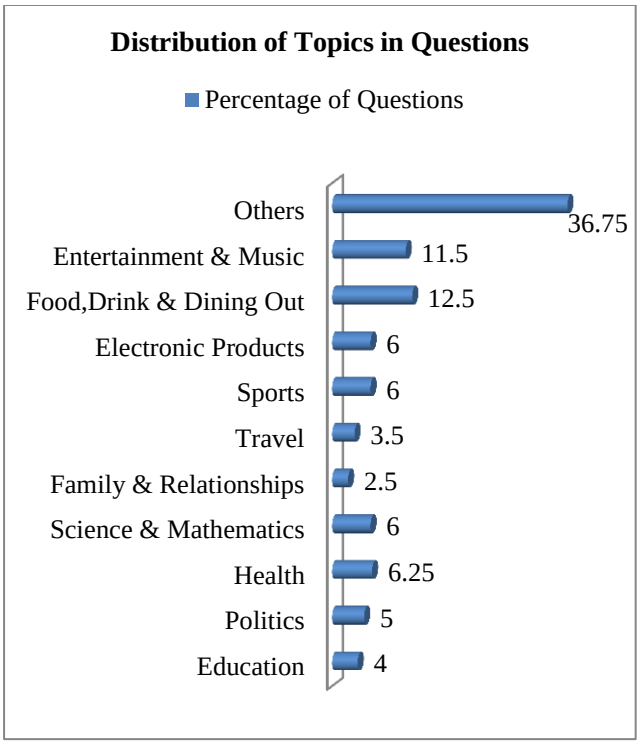

Fig. 4. Distribution of topics in why-questions

5-2- Distribution of Forms of Why-Questions

As mentioned in subsection 3.1, there are different forms of why-questions viz. 'Why is/was/were', 'Why do/does/did', 'Why should/would' and 'Why NP/PN' with their negative variants. Figure 5 illustrates the distribution of subcategories of why-type questions.

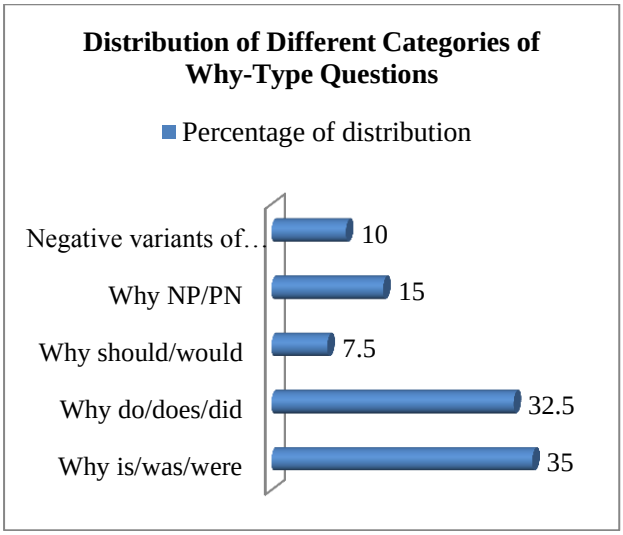

Fig. 5. Distribution of different forms of why-questions

6- Usage of Why-QA Dataset

This section highlights expected usage of why-QA dataset, some of which we have utilized in our research.

a) *Question Classification:* An open-domain why-QA dataset is utilized for classifying and assigning question and answer subtype. A taxonomy of why-type questions and their answers are proposed in the research [25,26, 36] which extracts lexical features of questions and assigns a question and answer type to them.
b) *Question to Query Reformulation:* The reformulated question helps in better document retrieval and thus answer candidate extraction. Rules are formulated for each type of why-question to convert them into appropriate user-oriented query which when posed on search engine helps in retrieving the accurate and relevant documents [37].
c) *Answer Re-Ranker:* With each why-type question, there are five possible answer candidates stored in the dataset. These answer candidates are re-ranked on the basis of similarity values and their relatedness with question [38].

4- Conclusions and Future Work

The paper proposes a dataset containing why-type questions and their answer candidates. Since the questions are sampled from various question answering websites such as Yahoo! Answers, Quora etc., the questions address the real-world problems usually faced by the user. The questions in dataset are collected with the help of web crawler in its original form, thus easier to predict properties and nature of why-type questions.
We believe that the developed open-domain why-QA dataset unfolds new avenues for research in improving performance of non-factoid QAS. With this, it also motivates to foster the techniques required for other applications like recommendation systems, chatbots, virtual assistants and many more.

