AUTOMATED TOURISM KNOWLEDGE GRAPH AND INTENT GENERATION FROM AUDIO CONTENT EXTRACTED FROM VIDEOS, BY UTILIZING NLP

Senuri Sarindi Seneviratne

199486X

Degree of Master of Science

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

July 2022

AUTOMATED TOURISM KNOWLEDGE GRAPH AND INTENT GENERATION FROM AUDIO CONTENT EXTRACTED FROM VIDEOS, BY UTILIZING NLP

Senuri Sarindi Seneviratne

199486X

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

July 2022

Declaration, copyright statement and the statement of the supervisor.

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Master's thesis under my supervision.

Signature of the supervisor: Date:

Abstract

Generating a knowledge graph for a chatbot is a time-consuming exercise which needs the help of an expert relevant to the field. This thesis presents our approach to synthesizes the creation of a knowledge graph and intents for a chatbot. Currently, the creation of a knowledge graph and intents for a chatbot is a tedious process and this process does not extract data from videos. Developing a chatbot also requires the support of experienced software engineers.

This platform allows a user to build a customized chatbot according to a specific requirement in any field, without the intervention of experts. It also allows for the seamless development of a comprehensive knowledge graph from the video content through a simple and less tedious approach. The platform uses Natural Language Processing (NLP) machine learning models such as Naive Bayes and Logistic Regression and grammar correction techniques to supplement the experience of the users.

The working process of this proposed system is Knowledge Extraction and generating the Knowledge Base. The user inserts keywords related to the chatbot's domain as the first step of the process. The system retrieves the search results from YouTube. Finally, NLP will be used to retrieve data contained in videos to create a preliminary knowledge graph and intents for a chatbot. A scheduler is then activated automatically from time to time to update the knowledge graph and intents. The knowledge graph and intents generated have been tested on a chatbot created using the Rasa framework, with the chatbot giving the correct answers when questioned by a user.

Acknowledgment

First and foremost, I would like to extend my sincere gratitude to Dr. Sagara Sumathipala for his invaluable support, guidance and motivation during the implementation of the project and preparation of the thesis. The constructive criticism given to me during the multiple rounds of supervision has helped me to improve and refine my thesis according to the guidelines shared by the University of Moratuwa.

Next, I would like to thank Thilanka Kumarasinghe from my workplace for helping me understand the importance and application of a chatbot for different industries, which motivated me to select this topic for my Master's thesis.

Finally, I would like to extend my gratitude to my parents and the sister for all the morale support given to me and for all sacrifices they made on my behalf to ensure that I succeed in this project. The guidance and motivation they gave me has enabled me to fuel my inner drive to do the absolute best in completing this thesis.

Table of contents

Declaration, copyright statement and the statement of the supervisor........................ i

