BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Intelligent assistant for holiday recommendations

– MIRPR report –

Team members
Finiti Sebastian, IE, 933
Dragoi Vlad, IE, 933
Bart Cramba, IE, 932

2019

Abstract

The online booking business is continuously rising with many applications, such as booking and airbnb trying to improve their system to better know their users and help them choose the best holiday place. We are targeting this problem by building a personal assistant that can process your text and images and understand your holiday needs. There is also an idea to develop a recommender system to better recommend holiday destinations based on previously collected data from you and other users.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 What? Why? How?

- What is the (scientific) problem?

  The problem tackles natural language processing and image processing. The user will upload a picture and will give some context to it. The personal assistant will process all this information and will make a recommendation based on it.

- Why is it important?

  It advances a specific area of a personal assistant and also improves the efficiency of searching for the perfect holiday. Scaled up, it can even grow the business of booking applications.

- What is your basic approach?

  We are first building the chatbot and making it responsive. Then we will start with basic algorithms for natural language processing and image processing.

There have been many intelligent agents that gather data from hotels and flights for learning the user's likes and dislikes but not a chatbot that can do what we want to do. In the end we will like to present all these recommendations in a more interactive way that are now presented.

1.2 Paper structure and original contribution(s)

The research presented in this paper advances the theory, design, and implementation of several particular models.

The main contribution of this report is to present an intelligent algorithm for solving the problem of ....

The second contribution of this report consists of building an intuitive, easy-to-use and user friendly software application. Our aim is to build an algorithm that will help ....

The third contribution of this thesis consists of ....

The present work contains *xyz* bibliographical references and is structured in five chapters as follows.

The first chapter/section is a short introduction in ....

The second chapter/section describes ....

The chapter/section 5 details ....

Chapter 2

Scientific Problem

2.1 Problem definition

Give a description of the problem. Explain why it must be solved by an intelligent algorithm. Details the advantages and/or disadvantages of solving the problem by a (some) given method(s).

Precisely define the problem you are addressing (i.e. formally specify the inputs and outputs). Elaborate on why this is an interesting and important problem.

Item example:

- content of item1
- content of item2
- content of item3

Figure example

... (see Figure 2.1)

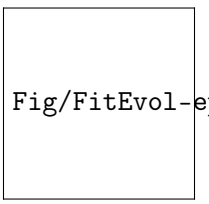

Figure 2.1: The evolution of the swarm size during the GA generations. This results were obtained for the f_2 test function with 5 dimensions.

Table example: (see Table 2.1)

Algorithm example

... (see Algorithm 1).

Table 2.1: The parameters of the PSO algorithm (the micro level algorithm) used to compute the fitness of a GA chromosome.

Parameter	**Value**
Number of generations	50
Number of function evaluations/generation	10
Number of dimensions of the function to be optimized	5
Learning factor c_1	2
Learning factor c_2	1.8
Inertia weight	$0.5 + \frac{rand()}{2}$

Algorithm 1 SGA - Spin based Genetic AQlgorithm

```
BEGIN
@ Randomly create the initial GA population.
@ Compute the fitness of each individual.
for i=1 TO NoOfGenerations do
  for j=1 TO PopulationSize do
    p ← RandomlySelectParticleFromGrid();
    n ← RandomlySelectParticleFromNeighbors(p);
    @ Crossover(p, n, off);
    @ Compute energy ΔH
    if ΔH satisfy the Ising condition then
      @ Replace(p,off);
    end if
  end for
end for
END
```

Chapter 3

Intelligent Algorithms used

3.1 Intent Classification

In order for our chatbot to respond according to the users query, we use "intent classification" and the categories in which a chatbot respond these are known as " intents ". So say you asked for a place to eat then it will respond under that category and if you asked for a place to stay then it will respond under that category and so on.