References

[1] S. Verberne, L.W.J. Boves, N.H.J. Oostdijk and P.A.J.M. Coppen, "Data for question answering: the case of why", 2006.
[2] shuzi, "GitHub - shuzi/insuranceQA: A question answering corpus in insurance domain", 2015. [Online]. Available: https://github.com/shuzi/insuranceQA. [Accessed Feb. 9, 2021].
[3] D. Cohen, L. Yang,, and W. B. Croft, "Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval", In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1165-1168.
[4] A. Dulceanu, T. Le Dinh, W. Chang, T. Bui, D.S. Kim, M.C.Vu, and S. Kim, "PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering", In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 2763-2770.
[5] Stack Overflow - Where Developers Learn, Share, & Build Careers. [Online]. Available: https://stackoverflow.com/. [Accessed Feb. 9, 2021].
[6] Adobe Support Community. [Online]. Available: https://forums.adobe.com/welcome. [Accessed Feb. 9, 2021].
[7] Graphic Design Stack Exchange. [Online]. Available https://graphicdesign.stackexchange.com. [Accessed Feb. 9, 2021].
[8] Super User Stack Exchange [Online]. Available https://superuser.com. [Accessed Feb. 9, 2021].
[9] Adobe Photoshop Family [Online]. Available https://feedback.photoshop.com. [Accessed Feb. 9, 2021].
[10] K. Jiang,, D. Wu and H. Jiang., "FreebaseQA: a new factoid QA data set matching Trivia-style question-answer pairs with freebase", In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), June 2019, pp. 318-323.
[11] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis and E. Grefenstette, "The narrativeqa reading comprehension challenge", Transactions of the Association for Computational Linguistics, vol. 6, pp. 317-328, 2018.
[12] H. Hashemi, M. Aliannejadi, H. Zamani and W.B. Croft, "ANTIQUE: A non-factoid question answering benchmark.". In European Conference on Information Retrieval, Springer, Cham, 2020, pp. 166-173.
[13]Yahoo! answers [Online]. Available https://answers.yahoo.com/. [Accessed Feb. 9, 2021].
[14] A. Colas, S. Kim, F. Dernoncourt,, S. Gupte, D.Z. Wang and D.S. Kim, "TutorialVQA: Question Answering Dataset for Tutorial Videos" [Online]. Available arXiv preprint arXiv:1912.01046, 2019.
[15] Answers [Online]. Available https://www.answers.com/. . [Accessed Feb. 9, 2021].
[16] S. Verberne, "Data Download," [Online]. Available: http://sverberne.ruhosting.nl/wordpress/research/data-download/. [Accessed Feb. 9, 2021].
[17] Quora [Online]. Available https://www.quora.com/ . [Accessed Feb. 9, 2021].
[18] Ask [Online]. Available https://www.ask.com/. [Accessed Feb. 9, 2021].

[19] Scrapy [Online]. Available https://scrapy.org/. [Accessed Feb. 9, 2021].
[20] BeautifulSoup [Online]. Available https://pypi.org/project/beautifulsoup4/. [Accessed Feb. 9, 2021].
[21] A. Rahman and V. Ng, "Coreference resolution with world knowledge", In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 814-824.
[22] A. Anandkumar, D. P. Foster, D. Hsu, S.M. Kakade and Y.K. Liu, "A spectral algorithm for latent dirichlet allocation", Algorithmica, vol. 72, no. 1, 2015, pp. 193-214.
[23] D. M. Blei, A.Y. Ng and M. I. Jordan, "Latent dirichlet allocation.", Journal of machine Learning research, 2003, pp. 993-1022.
[24] D.S. Chang and K.S. Choi, "Causal relation extraction using cue phrase and lexical pair probabilities", In International Conference on Natural Language Processing, Springer, Berlin, Heidelberg, 2004, pp. 61-70.
[25] M. Breja and S.K. Jain, "Why-type Question Classification in Question Answering System", In FIRE (Working Notes), 2017, pp. 149-153.
[26] M. Breja and S.K. Jain, "Analysis of Why-Type Questions for the Question Answering System", In European Conference on Advances in Databases and Information Systems, Springer, Cham, 2018, pp. 265-273.
[27] H. Fu and Y. Fan, "Music information seeking via social Q&A: An analysis of questions in music StackExchange community", In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, 2016, pp. 139-142.
[28] Financial Opinion Mining and Question Answering [Online] Available https://sites.google.com/view/fiqa/home/. [Accessed Feb. 9, 2021].
[29] A.F.U.R. Khilji, R. Manna, S.R. Laskar, P. Pakray, D. Das, S. Bandyopadhyay and A. Gelbukh, "Question classification and answer extraction for developing a cooking QA system", Computación y Sistemas, vol. 24, no. 2, 2020, pp. 921-927.
[30] V. Koeman, L. A. Dennis, M. Webster, M. Fisher and K. Hindriks, "The Why did you do that?" Button: Answering Why-questions for end users of Robotic Systems", In 7th International Workshop on Engineering Multi-Agent Systems (EMAs 2019), 2019, pp. 152-172.
[31] Yahoo! Language Data [Online]. Available https://webscope.sandbox.yahoo.com/catalog.php?datatype=l. [Accessed Feb. 9, 2021].
[32] E. Hovy, U. Hermjakob, and D. Ravichandran, "A question/answer typology with surface text patterns", In Proceedings of the Hum an Language Technology conference (HLT), San Diego, CA, 2002.
[33] A. Mishra and S.K. Jain, "A survey on question answering systems with classification", Journal of King Saud University-Computer and Information Sciences, vol. 28, no. 3, 2016, pp.345-361.
[34] The Baron [Online]. Available https://www.thebaron.info/archives/technology/reuters-technical-development-chronology-1991-1994, [Accessed March, 02, 2021]
[35] G. Smith, "Newspapers on CD-ROM", In Serials The Journal for the Serials Community, vol. 5, no. 3, 1992, pp. 17-22.
[36] M. Breja and S.K. Jain, "Analyzing Linguistic Features for Classifying Why-Type Non-Factoid Questions", International Journal of Information Technology and Web Engineering (IJITWE), vol. 16, no. 3, 2021, pp.21-38.
[37] M. Breja and S.K. Jain, "Why-Type Question to Query Reformulation for Efficient Document Retrieval", International Journal of Information Retrieval Research (IJIRR), vol. 12, no. 1, 2022, pp.1-18.
[38] M. Breja and S.K. Jain, "Analyzing Linguistic Features for Answer Re-Ranking of Why-Questions", Journal of Cases on Information Technology (JCIT), vol. 24, no. 3, 2022, pp.1-16.
[39] M. Breja and S.K. Jain, "A survey on non-factoid question answering systems", International Journal of Computers and Applications,2021, pp.1-8.
[40] Y. Niu , "Analysis of semantic classes: toward non-factoid question answering". University of Toronto, 2007 .