Abstract ii

Acknowledgment iii

Table of contents iv

List of figures viii

List of tables xi

List of abbreviation xii

1.0 INTRODUCTION 1

1.1 Prolegomena 1

1.2 Aims and objectives 2

1.3 Background and motivation 2

1.4 Problem definition 3

1.5 Your solution 3

1.6 Resource requirements 3

1.7 Summary 3

2.0 CHATBOT – DEVELOPMENTS AND ISSUES 4

2.1 Introduction 4

2.2 Early developments in Chatbots and knowledge bases 4

2.3 Breakthrough in chatbots (Latest/Modern) 7

2.4 Challenges (future trends/issues) in chatbots 9

2.5 Problem definition 11

2.9 Summary 12

3.0 TECHNOLOGIES USED FOR CHATBOTS 13

3.1 Introduction 13

3.2 Technologies of Chatbots 13

3.3 Technologies of Knowledge base development 14

3.4 Technologies of Knowledge graph development 15

3.5 Technologies of Knowledge extraction... 16

3.6 Chatbot Anatomy 16

3.7 Summary 16

4.0 APPROACH 17

4.1 Introduction 17

4.2 Hypothesis 17

4.3 Input.. 17

4.4 Output 17

4.5 Process 18

4.6 Users 19

4.7 Feature 20

4.8 Summary 20

5.0 DESIGN 21

5.1 Introduction 21

5.2 Data retrieval module 22

5.2.1 Data Retrieval from YouTube... 23

5.2.2 Video date filter.. 23

5.2.3 Audio downloader 24

5.2.4 Text converter 24

5.3 Data processing module 24

5.3.1 Data extractor 25

5.3.2 Feature weight assigner.. 25

5.3.3 Logistic model correct triplet extraction 26

5.4 Knowledge graph generation module... 26

5.4.1 Identify basic nodes and edges.. 27

5.4.2 Create knowledge graph.. 27

5.5 Intent generation module 28

5.5.1 User intent and example generator 28

5.3.2 Rasa nlu updater 29

5.6 Knowledge graph and chatbot intent update module 29

5.6.1 Knowledge update scheduler 29

5.7 Summary 30

6.0 IMPLEMENTATION 31

6.1 Introduction 31

6.2 Data retrieval module 31

6.2.1 Data Retrieval from YouTube 31

6.2.2 Video date filter 31

6.2.3 Audio downloader 32

6.2.4 Text converter 32

6.3 Data processing module 32

6.3.2 Feature weight assigner 33

6.2.3 Logistic model correct triplet extraction 38

6.4.1 Identify basic nodes and edges 39

6.4.2 Create knowledge graph 41

6.5 Intent generation module 43

6.5.1 User intent and example generator 43

6.5.2 Rasa nlu updater 43

6.6 Knowledge graph and chatbot intent update module. 47

6.6.1Knowledge update scheduler 47

7.0 EVALUATION 49

7.1 Introduction 49

7.2 Results 49

7.2 Evaluation using matrices 57

7.7 Summary 58

8.0 Conclusion and Further Work 59

8.1 Introduction 59

8.2 Conclusion 59

8.9 Summary 60

References 61

List of figures

Figure 4 1 High level architecture.. 18

Figure 5 1Top level architecture of knowledge base generator 22

Figure 5 2 Data retrieval module architecture.. 22

Figure 5 3 Architecture of data retrieval from YouTube 23

Figure 5 4 Architecture of video date filter.. 23

Figure 5 5 Architecture of audio downloader 24

Figure 5 6 Architecture of text converter... 24

Figure 5 7 Data processing module architecture... 24

Figure 5 8 Architecture of data extractor 25

Figure 5 9 Architecture of feature weight assigner... 25

Figure 5 10 Architecture of logistic model correct triplet extraction........................ 26

Figure 5 11 Knowledge graph generation module architecture 26

Figure 5 12 Architecture of identify basic nodes and edges 27

Figure 5 13 Architecture of create knowledge graph.. 27

Figure 5 14 Intent generation module architecture 28

Figure 5 15 Architecture of user intent and example generator................................ 28

Figure 5 16 Architecture of Rasa nlu updater 29

Figure 5 17 Architecture of knowledge update scheduler 29

Figure 6 1 Syntactic constraint pattern 32

Figure 6 2 Dataset Structure.. 37

Figure 6 3 Weight list.. 38

Figure 6 4 Loading dataset.. 38

Figure 6 5 Splitting data.. 38

Figure 6 6 Defining model 39

Figure 6 7 Training text data.. 39

Figure 6 8 Predict using a weight list... 39

Figure 6 9 Subject contain checker caller 39

Figure 6 10 Subject contain checker session creator 40

Figure 6 11 Subject contain checker query 40

Figure 6 12 Subject and keyword contain checker caller 40
Figure 6 13 Subject and keyword contain checker creator 40
Figure 6 14 Subject and keyword contain checker query 40
Figure 6 15 Check value none and node caller 41
Figure 6 16 Check value none and node creator 41
Figure 6 17 Check value none and node query 41
Figure 6 18 Keyword in subject node creator caller 41
Figure 6 19 Keyword in subject node creator 42
Figure 6 20 Keyword in subject node creator query 42
Figure 6 21 All nodes creator caller 42
Figure 6 22 All nodes creator 42
Figure 6 23 All nodes creator query 42
Figure 6 24 Wh word adder 43
Figure 6 25 Availability checker 43
Figure 6 26 Similarity checker with current intent 43
Figure 6 27 Threshold value exceed checker 44
Figure 6 28 Example writer 44
Figure 6 29 TP, FP, TN, FN values for different thresholds. 45
Figure 6 30 ROC Curve and threshold selection 46
Figure 6 31 Intent writer 47
Figure 6 32 Weight list 48

Figure 7 1 List of URL's 49
Figure 7 2 Keyword Text File. 49
Figure 7 3 Initial keyword on graph 49
Figure 7 4 Initial node 50
Figure 7 5 Dates 50
Figure 7 6 URL creation date 50
Figure 7 7 Audio downloading 50
Figure 7 8 Downloaded mp3 files 51
Figure 7 9 .mp3 to .wav 51
Figure 7 10 List of .wav files 51
Figure 7 11 List of chunks 52

Figure 7 12 Audio chunk text 52
Figure 7 13 List of text files.. 52
Figure 7 14 Audio converted text.. 53
Figure 7 15 Extracted triples 53
Figure 7 16 List of weights assigned... 53
Figure 7 17 Feature dataset 54
Figure 7 18 Logistic model training.. 54
Figure 7 19 Intents 55
Figure 7 20 nlu file with updated intents and examples if no semantic similarity with available intents.. 55
Figure 7 21 Similarity checker between available intents and created intent 55
Figure 7 22 nlu file with updated examples if semantic similarity with available intents 55
Figure 7 23 Knowledge graph... 56
Figure 7 24 Categorical data 56
Figure 7 25 Numerical data... 56
Figure 7 26 Training data and predicted value.. 57
Figure 7 27 Sample chat.. 57
Figure 7 28 Logistic regression matrices 57
Figure 7 29 Naive bayes matrices 58

List of tables

Table 2 1 Summary of the benefits and issues of the previous work......................... 11

Table 6 1 List of weights.. 33

List of abbreviation

Abbreviation	Description
FN	False Negative
FP	False Positive
FPR	False Positive Rate
NLP	Natural Language Processing
NLU	Natural Language Understanding
OIE	Open Information Extraction
RE	Relation Extraction
TN	True Negative
TP	True Positive
TPR	True Positive Rate
URL	Uniform Resource Locator