QUESTION CLASSIFICATION FOR THE TRAVEL DOMAIN USING DEEP CONTEXTUALIZED WORD EMBEDDING MODELS

Charmy Yasodhara Weerakoon

179356G

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

QUESTION CLASSIFICATION FOR THE TRAVEL DOMAIN USING DEEP CONTEXTUALIZED WORD EMBEDDING MODELS

Charmy Yasodhara Weerakoon

179356G

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgment 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 acknowledgment is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, 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 : 31/ 05/2021

Name: Charmy Yasodhara Weerakoon

The above candidate has carried out research for the Masters Dissertation under my supervision.

UOM Verified Signature

Signature of the supervisor: ……………………………… Date: 05/11/2021

Name of the supervisor: Dr. Surangika Ranathunga

Abstract

Keywords - Question Classification, Expected Answer Type, Ontology Learning, Transformers, RoBERTa

Question answering can be considered as a key area in Natural Language Processing and Information Retrieval, where users construct queries in natural language and receive suitable answers in return. In the travel domain, most questions are "content questions", where the expected answer is not the equivalent of "yes" or "no", but rather factual information. Replying to a free-form factual question based on a large collection of text is challenging. Previous research has shown that the accuracy of question answering systems can be improved by adding a classification phase based on the expected answer type. This research focuses on implementing a multi-level, multi-class question classification system focusing on the travel domain. Existing research for the travel domain is conducted using language-specific features and traditional Machine Learning models. In contrast, this research employs transformer-based state-of-the-art deep contextualized word embedding models for question classification. The proposed method improves the coarse class Micro F1-score by 5.43% compared to baseline. Fine-grain Micro F1-Score has also improved by 3.8%. This research also presents an empirical analysis of the effectiveness of different transformer-based deep contextualized word embedding models for multi-level multi-class classification.

DEDICATION

I dedicate this master's dissertation to my family and friends, who simply did not want me to quit.

ACKNOWLEDGMENT

I would like to express my profound gratitude to Dr. Surangika Ranathunga, for her invaluable support as my supervisor by providing relevant knowledge, materials, advice, supervision, state-of-art technology ideas, document reviews, and useful suggestions throughout this research work.

Further, I would like to thank all my colleagues in the MSC 2017 batch, for their help in sharing knowledge and encouragement. I would especially like to thank Hasangi Kahaduwa, Pathum Liyana Arachchi, Dilshan Pathirana, Vishma Dias, and Upali Kohomban for their research findings which would act as a baseline for this research.

Finally, I am indebted to my family, for all the sacrifices they have done throughout my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Abbreviation	Description
BERT	Bidirectional Encoder Representations
CNN	Convolutional Neural Network
EAT	Expected Answer Type
LSTM	Long-Short Term Memory
NLP	Natural Language Processing
POS	Part of Speech
Q&A	Question and Answer
RNN	Recurrent Neural Network
ROBERTA	Robustly Optimized BERT Pretraining Approach
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TREC	Text Retrieval Conference
UIUC	University of Illinois Urbana Champaign