3.1.1 What is Intent Classification ?

Think about how humans classify everything " jeans are an item of clothing, a guitar is an instrument, a song with a soft rhythm is relaxing. "
In the same way, an intent classifier is able to categorize text based on the intent, goal, or purpose expressed in its content. It does this by using machine learning algorithms that can associate words or expressions with a particular intent. For example, a machine learning model can learn that words such as buy or acquire are often associated with a Purchase intent.
However, these machine learning classifiers need to be trained first with examples. First, you need to define tags or categories that are relevant to the matter at hand. For intent classification, these tags typically refer to actions that customers intend to perform.
For example, if youâre analyzing customer emails, your tags might be something like Interested, Need Information, Unsubscribe, Wrong Person, Email Bounce, Autoreply, etc.
With tags in place, you can begin to train your model and feed it relevant examples for each tag. This way, youâll teach the classifier how to tag new data appropriately. For example: âI tried to make a purchase through the site but I donât know where to start, could you help me out? Iâm really interested in shopping the new collection.â This email can be tagged as Interested.

The more examples you provide the model, the smarter your intent classifier will be since it has more information to learn from.

3.1.2 Data Preparation

This step is very important because simplifying the data as much as we can, helps our model to train easily and faster.

3.1.2.1 Data Cleaning

We are using raw data by importing a CSV file, so we have to clean it before feeding it to our model. There is no definite method for doing that. In our case, we remove every punctuation and special characters (if any) from the data then we tokenize the senteces into words. After this we lowercase all the words and use lemmatization on them.

- Lemmatization

  " Lemmatisation (or lemmatization) in linguistics, is the process of grouping together the different inflected forms of a word so they can be analysed as a single item." In more simple words, a lemmatization is a process in which we get a lemma(actual words) of a word.

```
lemmatizer.lemmatize(" cats ") ==> cat
lemmatizer.lemmatize(" churches ") ==> church
lemmatizer.lemmatize(" abaci ") ==> abacus
```

  This is lemmatization that we are using it so that if someone writes a word differently, classifier can understand it and give us the best result possible.

3.1.2.2 Encoding

- Input Encoding

  After cleaning the data we get lists of words of sentences. To convert these words into indexes so that we can use them as input, we are using Tokenizer class of Keras.

- Output Encoding

  For outputs we do the same thing, first indexed those intents by using class of Keras.

3.1.2.3 Train and Validation set

Data is ready for model, so the final step that we do is to split the dataset into training and validation set. The model is created and trained using keras bidirectional gated recurrent unit algorithm.

Data will consist of 2 columns, one for the sentence and one for the intent. We currently have 21 unique intents and 1113 sentences. We are creating a list of intents, and for each intent we provide a set of training phrases representing what normal user may say for that intent.
The training was performed on 100 recursive passes thorugh the data and split into data sizes of 32 sentences.
Once the bot is trained, the bot's intent classification is evaluated using testing phrases to see if the bot detects the intents correctly.
The benchmark shows that we can improve the intent classification by increasing the size and the quality of the training phrase.

Chapter 4

State of art/Related work

The theory of the methods utilised until now in order to solve the given problem.

Answer the following questions for each piece of related work that addresses the same or a similar problem.

- What is their problem and method?

  Netflix and Youtube are widely known for developing a recommender system that is based on user interaction. Their method is going for explicit feedback, the users are asked to give feedback on the content they watched. This gives them a different approach on how to build a more suitable recommendation system.

- How is your problem and method different?

  For starters, we do not have the possibility of having a huge pool of users for our chatbot and our data will be crawled/collected from the web. This is known as implicit data, we will be training our intelligent agents with data we collect from the internet rather than from users.

- Why is your problem and method better?

  Sometimes, the user may not be honest about the feedback they give. Humans are humans after all, our data will lean more towards statistical data collected by different people that are working on such system. With a large enough pool we can train our intelligent agents to work even better in some cases.

Bibliography

[1] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

Chapter 5

Proposed approach

One idea involved in tackling this problem is understanding the sentences given by the user, that is, representing the tree of the sentence, and performing information extraction, given a known structure of the information.

The upside of this approach is the easiness in understanding the behaviour of the chatbot and in the implementation - the downside may be that, for some users with chaotic style of describing, the bot would not understand the required details in their entirety - however, that would generally be true for any statistical approach.

Chapter 6

Application (numerical validation)

For the first approach, there is no thorough validation to be performed - the intelligent agents are used for POS tagging and parsing into a sentence tree - for this we leverage Natural Language Toolkit (NLTK).

