CSCE 771: NLP

Chatbot COVID Generator

Francisco Vilchez

November 14, 2020

1 Background

The usage of chatbots for enabling access to information for regular users have already been analyzed in previous works. Its abilities to adapt to different domains and data sources make them a good choice to engage different type of users. Due to the increment of open data resources, they are considered a good option where insights from the data can be helpful for different users and its information should be easily accessible. This approach fits perfectly for solving our problem since the system can dynamically create one or more queries to answer the user's inquiry [1].

Additionally, it is been analyzed scenarios where chatbots need to be configured based on user's specific preferences, like language or locations, which is referred as *personal chatbots*. Previous works have analyzed different approaches and architectures to achieve this result and the requirements that a personal chatbot should comply in order to be useful for the users [2].

On the other hand, Python[1] is a language with a set of tools that can help us to achieve our goal. Libraries like Tensorflow[2] provides us the ability to create *Neural Networks* for finding patterns in a predefined set of responses to simulate the conversational behavior of our chatbot. Spacy[3] allows us to process text through different NLTK processes like POS Tagging, Dependency Parse and Named Entities. NLTK[4] helps us with preprocessing tasks like lemmatization. Finally, libraries like Numpy[5] or Pandas[6] help us to manipulate dataset of datas in a more easy and efficient way.

2 Problem

Due to the pandemic, a high number of resources and datasets regarding COVID-19 appeared, making it one the topics that most of the population is interested in the current days. However, despite of the great amount of information, the data is not easily accessible for the needs of every user. If a user needs to get a specific information for COVID-19, they will likely have to learn how to use a specific user interface, which is not always easy to learn or does not always provide the functionality that they need. On other scenarios, the users are even expected to know a programming language in order to be able to fetch the data that they need and do a further analysis.

A chatbot provides a more natural way for users to interact with data and use it according to their needs. Our chatbot generator will allow people to generate chatbots focused on a predefined set of resources targeted to the userś location and be able to communicate with it using the language that they prefer.

[1] Python Programming Language: python.org

[2] Machine Learning platform developed by Google: tensorflow.org

[3] Natural Language Processing library: spacy.io

[4] Text Processing library: nltk.org

[5] numpy.org

[6] pandas.pydata.org

3 Method/Solution Steps/Algorithm

In this project, a retrieval-based approach was used for the creation of the chatbot, with a neural network for the response decision process. The details for each component and the algorithms used will be detailed in this section. Each of the steps followed during this work could be improved in the future in order to get better accuracy results.

3.1 Responses dataset

A dataset in JSON format was created manually in English language. Each JSON item represented a response that our chatbot should use for the user's utterance. The structure of the JSON object goes as follow:

Listing 1: JSON object for a chatbot response

```
{
  ”tag”: ”An ID for the chatbot response”,
  ”patterns”: ”List of questions that trigger this response”,
  ”responses”: ”List of responses for the questions”,
  ”context”: ”(Optional) Next tag that follows this response”,
  ”action”: ”(Optional) An action performed by chatbot”
}
```

3.2 Translation

This step generates a dataset in a different language than English. This process is only performed in case the language selected by the user for the chatbot is different than English. If that is the case, this new dataset will be the one used in the following sections instead of the English dataset.

3.3 Preprocessing

The set of possible chat responses are preprocessed in order to use them for feeding the neural network and make a more accurate prediction. Since the input to the chatbot will be the user's utterance, we need to preprocess it and use it later in order to determine to which tag it corresponds.

From our dataset of responses, we already have a set of *patterns* (list of questions in our dataset) with their respective *tag* (an ID assigned to the response). So can use them to train a Neural Network and predict with response will correspond to a different question from the user. However, we need to clean our questions in order to improve the accuracy of our responses. For that we used two process:

- Stemming: For reducing each of the words in the question to their *stem* and that way generalize the question to other similar questions with a few word tense differences.
- Bag of Words: We encode the questions using a Bag of Words representation, which will be fed to the Neural Network

3.4 Training and Testing dataset

Our initial dataset is separated into two different datasets. Their details are explained bellow.

- Training dataset: It contains all the questions from our original dataset except one for each tag. These questions will be used to train the neural network.
- Test dataset: It contains only one question per tag. These questions will be used to test the Neural Network trained with the Training dataset.

3.5 Neural Network

Our neural network is trained based on our dataset of questions so it can predict the proper response for any utterance that the user may ask. The components of our Neural Network are described bellow.

