Initial Lessons from Building an IVR-based Automated Question-Answering System

Pranav Bhagat
Indian Institute of Technology
Delhi, India
cs1160352@iitd.ac.in

Sachin Kumar Prajapati
Indian Institute of Technology
Delhi, India
cs1160355@iitd.ac.in

Aaditeshwar Seth
Indian Institute of Technology
Delhi, India
aseth@cse.iitd.ac.in

ABSTRACT

With improvements in speech recognition and natural language processing capabilities, voicebot systems show promise to run interactive information services for less-literate populations in developing regions. In this context, we describe our initial experiences towards building an automated question-answering system in the domain of sexual and reproductive health and rights. This system is trained on data acquired from an IVR (Interactive Voice Response) platform on which users could record questions, which were then moderated and sent to an expert to get answers. Our goal is to now use this data to build an automated answer retrieval system so that questions can be answered in real time by retrieving an appropriate answer from the corpus of questions and answers available so far. Our insights are likely to be useful for several initiatives using IVR systems and looking to automate their search and retrieval functionality.

KEYWORDS

FAQ retrieval, question-answering, voicebot, Interactive Voice Response (IVR) systems, voice forums, Sexual and Reproductive Health and Rights (SRHR)

ACM Reference Format:
Pranav Bhagat, Sachin Kumar Prajapati, and Aaditeshwar Seth. 2020. Initial Lessons from Building an IVR-based Automated Question-Answering System. In *Information and Communication Technologies and Development (ICTD '20), June 17–20, 2020, Guayaquil, Ecuador*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3392561.3397581

1 INTRODUCTION

Domain-specific automated question-answering (QA) systems may provide a more user-friendly means of information search than keyword based queries [10, 23]. FAQ (Frequently Asked Questions) retrieval is a particular form of QA systems where incoming questions are matched against a database of FAQs and a proximate answer is returned [11, 17]. These are particularly useful in settings where many users are likely to have similar queries, and answers manually prepared for common questions can thus address the needs of many users. Further, with improvements in automated speech recognition, voice-based QA systems are likely to have a strong potential as well, especially to service less-literate populations in developing regions of the world who prefer to use voice instead of text as an interaction modality [5, 19, 20]. In this paper, we report initial experiences with building a FAQ retrieval system in the domain of sexual and reproductive health and rights (SRHR), with a goal to eventually run it as a voicebot over IVR (Interactive Voice Response) systems for adolescent and young girls and boys in India.

An IVR system on SRHR called Kahi Ankahi Baatein (KAB) has been running since almost five years in India, and services calls from across the country (in Hindi) [5]. A popular use-case on the platform is a manually operated QA programme on which young callers ask questions about masturbation, menstruation, contraceptives, love affairs, and sexual insecurity. Interesting questions are selected by a team of content moderators and a few selected ones are answered each week by an SRHR specialist, through an underlying framing of sexual independence and a non-judgmental approach, to impress upon young minds an agenda of gender equality, a right over their own bodies, importance of consent, and a scientific understanding of sexual processes to counter myths and misconceptions [4]. The initial traction for KAB was built through promotional advertisements on a network of community radio stations in India, and the platform has since then sustained itself through word-of-mouth publicity and workshops by a large network of partners led by the feminist organization CREA. Over the years, a large database of SRHR questions and answers has been curated, and our goal was to use this database to build an automated FAQ retrieval system so that questions can be answered automatically and instantly, to bypass the moderation step.

We report on our initial experiences of using the KAB dataset of questions and answers to build a FAQ retrieval system. We identified several issues which are likely to be relevant for other ICT4D initiatives that use IVR systems for manual QA and would be looking towards automating the process in the future [16, 18–20]. First, we find that popular speech recognition APIs offered by Google and Amazon often make mistakes in their transcriptions, and we quantify the loss in accuracy with using such APIs, benchmarked against manual transcriptions of the audio. Second, we find that on systems like KAB where users are provided a single voice-recording opportunity to record their question, users tend to provide considerable non-relevant information such as their name, their personal details, etc, and this superfluous information causes an accuracy loss for information retrieval systems. We quantify this loss, benchmarked against an annotated dataset that separates out the non-relevant portions of the question from its relevant portions. Finally, we make recommendations for a voicebot style user-interface design for such systems.