Therefore, for such a sentence structure the understanding rate is 100%.

6.1 Methodology

- What are criteria you are using to evaluate your method?
- What specific hypotheses does your experiment test? Describe the experimental methodology that you used.
- What are the dependent and independent variables?
- What is the training/test data that was used, and why is it realistic or interesting? Exactly what performance data did you collect and how are you presenting and analyzing it? Comparisons to competing methods that address the same problem are particularly useful.

6.2 Data

Describe the used data.

6.3 Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are

they statistically significant?

6.4 Discussion

- Is your hypothesis supported?
- What conclusions do the results support about the strengths and weaknesses of your method compared to other methods?
- How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

Chapter 7

Conclusion and future work

Try to emphasise the strengths and the weaknesses of your approach. What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.

Briefly summarize the important results and conclusions presented in the paper.

- What are the most important points illustrated by your work?
- How will your results improve future research and applications in the area?

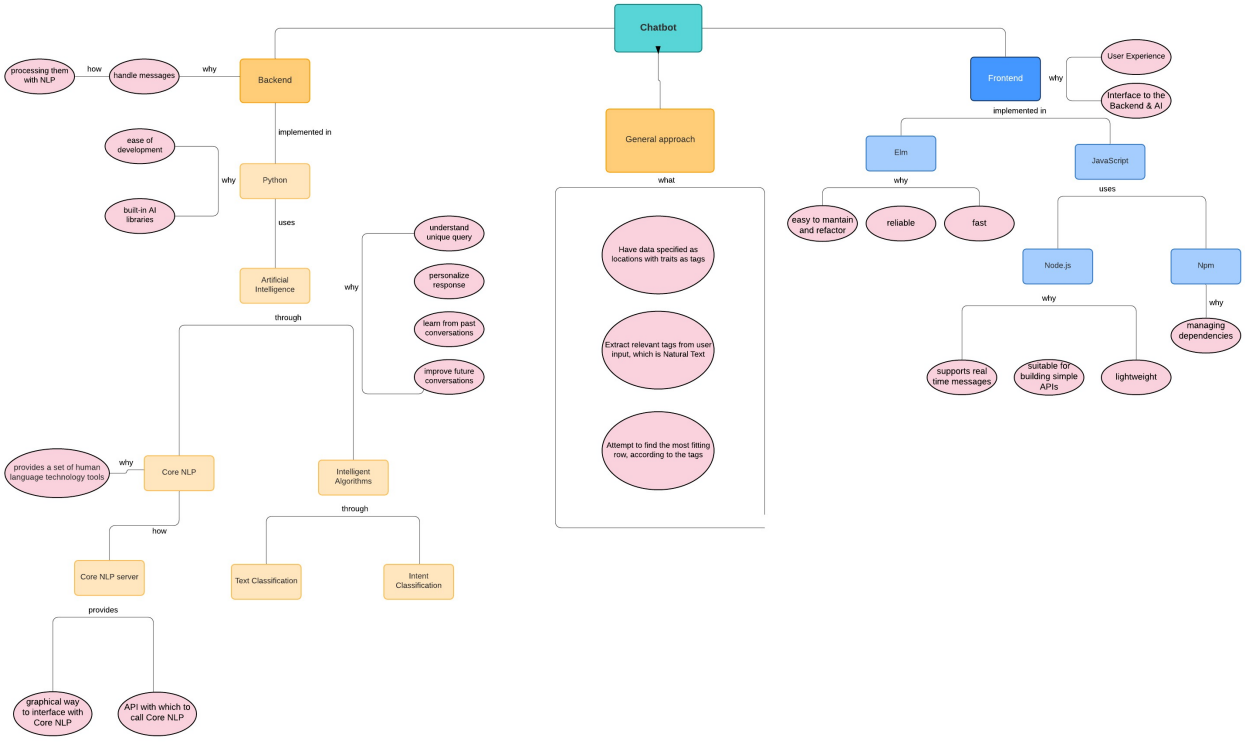

Figure 7.1: The mindmap of the chatbot