AgriBot: Agriculture-Specific Question Answer System

Naman Jain*, Pranjali Jain*, Pratik Kayal*, Jayakrishna Sahit*, Soham Pachpande*
Jayesh Choudhari, Mayank Singh
Indian Institute of Technology Gandhinagar, Gujarat, India 382 355
{naman.j,pranjali.jain,pratik.kayal,jayakrishna.sahit,pachpande.soham,choudhari.jayesh,singh.mayank}@iitgn.ac.in

ABSTRACT

India is an agro-based economy and proper information about agricultural practices is the key to optimal agricultural growth and output. In order to answer the queries of the farmer, we have build an agricultural chatbot based on the dataset from Kisan Call Center. This system is robust enough to answer queries related to weather, market rates, plant protection and government schemes. This system is available 24*7, can be accessed through any electronic device and the information is delivered with the ease of understanding. The system is based on a sentence embedding model which gives an accuracy of 56%. After eliminating synonyms and incorporating entity extraction, the accuracy jumps to 86%. With such a system, farmers can progress towards easier information about farming related practices and hence a better agricultural output. The job of the Call Center workforce would be made easier and the hard work of various such workers can be redirected to a better goal. The prototype of this system is hosted at https://agribotchat.appspot.com/.

This work received the third best paper award and was accepted at the International Conference on STEM, Vibrant Gujarat, Gujarat, India, 2019. For full paper: https://indiarxiv.org/3qp98/.

1 INTRODUCTION

In India, agriculture plays an important role in the economic development by contributing about 16% to the overall GDP and accounting for employment of approximately 52% of the Indian population[2]. However, most farmers do not have access to authentic information about the latest farming practices and trends. One of the reasons is that the people involved in the occupation of farming are comparatively slow adopters of latest technology.

Traditionally, field officers visit the farmlands and provide training, advice, and support to the farmers. Many of the rural villages lack the ease of accessibility resulting in the wastage of time and money spent on obtaining information or contacting officials. Hence, farmers are often unable to obtain agricultural information which can help them in taking better decisions related to the crops that they cultivate. This leads to reduced crop yield, increased wastage of valuable labor, and market inefficiency. This problems worsens due to the spread of misinformation.

To address the issue, Government has initiated the Kisan Call Center (KCC). KCC is a helpline service for farmers to clarify their queries over the phone. Since the service facilitates a telephonic conversation, this service provides a customised solution relevant to each farmer in their native language. This service builds the trust of the farmer on the Government by removing the information barrier. However, KCC services are only available from 6 AM to 10 PM, and skilled labor with good knowledge of agricultural practices is required to operate the Call Center. Our analysis of the KCC's data showed that about 1.36 million calls were made to KCC in 2017 which increased to about 1.72 million calls in 2018. Exponential increase in the queries to KCC have generated the need to set up new call centers which will require massive cost along with training the human resource. Interestingly, only 5% unique new queries were made in 2018 compared to 2017. To account for the redundancy and manual effort, there is an opportunity for automated mobile-based systems to support the industry.

In a nutshell, right information is crucial for social and economic activities which fuel the development process of a nation. In order to achieve it, we require a decision support solution as simple as a messaging app which makes use of Internet to ease accessibility and automate the process of the conversation with an operator to avoid redundancy. Also, the system should integrate features like real-time outputs, farmer-friendly interface, information delivery in multiple languages, and cost-effectiveness for both the farmer and the operating authority. Such a solution can potentially bridge the information gap for the farmers to facilitate in building a productive market.

2 DATA COLLECTION AND ANALYSIS

We collected data from all Indian states for a period of 5 years from https://data.gov.in through an automated program in multiple files containing *the query ID, the query, query-type, query creation time, state name, district name, season* and *the answer to a given query*. The data is not properly formatted and machine readable because it is a two-line summary of the telephonic conversation between the KCC employee and the farmer.

One of the most critical aspects of the data is that it is multilingual. We observe that some words in the data were written in the native language of a particular state and in some cases the entire data entry had been written in the native language. In addition to that, the data entries do not have proper grammar, spelling or punctuation. Another important aspect about the data is the ambiguity in the responses to the queries. Most of the answers do not completely describe the information asked in the question some also being just numerical and same questions had different responses in different states. For our study, we focused only on English queries, pre-process data for machine readability and assumed that the data is correct in terms of the information.

We explore various features to understand and gain statistical insights of the data. The analysis give a good picture of the agricultural landscape of India regarding which crops are popular in which state, what kind of queries are most commonly asked, and the different sectors the queries were related to. We notice that the number of queries related to weather is about 64.4% of the total number of queries asked. For such queries, our model deals differently by through a weather API.

*Equal Contribution.