ICTD '20, June 17–20, 2020, Guayaquil, Ecuador

ACM ISBN 978-1-4503-8762-0/20/06.
https://doi.org/10.1145/3392561.3397581

2 RELATED WORK

With improvements in natural language processing methods, chatbots have emerged as a novel interaction modality to guide users in content search and retrieval, and have also found applications in social development contexts. Mental health is one such area where users seem to feel comfortable in interacting anonymously with a non-judgemental machine [1], and has seen innovations such as Woebot [9] which first asks the user a series of questions and then recommends to them appropriate tools and strategies for behavioral therapy. Doctor Vdoc is another chatbot for medical purposes to get referral advice [15]. Attempts have also been for voicebots in the social development context, such as FarmChat [12], to answer farmer queries in Hindi. Our work is related but addresses a different problem. One, we evaluate the feasibility of using IVR systems for voice input, which tends to be of a lower audio quality than voice recordings done via a smartphone mobile application. Two, our dataset from KAB is not in a conversational format through which we can derive a dialogue flow, rather our setting is of FAQ retrieval where we need to match an incoming query question against an FAQ database and return an appropriate answer.

FAQ retrieval is a form of automated QA. Research in automated QA can itself be divided into two broad areas: factoid and non-factoid question-answering. Factoid QA tackles factual questions that are typically answerable in short answers, such as "*who is India's prime minister*", and research looks at learning or retrieving such facts from document corpuses like the Wikipedia [3]. Our problem is closer to that of non-factoid QA, which tries to answer questions like "*how can I delete my Facebook account*". Past work has tried to solve this task by building knowledge bases of possible question trees and answer trees [24], matching questions and answers using the word embedding based features of uni/bi/tri-grams and other linguistic units [8], and other matching techniques such as through the use of a bi-LSTM architecture with an attention model to learn better word representations [22], and a BERT based architecture to fine-tune representations during the training phase as well [6]. In this paper, we use the BERT model and compare it against a baseline model that uses simple Jaccard's similarity to match questions and answers.

Conversational interaction and FAQ retrieval in the IVR setting has not been actively researched as yet. Several IVR systems, especially voice forums where users can listen to audio messages and record their own messages, have been actively used in social development programmes such as for agricultural advisory [20], grievance reporting [16, 18], behavior change communication [2], and support groups for physically disabled people [7] and HIV/AIDS patients [13], among others. None of these platforms have however so far attempted to automate their search and retrieval functions for greater scalability. Google's Dialogflow[1] and Amazon's Alexa[2] have emerged as popular frameworks to develop chatbots and voicebots with which such IVR systems can integrate to build a conversational capability that goes beyond the current keypress-based interaction modality used in most such systems. Alexa is also building a crowdsourced database for factoid QA [14], but cannot handle non-factoid QA like FAQ retrieval. Our insights are therefore likely to be useful to many IVR systems for social development that want to build conversational FAQ retrieval systems, and have a dataset that has been acquired in a non-conversational setting.

[1] https://dialogflow.com/
[2] https://www.alexa.com/

3 KAB DATASET

The KAB dataset we use is a collection of audio files of 90 answers and 516 questions, that came up in the last two years of the running of KAB. Whenever a question is recorded on KAB, content moderators first try to identify an earlier answer that can be served in response to the question, else they refer the question to an SRHR expert. The expert then provides an answer to this new question. In this way, each answer in the dataset was mapped to multiple questions, essentially capturing different ways in which the question could have been asked.

