

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 its 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

Declaration	ii
Dedication	iii
Acknowledgement	iv
Abstract	v
Table of Contents	vi
List of Figures	xi
List of Tables	xiv
List of Abbreviations	xv
1. Introduction	1
1.1 Prolegomena.....	1
1.2 Aim and Objectives.....	2
1.2.1 Aim.....	2
1.2.2 Objectives.....	2
1.3 Background and Motivation.....	3
1.4 Problem in Brief.....	3
1.5 Generic Information Extraction Framework for Document Processing	3
1.6 Resource Requirements.....	4
1.7 Structure of the Thesis	4
1.8 Summary	5
2. State of the Art in Document Information Extraction	6
2.1 Introduction.....	6
2.2 Major Researches in Document Information Extraction	6
2.2.1. OCR Handwritten Documents	6

2.2.2 Derive Concept-Relation.....	7
2.2.3 Derive Semantic Structure	8
2.2.4 Document Entity Recognition.....	8
2.2.5 Question Answering from Documents.....	9
2.3 Limitation in Current Approaches	13
2.4 Future Trends	14
2.5 Problem in Brief.....	14
2.6 Summary	14
3. Theoretical Foundation for GIEF	15
3.1 Introduction.....	15
3.2 TensorFlow	15
3.2.1 Tensor Processing Unit (TPU) [19]	15
3.2.2 Features [20].....	16
3.2.3 Applications [21].....	18
3.2.4 TensorFlow 2.0 [22].....	18
3.1.5 Why TensorFlow for GIEF?	19
3.2 Transformers	20
3.3.1 Encoder	20
3.3.2 Decoder	26
3.3.3 Why Transformers for GIEF?	26
3.4 ELMo [25] [26] [18]	27
3.4.1 Architecture.....	27
3.4.2 Why ELMo for GIEF?	28
3.5 BERT [27].....	28
3.5.1 BERT Encoder	29

3.5.2 Training the Model.....	29
3.5.3 Inputs and Outputs	30
3.5.4 Applications	32
3.5.5 Why BERT for GIEF?	32
3.6 Summary	33
4. Concept Relation Extraction Approach for Document Processing.....	34
4.1 Introduction	34
4.2 Hypothesis.....	34
4.3 Input	34
4.4 Output.....	34
4.5 Process	35
4.6 Potential Users of the System	36
4.7 Features	36
4.8 Summary	36
5. Design of GIEF for Document Processing	37
5.1 Introduction	37
5.2 Top Level Design.....	37
5.3 Framework Components	38
5.3.1 Concept-Relation Generator.....	38
5.3.1.1 Layout Analyzer.....	39
5.3.1.2 Knowledge-Map Generator.....	41
5.3.1.3 Entity Extractor	43
5.4.2 Pre-Trained Model	44
5.4.3 Query Extractor.....	44
5.5 Summary	44

6. Implementation of GIEF for Document Processing	45
6.1 Introduction	45
6.2.1 Layout Analyzer	45
6.2.1.1 Datasets	45
6.2.1.2 OCR Preprocessing	48
6.2.1.3 Model Generation.....	49
6.2.2 Knowledge-Map Generator.....	50
6.2.2.1 Dataset.....	50
6.2.2.2 Preprocessing	51
6.2.2.3 Model Generation.....	52
6.2.3 Entity Extractor	55
6.2.3.1 Datasets	55
6.2.3.2 Preprocessing	57
6.2.3.3 Model Generation.....	58
6.2.4 Query Extractor Module	59
6.2.4.1 Datasets	59
6.2.4.2 Preprocessing	60
6.2.4.3 Model Generation.....	60
6.3 Summary	61
7. Evaluation	62
7.1 Introduction	62
7.2 Experimental Design.....	62
7.3 Evaluation Strategy	62
7.3.1 Benchmark Datasets.....	62
7.3.1.1 SQuAD [39]	63

7.3.1.2 Why SQuAD?	64
7.3.1.3 DocVQA [37].....	64
7.3.1.4 Why DocVQA?	65
7.3.2 Benchmark Models	66
7.3.2.1 BERT base	66
7.3.2.2 BERT large.....	67
7.3.3 Metrics	67
7.3.2.2 Precision and Recall.....	68
7.3.2.3 F1 Score	69
7.3.2.4 Average Normalized Levenshtein Similarity (ANLS).....	69
7.4 Experimental Results	70
7.5 Discussion on Results	72
7.6 Summary	73
8. Conclusion and Future Work	74
8.1 Introduction.....	74
8.2 Concluding Remarks.....	74
8.3 Limitations and Future Works	75
8.4 Summary	75
References	76