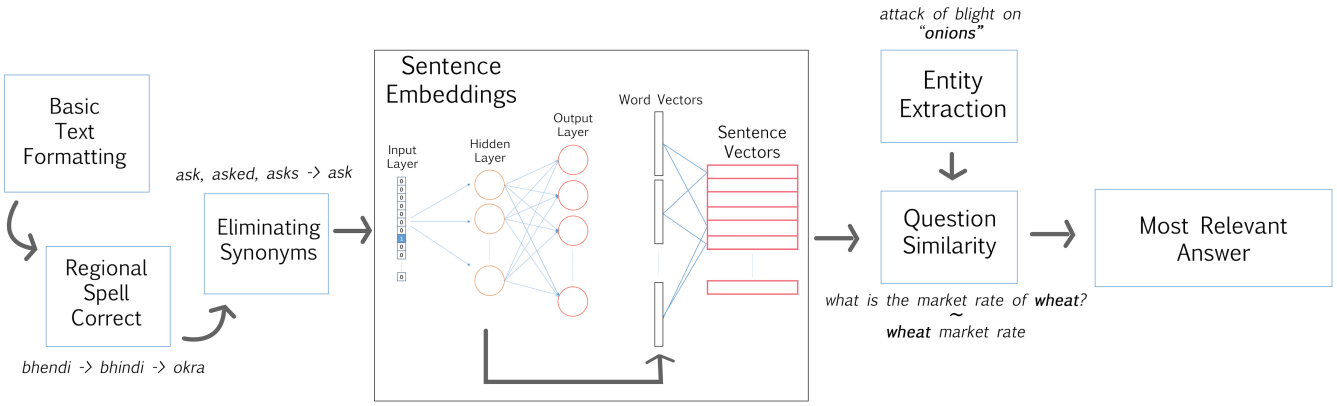

Figure 1: Overview of the Sen2Vec Based Model

In order to check the accuracy of our system, we needed a dataset with ground truth corresponding to each query which does not exist. Lack of truth values made it difficult to determine if the answer for a given input query is correct or not. Such a metric is necessary to measure the reliability of our model. Hence, the determination of a suitable metric for the model is a significant task.

3 METHODOLOGY

Considering the improper format of the queries, we attempt to match input question to all the questions which are present in our dataset rather than processing the answers. The reason behind this is that due to the amount of redundancy, the question is highly likely to be present in the database. We divided the collected data into - train and test. Using the training data, we train our model based on Sen2Vec[1] and then for each query in test data we find the most similar question indexed in the training data.

The overview of the system is shown in Figure 1. We clean the data, develop our spell correct for local language words and normalise words using synonyms. To remove redundancy, we group all answers for a particular question into a list. Finally, the queries were separated into the train (80%) and test (20%). These queries are converted to sentence embedding using the method described by Arora et al. [1]. To enhance the model, we identify crop names using the Entity Extractor and give them higher values as shown in Figure 1. This helps in distinguishing similar questions with different crop names. The model outputs the most similar query from the training data by comparing the embedding vectors using cosine similarity. We apply an answer ranking method to output the best answer from the list of answers corresponding to the model's output.

As far as we know, none of the standard scientific metrics to be suitable for evaluating our model due to the improper and inconsistent structure of question-answer pairs regarding language usage. We would have to design a new metric. We evaluated our model by taking inspiration from Jaccard and Lesk similarity metrics. In order to evaluate our metric, we manually labeled 100 test data queries and calculated our modified Jaccard scores and modified Lesk scores for the prediction of the test data questions. Using these predictions and the ground truth, we then define a threshold for both scores. The threshold tells the model which predictions are to be considered as good results.

4 RESULTS AND DISCUSSION

Using the defined metrics, our model is able to obtain an accuracy of about 56% without synonym normalisation and entity extraction. One key observation was that the crop names were important determiner while comparing the most similar queries. We observe that the accuracy jumps from 56% to 86% after using entity extraction. Also, we note that our chatbot can only answer pre-existing questions in the database. As absolutely new questions cannot be answered by our system, we plan to re-route such queries to human employees for answers and newly created question-answer pair can be then added to the database.

Our chatbot can positively impact under served communities by solving queries related to agriculture, horticulture and animal husbandry using natural language technology. The farmer will be able to receive agricultural information as well as localized information such as the current market prices of various crops in his/her district and weather forecast through a messaging app. Our system would enable the farmer to ask any number of questions, anytime, which will in turn help in spreading the modern farming technology to many farmers rapidly.

Moreover, we found that most of the queries related to localized information such as weather and market prices were redundant which meant our system can answer maximum queries on its own without any human intervention with high accuracy. This will lead to better utilization of human resource and avoid unnecessary costs in setting up new call centers. The system also provides an option that enables the farmer to ask questions directly to the KCC employees if and when necessary. Above all, we believe that the system helps in analyzing the farmers' mindset and the structure of the agricultural sector in India by helping policy makers to understand the needs and concerns of the farmers. The data analysis also provides an understanding of which sector or season farmers requires attention.

For the future, we plan to implement multilingual support for the chatbot with voice-over support and entity extraction from answers for generating knowledge graphs to answer new questions.

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. (2016).

[2] Government of India. 2013. Farmers Portal. https://www.farmer.gov.in/