Since questions were recorded over IVR, some of them had a poor audio quality. We worked with an agency to manually transcribe all questions and answers, and also obtained transcripts through the Google and Amazon speech APIs as well. We found that this automatically transcribed text had two sources of errors: (a) Several parts of the audio recording being of very poor quality could not be transcribed at all, resulting in very short transcripts for some audio, and (b) some words were misspelled, especially sexual terms or Hindi slangs as shown in Figure 1. For this second source of error, we built a custom vocabulary with help from the KAB team and also by identifying words having a high TF/IDF score in several online manuals related to SRHR provided to us by the KAB team. The Amazon APIs allow such a custom vocabulary to be specified. We then evaluated the quality of the automatically generated transcripts using two metrics. First, we computed the words/sec for the question audios by dividing the number of words in the automatically generated transcript by the length (in seconds) of the audio. Hindi is typically spoken at a rate 1.5 words/sec, and transcripts which are considerably shorter than this ratio are likely to be for cases where the transcription failed. Figure 2a shows a CDF of the words/sec for transcripts generated through the different methods. Second, we compare the automatically generated transcripts with the manual transcripts by computing word error rate based on the length normalized edit distance between the ground-truth sentence and the hypothesis sentence. Figure 2b shows a CDF of the word error rates for the different transcripts. Both metrics demonstrate the superiority of the transcripts generated by the Amazon APIs with a custom vocabulary, and we use this for subsequent experiments.

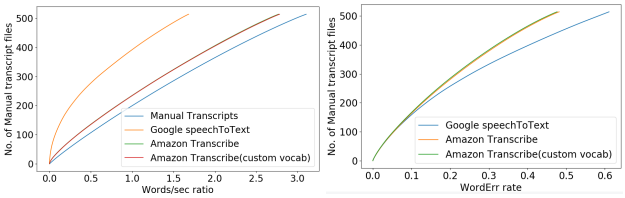
Finally, since several questions were very descriptive, content moderators familiar with KAB for several years helped us break each question into three parts: Likely non-relevant information given by the users such as their name or location, possibly relevant contextual information such as the user's age or gender, and the actual question. An example is shown in Figure 3.

4 METHODOLOGY

Our FAQ retrieval task can be framed as follows. The FAQ database is organized in terms of (Q, A) pairs. When a user ask a new query (q), it needs to be matched against the (Q, A) database to find an appropriate answer. In other contexts, Question-question (Q-q) similarity is known to perform better than (A-q) similarity [21];

सेक्स(sex) -> टेस्ट(Test) , सेट(Set)
यौन सम्बन्ध(Yaun Sambandh) -> जो समन(Joe Summon)
श्रेणी(Shreni) -> सिनेमा(Cinema)
गर्भधारण(Garbhadharana) -> घर उधारण(Ghar Udharan)
एड्स(AIDS) -> एट(ET), एक(Ek)
हस्तमैथुन(Hastmaithun) -> हफ्ता में(Hafta me)
माहवारी(Mahavari)-> वहाँ भारी (vaham Bhari)
एचआईवी एड्स(HIV AIDS) -> एक भी एड(Ek Bhi ed)

Figure 1: Incorrectly transcribed words in Hindi. Words in red were transcribed correctly after listing them in a custom dictionary



(a) CDF for words per sec ratio of questions **(b) CDF for word error rate of questions**

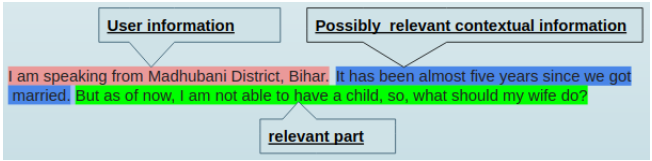


Figure 3: An example of a typical question

we go with this thumbrule and use two methods to compute (Q-q) similarity.

As a baseline, we use Jaccard similarity[3] to find the best (Q-q) match. Jaccard similarity between the query (q) and the candidate set of Questions (Q) can be calculated[4]. We remove Hindi stop words, then we select from the database the best matching question to the query question, and return the corresponding answer.

We compare the baseline with a state-of-the-art FAQ retrieval system that uses the BERT (Bidirectional Encoder Representations from Transformers [6]) model. Unlike other language representation models, BERT first uses a large corpus of data to learn deep bidirectional representations for words (using both the left and right side context of the words) and then further fine-tunes these representations when it is trained in specific domains. The word embeddings generated by BERT are therefore better as they capture the context in which the words are used in the training data, as compared to other pre-trained word representation models like Word2vec which only develop static word embeddings. We then use BERT to compute Q-q similarity by applying the embeddings to a sentence-pair classifier. This sentence-pair classifier is developed as follows: For each positive example (Q,q) in the training data, we randomly select q^- and produce negative training data (Q,q^-), then use these positive and negative examples to train the sentence-pair classifier. BERT, being pre-trained on a large corpus of data, has been shown to achieve good results especially in cases where the data size for tasks within specific domains is small, quite similar to our setting.

