Chatbot for IEEE SB NIT Durgapur

Ridam Hazra

May 11, 2021

Background:

The usage of chatbots for enabling access to information for regular users have already been analysed 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]

Problem:

The visitors of the website have standard questions that usually cover the same set of topics. Answering questions takes much times and the answers may be delayed for many days. So, an automated system must be created for the smooth run of the whole thing.

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.

Response Dataset: -

- NLU Dataset:

A dataset in YML format was created manually in English language for the NLU part. Each YML item represented an intent that the user of the chatbot might use. The structure of the YML object goes as follows:

```
- intent: contact_us
  examples: |
    - how can I contact you?
    - how can I reach you?
    - how to reach you through social media?
    - tell me how to reach you
    - tell me how to contact you
    - are you available in social media?
    - can I get your social media links?
    - can I get your social media handles?
```

- Rules Dataset:

A dataset in YML format was created manually in English language for the pre-defined rules for the chatbot. Each YML item represented a rule for the chatbot. The structure of the YML object goes as follows:

```
rules:

- rule: Say goodbye anytime the user says goodbye
  steps:
  - intent: goodbye
  - action: utter_goodbye

- rule: Say 'I am a bot' anytime the user challenges
  steps:
  - intent: bot_challenge
  - action: utter_iamabot
```

- Stories Dataset:

A story is a representation of a conversation between a user and the chatbot, converted into a specific format where user inputs are expressed as intents (and entities when necessary), while the assistant's responses and actions are expressed as action names. A dataset in YML format was made for stories which goes as follows:

```
- story: contact path 1
  steps:
  - intent: greet
  - action: utter_greet
  - intent: contact_us
  - action: utter_contact_us
  - action: utter_did_that_help
  - intent: affirm
  - action: utter_happy

- story: contact path 2
  steps:
  - intent: greet
  - action: utter_greet
  - intent: contact_us
  - action: utter_contact_us
  - action: utter_did_that_help
  - intent: deny
  - action: utter_no_more
```

- Domain Dataset:

Under domain dataset all possible intents and responses of the Chatbot was given. Example:

```
intents:
  - greet
  - goodbye
  - affirm
  - deny
  - mood_great
  - mood_unhappy
  - bot_challenge
  - event
  - contact_us
  - contact_fb
  - contact_insta
  - contact_linkedin

responses:
  utter_greet:
  - text: "Hey! Welcome to the website of IEEE SB NIT Durgapur. How can I help you?"

  utter_cheer_up:
  - text: "Here is something to cheer you up:"
    image: "https://i.imgur.com/nGF1K8f.jpg"

  utter_event:
  - text: "Here are all of our [activities](https://www.ieeesbnitdgp.com/thereg/)"

  utter_contact_us:
  - text: "Here are all of our social media links: 1.
[Facebook](https://www.facebook.com/nitdgpieeesociety/) 2.
[Instagram](https://www.instagram.com/ieeesb_nitdgp/) 3.
[LinkedIn](https://www.linkedin.com/company/ieeesbnitdgp/)"

  utter_contact_fb:
  - text: "Here is our facebook link: [Facebook](https://www.facebook.com/nitdgpieeesociety/)"

  utter_contact_insta:
  - text: "Here is our instagram link: [Instagram](https://www.instagram.com/ieeesb_nitdgp/)"

  utter_contact_linkedin:
  - text: "Here is our linkedin link: [LinkedIn](https://www.linkedin.com/company/ieeesbnitdgp/)"

  utter_did_that_help:
  - text: "Did that help you?"

  utter_happy:
  - text: "Great, carry on!"

  utter_goodbye:
```

```
- text: "Bye"

utter_iamabot:
- text: "I am a bot, working for IEEE SB NIT Durgapur, powered by Rasa."

utter_no_more:
- text: "Sorry, that's all I have!"
```

Pre-processing: -

The set of possible chat responses are pre-processed 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 pre-process 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 examples (list of questions in our dataset) with their respective tag (an ID assigned to the response). So that we can use them to train a Neural Network and predict which 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 words tense differences.
- Bag of Words: We encode the questions using a Bag of Words representation, which will be fed to the Neural Network.

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.