LIST OF FIGURES

Figure 2.1: Overview of the Text Recognition System [4]	6
Figure 2.2: System Diagram of Descriptors [7]	7
Figure 2.3: Multimodal Fully Convolutional Neural Network [9]	8
Figure 2.4: Process of WAD Methodology [15]	10
Figure 3.1: TensorFlow Visualizer with TensorBoard [20]	16
Figure 3.2: TensorFlow Feature Columns [20]	17
Figure 3.3: TensorFlow: Parallel Neural Network Training [20]	17
Figure 3.4: Massive Multitask [21]	18
Figure 3.5: TensorFlow 2.0 Overview [22]	19
Figure 3.6: Illustrated Transformer [23]	20
Figure 3.7: Encoder Architecture [23]	21
Figure 3.8: Training parameters of self-attention layer [23]	22
Figure 3.9: Dot product attention	22
Figure 3.10: Self-Attention Steps Summary [23]	23
Figure 3.11: Self-attention calculation in matrix form [23]	23
Figure 3.12: Multi Attention Heads [23]	24
Figure 3.13: Self Attention steps [23]	24
Figure 3.14: Position Encoding Example [24]	25
Figure 3.15: Encoder-Decoder Architecture [23]	26
Figure 3.16: Model Inputs in BERT [28]	28
Figure 3.17: BERT Encoder [28]	29
Figure 3.18: Masked Language Model (MLM) [27]	30
Figure 3.19: BERT [29]	30

Figure 3.20: Spam detention with BERT [28]	31
Figure 3.21: BERT Usability [28].....	32
Figure 5.1: Top-level Design of GIEF for Document Processing	38
Figure 6.1: Receipts Dataset Structure.....	46
Figure 6.2: Receipts dataset statistics	46
Figure 6.3: Word grouping and semantic entity labelling [33].....	47
Figure 6. 4 Example image from FUNSD dataset	47
Figure 6.5: Annotations in JSON for FUNSD example image.....	47
Figure 6.6: cheesecake-20191221_003.pdf from the dataset.....	48
Figure 6.7: Preprocessed Information from receipt	48
Figure 6.8: Training log of layout analyzer	49
Figure 6.9: SNLI Dataset usage statistics	50
Figure 6.10: SNLI column statistic	51
Figure 6.11: Dataset Sample	51
Figure 6.12: Segmentation of the example sentence.....	52
Figure 6.13: Entity Extraction result for the example.....	52
Figure 6.14: Extracted entity pairs sample.....	53
Figure 6.15: relation result for the example sentence	53
Figure 6.16: Relations frequency	54
Figure 6.17: Knowledge Graph Representation.....	54
Figure 6.18: med_train dataset statistics	55
Figure 6.19: entity-annotated-corpus dataset statistics	56
Figure 6.20: Number of tagged entities	56
Figure 6.21: NER dataset annotation	57
Figure 6.22: Example of a sentence	58

Figure 6.23: Labels for the example sentence.....	58
Figure 6.24: Entity Extractor training log	58
Figure 6.25: Example Document from DocVQA with Question-Answer.....	59
Figure 6.26: Question-Answer Format	60
Figure 6.27: Dataset Visualization.....	60
Figure 7.1: Question Answer Pair from SQuAD dataset [39]	63
Figure 7.2: Sample Question-Answer pairs from DocVQA dataset [37]	65
Figure 7.3: Confusion Matrix.....	67
Figure 7.4: ANLS Results on Validation and Test sets	71
Figure 7.5: Accuracy Results on Validation and Test Sets.....	71
Figure 7.6: F1 Score Graph on SQUAD Dataset	72

LIST OF TABLES

Table 7.1: BERT-Large Models Accuracies [41]	67
Table 7.2: ANLS and Accuracy Comparison on DocVQA dataset	70
Table 7.3: F1 Scores on SQUAD 2.0 Dataset	71

LIST OF ABBREVIATIONS

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