- Input Layer: The number of input units is equal to the length of our vocabulary, i.e. the length of the representational array in our bag of words.
- Hidden Layers: A first hidden layer with 128 neurons and a second hidden layer with 64 neurons were included. Each of the have a *relu* activation function and a 50% dropout.
- Output Layer: The number of units in our output layers is equal to the number of possible *tags* for our questions. A *softmax* activation layer was included.
- Loss function: *Categorical crossentropy* was used due to its good results in multi-task classification tasks.
- Optimizer: *Stochastic Gradient Descent* (SGD) was used as the iterative method for optimizing the objective function.

3.6 Data Sources

The conversational capabilities of our chatbot are based on data sources for three different aspects:

- COVID: Information related to COVID, its symptoms and recommendations were taken from CDC Website[7].
- Statistics: Information regarding COVID cases per country, including information as total cases, today cases, deaths, today death, etc. were taken from Novel COVID19 API[8]
- Hospitals: Information about hospitals available for each country (not all countries supported), were taken from Healthsites API[9].

4 Demonstration

The capabilities of our chatbot for both languages are displayed below:

Listing 2: English Chatbot Test Case

```
Welcome to COVID Chatbot Generator. Create your chatbot as desired.
Type ‘quit‘ to exit or press ‘enter‘ to continue

System (Covid:Config)   $: Type your language preference (EN/ES)
User                    $: EN

System (Covid:Config)   $: Type your country:
User                    $: India

Training model...
[..Model training output...]
[Screen clears]
Welcome to the COVID chatbot

User                    $: Hi
System (Covid:EN:India) $: Good to see you again
```

[7] CDC COVID Information website: https://www.cdc.gov/coronavirus/2019-ncov/index.html

[8] Documentation available at disease.dh/docs

[9] Documentation available at https://healthsites.io/api/docs/

```
User                    $: What is COVID
System (Covid:EN:India) $: COVID-19 is an infectious disease caused by
    severte acute respiratory syndome coronaravirus 2

User                    $: What are covid symptoms
System (Covid:EN:India) $: COVID-19 affects different people in different
    ways. Infected people have had a wide range of symptoms reported  from
    mild symptoms to severe illness.

User                    $: Show me covid statistics
System (Covid:EN:India) $: Loading COVID metrics for a specific location...
Displaying COVID metrics for India as of 2020-10-26
Today cases: 45158
Total cases: 7909050

User                    $: Show me hospitals near me
System (Covid:EN:India) $: Loading COVID hospitals...
Displaying hospitals for India. Hospital 1 of 100:
> Dr.Vaibhav clinic
Type ‘n‘ to display the next result or ‘q‘ to end: n
Displaying hospitals for India. Hospital 2 of 100:
> Government Hospital Sainthal
Type ‘n‘ to display the next result or ‘q‘ to end: q

User                    $: Bye
System (Covid:EN:India) $: Have a nice day

Returning to Chatbot Generator Configuration Mode
[Process returns to the beginning]
```

Listing 3: Spanish Chatbot Test Case

```
Welcome to COVID Chatbot Generator. Create your chatbot as desired.
Type ‘quit‘ to exit or press ‘enter‘ to continue

System (Covid:Config)   $: Type your language preference (EN/ES)
User                    $: ES

System (Covid:Config)   $: Type your country:
User                    $: Peru

Generating dataset...
Translating intents (11 left)
Translating intents (10 left)
Translating intents (9 left)
Translating intents (8 left)
Translating intents (7 left)
Translating intents (6 left)
Translating intents (5 left)
Translating intents (4 left)
Translating intents (3 left)
Translating intents (2 left)
Translating intents (1 left)
Saving intents file...
Dataset created!
Training model...
[..Model training output...]
[Screen clears]
User                    $: Que es COVID
System (Covid:ES:Peru)  $: COVID-19 es una enfermedad infecciosa causada
    por el coronavirus 2 sindrome respiratorio agudo severo

User                    $: Cuales son los sintomas del covid
System (Covid:ES:Peru)  $: COVID-19 afecta a diferentes personas de
    diferentes maneras. Se ha informado de una amplia gama de sintomas de
    personas infectadas, desde sintomas leves hasta enfermedades graves.

User                    $: Muestrame los casos covid
System (Covid:ES:Peru)  $: Cargando metricas de COVID para una ubicacion
    especifica ...
Mostrando metricas de COVID para Peru al 2020-10-26
Casos totales: 2501
Casos de hoy: 888715

User                    $: Muestrame hospitales cercanos
System (Covid:ES:Peru)  $: Cargando hospitales COVID ...
Mostrando hospitales para Peru. Hospital 1 de 100:
> Botica SanaFarma
Escriba ‘n‘ para mostrar el siguiente resultado o ‘q‘ para terminar: n
Mostrando hospitales para Peru. Hospital 2 de 100:
> Boticas Peru
Escriba ‘n‘ para mostrar el siguiente resultado o ‘q‘ para terminar: q

User                    $: Adios
System (Covid:ES:Peru)  $: Nos vemos!

Regresando al modo de configuracion de Generacion de Chatbot
[Process returns to the beginning]
```