[3]https://en.wikipedia.org/wiki/Jaccard_index
[4]https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

5 RESULTS

We conduct the following experiments to study FAQ retrieval feasibility in our setting. For each experiment, we break the question database into a training set (Q) and a test set (q). For answers with more than 10 questions we used a 50/50 training/test split, and for answers with fewer questions we used a 70/30 training/test split. The results are in Table 1.

(1) Both (Q, A) and (q) use transcriptions generated by the Amazon APIs with the custom vocabulary.
(2) Both (Q, A) and (q) use manual transcriptions.
(3) Only the relevant portions of the questions and answers is used with the manual transcripts for (Q, A) and (q).
(4) We divide the training and test sets into ten broad themes that KAB covers: Relationships, Conception and Contraception, Menstruation, Human anatomy, Sexual intercourse related information, Types of sexual behaviour, STDs and HIV, Homosexuality, Abortion, and Breastfeeding. The evaluation is done within each theme, to evaluate a setup where the user first specifies a broad theme and then asks a question. This restricts the search space and can potentially lead to more accurate answers. We do this evaluation using only the relevant portions of the (Q, A) and (q) datasets.
(5) Considering the theme-wise segmentation, we evaluate a case where (q) is automatically transcribed but (Q, A) is manually transcribed. This mimics a real-world setting where in real-time the question asked by the user can be automatically transcribed but the database can be curated and manually transcribed separately.

For each of the five experiments, we also evaluate an approximate match of returning answers that are similar to the actual answer but are still likely to be satisfactory for the user. These approximate matches were developed with help from the KAB team.

All the experiments are evaluated using the following metrics: SR@1 (Success Rate in returning the correct answer), SR@3 (Success Rate in the top-3 results, of having at least one correct answer among the top-3), MRR (Mean Reciprocal Rank, to give greater weight to correct answers ranked higher), and NDCG (Normalized Discounted Cumulative Gain, for a measure of usefulness through graded relevance. Results for both the BERT model and the baseline Jaccard similarity model are given.

We can see that the first experiment which only uses automatically generated transcripts is not able to perform well, indicating the need for manually transcribed and curated data. The second experiment of using manual transcripts works better, but greater success is achieved in the third experiment that eliminates superfluous information from both the questions and answers. The results further improve when the search space is restricted by having the user select in advance the broad theme of interest to them. The approximate-match evaluation of this setup gives an SR@3 close to

Model Used	Exact match					Approximate match				
1) Data transcribed using Amazon Transcribe (with custom vocabulary)										
	SR@1	SR@3	SR@5	MRR	NDCG	SR@1	SR@3	SR@5	MRR	NDCG
BERT_multilingual	0.035	0.082	0.143	0.077	0.031	0.026	0.074	0.121	0.061	0.018
Jaccard Similarity	0.1	0.186	0.229	0.159	0.061	0.108	0.199	0.26	0.174	0.082
2) Manually transcribed data										
	SR@1	SR@3	SR@5	MRR	NDCG	SR@1	SR@3	SR@5	MRR	NDCG
BERT_multilingual	0.117	0.299	0.359	0.227	0.088	0.355	0.511	0.593	0.452	0.143
Jaccard Similarity	0.229	0.329	0.394	0.308	0.113	0.268	0.394	0.463	0.356	0.116
3) Manually transcribed data (using only relevant portions)										
	SR@1	SR@3	SR@5	MRR	NDCG	SR@1	SR@3	SR@5	MRR	NDCG
BERT_multilingual	0.312	0.502	0.57	0.429	0.16	0.452	0.593	0.697	0.545	0.176
Jaccard Similarity	0.29	0.466	0.557	0.402	0.155	0.353	0.548	0.629	0.466	0.158
4) Manually transcribed data (using only relevant portions), segmented into broad themes										
	SR@1	SR@3	SR@5	MRR	NDCG	SR@1	SR@3	SR@5	MRR	NDCG
BERT_multilingual	0.389	0.606	0.71	0.525	0.191	0.471	0.683	0.751	0.59	0.19
Jaccard Similarity	0.362	0.615	0.733	0.513	0.191	0.448	0.697	0.787	0.583	0.197
5) Same as previous with manually transcribed training data, but query data transcribed using Amazon										
	SR@1	SR@3	SR@5	MRR	NDCG	SR@1	SR@3	SR@5	MRR	NDCG
BERT_multilingual	0.271	0.489	0.593	0.41	0.159	0.353	0.548	0.643	0.474	0.157
Jaccard Similarity	0.226	0.416	0.543	0.36	0.136	0.276	0.502	0.633	0.425	0.143