The following default pipeline was used to train our model:

```
- name: WhitespaceTokenizer
- name: RegexFeaturizer
- name: LexicalSyntacticFeaturizer
- name: CountVectorsFeaturizer
- name: CountVectorsFeaturizer
  analyzer: char_wb
  min_ngram: 1
  max_ngram: 4
- name: DIETClassifier
  epochs: 100
  constrain_similarities: true
  model_confidence: cosine
- name: EntitySynonymMapper
- name: ResponseSelector
  epochs: 100
  constrain_similarities: true
  model_confidence: cosine
```

```
- name: FallbackClassifier
  threshold: 0.3
  ambiguity_threshold: 0.1
```

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 below.

• 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 cross entropy 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.

Demonstration:

The capabilities of our chatbot are displayed below:

```
Bot loaded. Type a message and press enter (use '/stop' to exit):
Your input ->  hello
Hey! Welcome to the website of IEEE SB NIT Durgapur. How can I help you?
Your input ->  tell me about your events
Here are all of our [activities](https://www.ieeesbnitdgp.com/thereg/)
Did that help you?
Your input ->  yeah
Great, carry on!
Your input ->  fb contact please
Hey! Welcome to the website of IEEE SB NIT Durgapur. How can I help you?
Your input ->  give me your facebook contact
Here is our facebook link: [Facebook](https://www.facebook.com/nitdgpieeesociety/)
Did that help you?
Your input ->  instagram id
Here is our instagram link: [Instagram](https://www.instagram.com/ieeesb_nitdgp/)
Did that help you?
Your input ->  thank you
Great, carry on!
Your input ->  /stop
2021-05-12 00:17:39 INFO     root  - Killing Sanic server now.
```

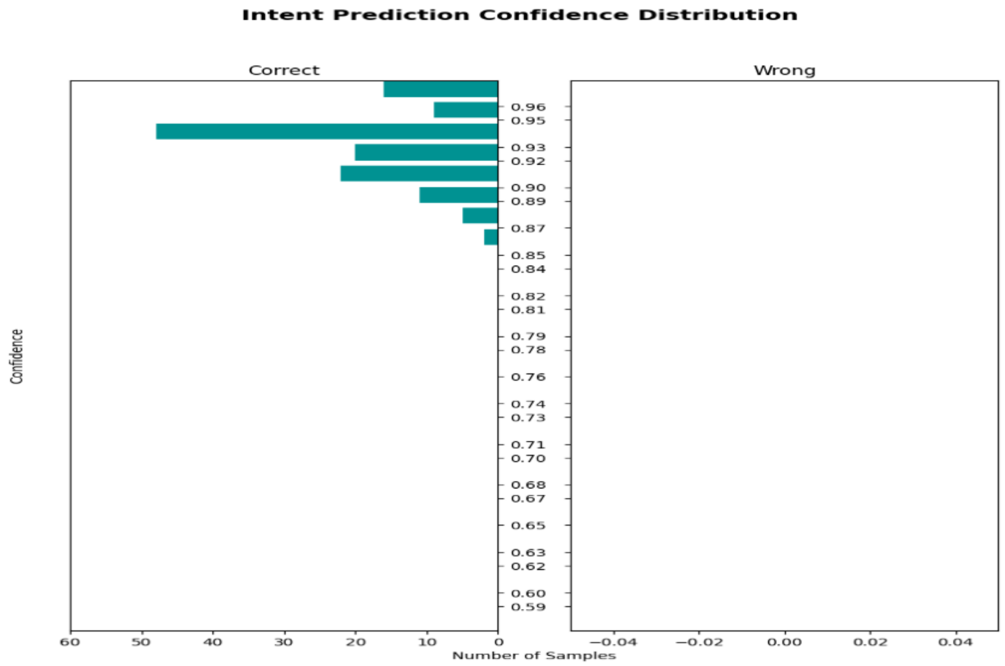
Evaluation:

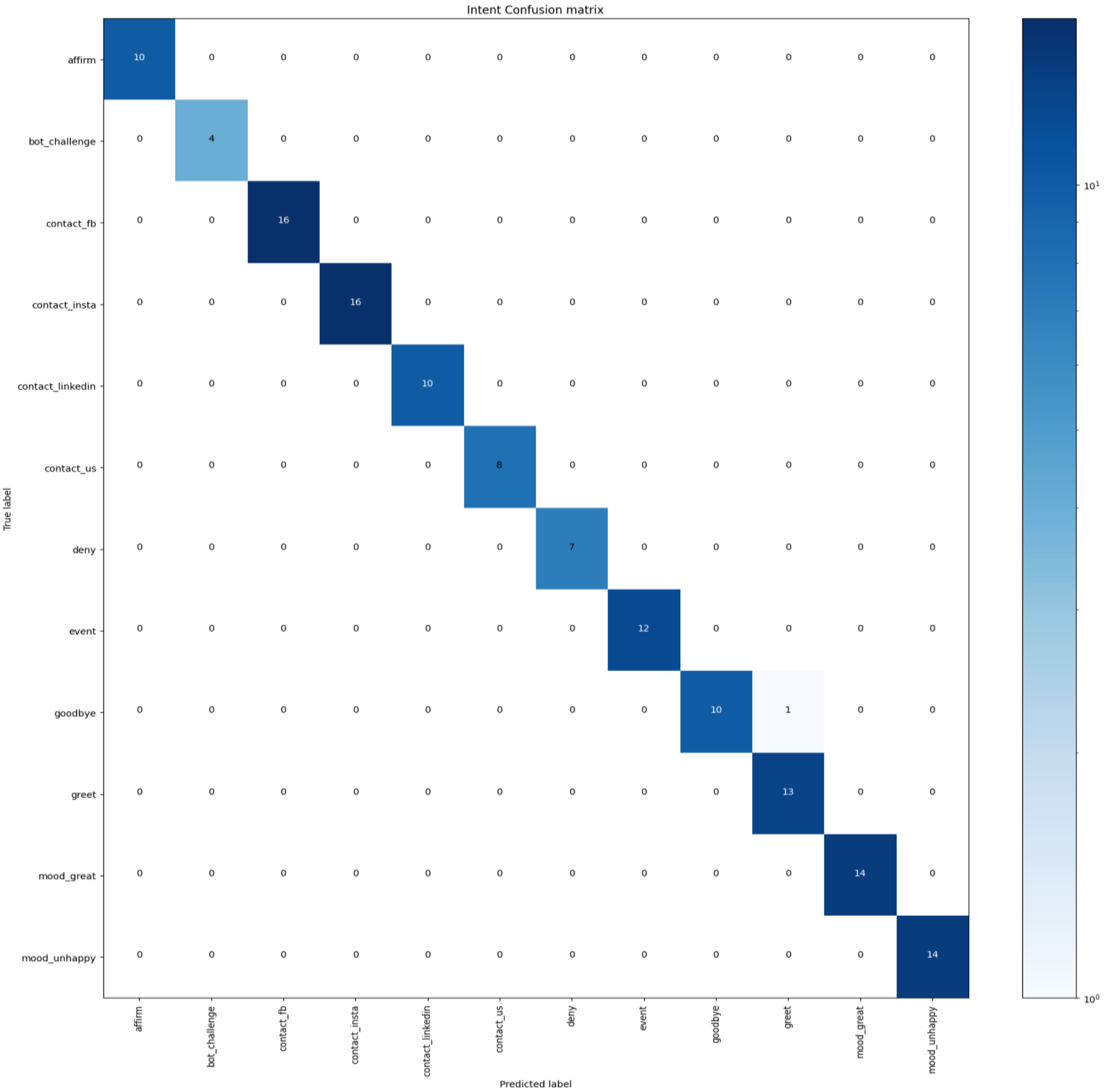
As mentioned in the previous section, part of the dataset was randomly used to test the accuracy of our Neural Network model. Based on the approach followed an average of 99.26% accuracy was obtained by our model. An example of the evaluation results for each language model are presented below:

```
- story: happy path 2
  steps:
  - user: |
      hello there!
    intent: greet
  - action: utter_greet
  - user: |
      amazing
    intent: mood_great
  - action: utter_happy
  - user: |
      bye-bye!
    intent: goodbye
  - action: utter_goodbye
```

Furthermore, this report was obtained:


```
"accuracy": 0.9925925925925926,
  "macro avg": {
    "precision": 0.9940476190476191,
    "recall": 0.9924242424242425,
    "f1-score": 0.9929453262786595,
    "support": 135
  },
  "weighted avg": {
    "precision": 0.993121693121693,
    "recall": 0.9925925925925926,
    "f1-score": 0.9925533999608073,
    "support": 135
  }
```


Intent Prediction Confidence Distribution
Correct
Wrong
Confidence
Number of Samples


Intent Confusion matrix
True label
Predicted label
affirm
bot_challenge
contact_fb
contact_insta
contact_linkedin
contact_us
deny
event
goodbye
greet
mood_great
mood_unhappy

Discussion:

This project provided the ability for users to create chatbots for giving general query-based information based On the need of the user. 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 this particular topic. This flexibility is important since the number of 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.

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.
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.