Evaluation

As mentioned in the previous section, part of the dataset was randomly used to test the accuracy of our Neural Network model. Based on the approach followed an average of 60% accuracy was obtained by our model. An example of the evaluation results for each language model are presented below:

Table 1: Evaluation for English chatbot

Question asked	Answer category returned by chatbot	Answer category that should be returned
Hi	greeting	greeting
Hello	greeting	greeting
What is COVID	covid_information	covid_information
Give me information about COVID	covid_metrics_in_location	covid_information
What are COVID symptoms	covid_symptoms	covid_symptoms
Tell me about COVID symptoms	covid_symptoms	covid_symptoms
Show me COVID statistics	covid_metrics_in_location	covid_metrics_in_location
Give me COVID statistics	covid_metrics_in_location	covid_metrics_in_location
Can you tell me hospitals nearby	search_covid_hospitals_by_country	search_covid_hospitals_by_country
Show me hospitals close to me	search_covid_hospitals_by_country	search_covid_hospitals_by_country
Bye	goodbye	goodbye
See you later	options	goodbye

Table 2: Evaluation for Spanish Chatbot

Question asked	Answer category returned by chatbot	Answer category that should be returned
Hola	greeting	greeting
Cómo estás	greeting	greeting
Qué es el COVID	covid_information	covid_information
Háblame del COVID	covid_information	covid_information
Cuáles son los síntomas de COVID	covid_symptoms	covid_symptoms
Dime los síntomas de COVID	covid_symptoms	covid_symptoms
Puedes decirme las estadísticas de los casos de COVID	covid_metrics_in_location	covid_metrics_in_location
Dame estadísticas de COVID	covid_metrics_in_location	covid_metrics_in_location
Puedes decirme los hospitales cercanos	search_covid_hospitals_by_country	search_covid_hospitals_by_country
Muéstrame los hospitales cerca de mí	covid_symptoms	search_covid_hospitals_by_country
Adiós	goodbye	goodbye
Nos vemos más tarde	greeting	goodbye

For both cases, we could see that from the total of 12 questions asked, it retuned 10 correct answers which is equal to 83% of accuracy. These results were randomly realized, because of that the accuracy could change.

Discussion

This project provided the ability for users to create chatbots for giving COVID information based on the configuration they selected to match their needs, which provides different benefits. On one side, it gives users from different locations the ability to interact with different datasets (in this case, COVID informations and data) in a more intuitive way. On the other side, it generalizes the way chatbots are created in order to make it feasible to adapt it to any other type of information different than COVID. This flexibily is important since the ammount on datasets and information that appear in internet will keep growing, so it is important to give people the ability to interact with any new data source that may appear in an easy and intuitive way.

During its developments, we faced an issue with the dependency with the different structures for the data received by external APIs, which control the functionalities of covid statistics and hospitals offered in our chatbot. This involved an additional data structure customization in our chatbot that could enable us to still have the ability to generate chatbots in different languages without compromising the functionality of the APIs. Additionally, the dependency on external tools for NLP features like spacy.io, which its performance changes drastically between languages, made us keep using features from English spacy.io tool even for different other languages like entity recognition.

We are expecting to keep improving this project so the accuracy is increased and get better results. We are considering using a different data representation model for our chatbot sentences. Additionally, a deeper analysis in the structure of our neural network could improve the results obtained. On the other hand, support for a broader amount of languages are easily achievable with the usage of our translation APIs. Futhermore, we expect to make the APIs for retrieving information customizable in order to allow users to decide which datasource should retrieve their COVID information or statistics. Finally, we consider that users should also have a learning mode in which they can increase the capabilities of the chatbots generated and that way expand the options they can offer.

References

[1] B. Srivastava, “Decision-support for the masses by enabling conversations with open data,” *arXiv preprint arXiv:1809.06723*, 2018.

[2] F. Daniel, M. Matera, V. Zaccaria, and A. Dell’Orto, “Toward truly personal chatbots: on the development of custom conversational assistants,” in *Proceedings of the 1st International Workshop on Software Engineering for Cognitive Services*, pp. 31–36, 2018.