Table 1: Results

70% and SR@5 of up to 75%, which is quite good considering that KAB users tend to browse up to six messages on average when they call the IVR. In the final setting when the query question uses the automatically transcribed text, the performance falls by about 20%. The BERT model does somewhat better than the Jaccard similarity model, but the difference is not large, indicating that the lightweight Jaccard model may be a reasonably satisfactory choice in practice.

6 DISCUSSION AND FUTURE WORK

Our attempt towards building an IVR-based FAQ retrieval system showed that it is sensitive to the accuracy of the speech recognition APIs. It can also benefit from narrowing the search space by asking users to specify the topic in more detail, and by constraining the user to ask crisp questions without stating too much unnecessary information in their questions. These are important lessons and give hints towards what an eventual voicebot design could look like. We show a tentative dialogue flow for the voicebot in Figure 4. To prevent the user from giving superfluous information in their question, we suggest that the voicebot could start with some customary greetings and questions about user familiarization such as asking users their name and location, which will reduce the chances of the user repeating this information again later. This can be followed by a multiple-choice or voice-input question to have the user choose a broad topic from among a few themes. Such intent classification can be done on simple keyword matches. Depending upon the topic that has been selected, some specific context information might also be needed. For example, the KAB team suggested that in the Conception and Contraceptives topic, it could be useful to frame the answer differently depending upon the age of the couple and any history of smoking. Similarly, the age of the user for the topic of Menstruation, and whether or not the user has a stressful lifestyle to handle questions on the topic of Sexual intercourse, can be useful contextual information to seek. Finally, the user can ask the question, and an answer can be returned. We are building the framework to run such a voicebot on IVR systems and integrate it with KAB. To handle the problem with speech recognition on low-quality IVR audio, we also plan to run the voicebot as a smartphone application. We are also considering to develop a text-based chatbot on Facebook Messenger and Whatsapp (using their business APIs). We hope to report studies on user experience in the future.

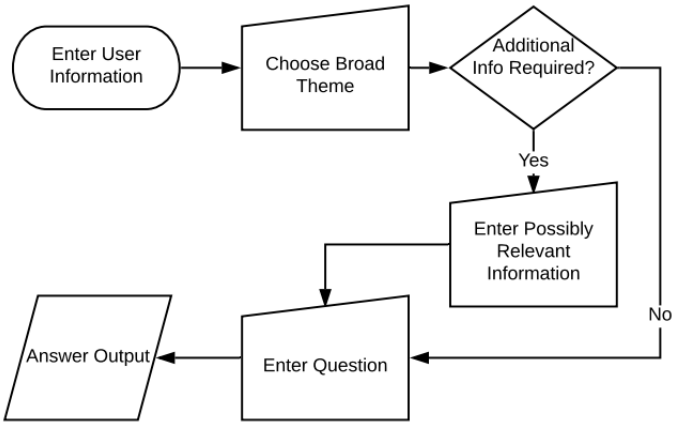


Figure 4: Conversational dialogue flow for the chatbot

REFERENCES

[1] Alaa A. Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features

of chatbots in mental health: A scoping review. *International Journal of Medical Informatics* 132 (2019), 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

[2] Dipanjan Chakraborty, Akshay Gupta, and Aaditeshwar Seth. 2019. Experiences from a mobile-based behaviour change campaign on maternal and child nutrition in rural India. 1–11. https://doi.org/10.1145/3287098.3287110

[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051* (2017). https://arxiv.org/pdf/1704.00051.pdf

[4] CREA. 2005. Adolescent Sexual and Reproductive Health and Rights in India. Working paper. http://www.nipccd-earchive.wcd.nic.in/sites/default/files/PDF/18%20adolescent%20working%20paper.pdf.

