GENERIC INFORMATION EXTRACTION FRAMEWORK FOR DOCUMENT PROCESSING

Agampodi Kanishka Gayathri Silva

189395H

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

May 2021

DECLARATION

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).

Name of the Student

A.K.G. Silva

Signature of the Student

Date: 31/5/2021

Supervised By

Dr A.T.P. Silva

Signature of the Supervisor

Date: 31/5/2021

DEDICATION

I dedicate this thesis to my family and friends. I will always appreciate the help of friends for the things all they have done and their valuable thoughts. I dedicate this work and give many thanks to people at the University of Moratuwa for their help and especially for lecturer for guiding this research.

ACKNOWLEDGEMENT

First and foremost, I would like to extend the sincere gratitude to my supervisor Dr A.T.P. Silva, for the continuous support of the research, for the patience, motivation, enthusiasm, and immense knowledge. I would like to thank the University of Moratuwa for allowing carrying out this research and it continuing support during the research.

I would like to pay gratitude for all the academic and non-academic staff members of the University of Moratuwa and the batchmates for their generous support, comments and encouragement throughout the project. I am grateful to all expert and novice meditators who were involved in this research project.

ABSTRACT

Information extraction from documents has become great use of novel natural language processing areas. Most of the entity extraction methodologies are variant in a context such as medical area, financial area, also come even limited to the given language. Rather than tackling this problem in such manner, it is better to have one generic approach which is applicable for any of such document types to extract entity information regardless of language, context and structure. Also, the great barrier in such research is exploring the structure while keeping the hierarchical, semantic and heuristic features. Another problem identified is that usually, it requires a massive training corpus. Therefore, this research focus on mitigating such problems.

Throughout the research timeline, several approaches have been identifying towards building document information extractors focusing on different disciplines. Starting from optical character recognition of document images to data mining of large corpus of documents this research area has been contributed to the development of natural language processing, semantic analysis, information extraction and conceptual modelling. Although in separate ways those are trying to achieve the generic ability to process any kind of document which unfortunately not being achieved successfully due to the approach and technical limitations.

As per the approach within this research, it can process any kind of document in any domain by simply adhering the conceptual relations without being trying to extract component-wise and mapping into known structures. Just as a human being look at any unknown document and going through the relations and making best guesses on answering the queries, this system will also mimic the same behaviour. As per the output, it can either document Concept-Relation or some answer for the given query.

The experimental strategy has partaken with regards to several different datasets originated from SQUAD 2.0, DOCVQA dataset, SQUAD 2.0 dataset and Kaggle based datasets. Based on F1 evaluation metric it performs with overall 87.01 performance rate on SQUAD 2.0 dataset showcasing its capable of question-answering task with higher accuracy.

Upon diving into experimental design, starting from the dataset evaluation several experiments have been carried out. Datasets such as SQUAD 2.0 and DocVQA has been used to evaluate the overall performance over metrics such as F1 score, accuracy and ANLS providing scores 87.01,52.78 and 0.583 respectively. The F1 score, which is 87.01 showcase that the provided solution achieves the expected objectives in deriving a generic model fitting for any question-answering task based on documents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
GIEF	Generic Information Extraction Framework
DLS	Document Concept-Relation
NLP	Natural Language Processing
OCR	Optical Character Recognition