[5] CREA. 2013. Kahi-Ankahi-Baatein. https://creaworld.org/events/kahi-ankahi-baatein-mobile-phone-based-info-line.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[7] Karn Dubey, Palash Gupta, Rachna Shriwas, Gayatri Gulvady, and Amit Sharma. 2019. Learnings from deploying a voice-based social platform for people with disability. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. 111–121.

[8] Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 813–820.

[9] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health* 4, 2 (06 Jun 2017), e19. https://doi.org/10.2196/mental.7785

[10] Borut Gorenjak, Marko Ferme, and Milan Ojsteršek. 2011. A question answering system on domain specific knowledge with semantic web support. *International journal of computers* 5, 2 (2011), 141–148.

[11] Kristian Hammond, Robin Burke, Charles Martin, and Steven Lytinen. 1995. FAQ Finder: a case-based approach to knowledge navigation. In *Proceedings the 11th Conference on Artificial Intelligence for Applications*. IEEE, 80–86.

[12] Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. 2018. FarmChat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–22.

[13] Anirudha Joshi, Mandar Rane, Debjani Roy, Nagraj Emmadi, Padma Srinivasan, N. Kumarasamy, Sanjay Pujari, Davidson Solomon, Rashmi Rodrigues, D.G. Saple, and et al. 2014. Supporting Treatment of People Living with HIV / AIDS in Resource Limited Settings with IVRs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1595–1604. https://doi.org/10.1145/2556288.2557236

[14] Courtney Linder. 2019. Amazon Is Crowdsourcing Alexa's Answers, So This Should Be Fun. https://www.popularmechanics.com/technology/a29086631/alexa-answers-crowdsourcing/.

[15] Saurav Kumar Mishra, Dhirendra Bharti, and Nidhi Mishra. 2018. Dr. Vdoc: A Medical Chatbot that Acts as a Virtual Doctor. *Research & Reviews: Journal of Medical Science and Technology* 6, 3 (2018), 16–20.

[16] Aparna Moitra, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. 2016. Design lessons from creating a mobile-based community media platform in Rural India. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. 1–11.

[17] Panitan Muangkammuen, Narong Intiruk, and Kanda Runapongsa Saikaew. 2018. Automated Thai-FAQ Chatbot using RNN-LSTM. In *2018 22nd International Computer Science and Engineering Conference (ICSEC)*. IEEE, 1–4.

[18] Preeti Mudliar, Jonathan Donner, and William Thies. 2012. Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*. 159–168.

[19] PAJ. 2018. Labour Reform is Fine But Who Holds Employers to Account When Government Fails? https://thewire.in/labour/rights-at-work-who-holds-employers-to-account-when-the-government-fails.

[20] Neil Patel, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan Parikh. 2010. Avaaj Otalo - A field study of an interactive voice forum for small farmers in rural India. *Conference on Human Factors in Computing Systems - Proceedings* 2, 733–742. https://doi.org/10.1145/1753326.1753434

[21] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1113–1116.

[22] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. LSTM-based Deep Learning Models for Non-factoid Answer Selection. *arXiv preprint arXiv:1511.04108* (2015).

[23] DS Wang. 2010. A domain-specific question answering system based on ontology and question templates. In *2010 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*. IEEE, 151–156.

[24] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer Extraction as Sequence Tagging with Tree Edit Distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 858–867. https://www.aclweb.org/anthology/N13-1106

ACKNOWLEDGEMENT

We are extremely thankful to the CREA and Gram Vaani teams working on the Kahi Ankahi Baatein project, in particular Rupsa Malik from CREA for allowing us access to the data, and the Gram Vaani team of Sangeeta Saini, Paramita Panjal, and the moderators, who helped build the QA dataset. We are also thankful to the IIT Delhi HPC (High Performance Computing) facility for the computation resources, and to Navneel Mandal and Aman Khullar for helping think through the voicebot design.