A Review on Document Information Extraction Approaches

Kanishka Silva[1] **Thushari Silva**[2]
[1,2]Faculty of Information Technology, University of Moratuwa, Moratuwa, Sri Lanka

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. It is better to have one generic approach applicable for any document type to extract entity information regardless of language, context, and structure. Also, another issue in such research is structural analysis 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 barriers. Several approaches have been identifying towards building document information extractors focusing on different disciplines. This research area involves natural language processing, semantic analysis, information extraction, and conceptual modelling. This paper presents a review of the information extraction mechanism to construct a generic framework for document extraction with aim of providing a solid base for upcoming research.

1 Introduction

Digitization of documents and moving forward for the paperless community has been in crucial discussion in the past few decades, contributed to many research opportunities advancing from image processing, Natural Language Processing, ontology, semantic analysis. Such research has the focus on one particular narrowed problem. But none or fewer are supporting to provide a generalized approach for information extraction from documents. Document processing is tremendous due to its diversity; even document generation could be in different ways, i.e., handwritten, scanned-handwritten, typewritten, machine-generated, machine-printed. In terms of content, again it is possible to divide among purpose (business, legal, technical, medical, academic, historical) (Ritter et al., 2011) (An and Park, 2018), language (English, Hindi, Arabic, Multi-lingual) (Tao et al., 2014).

It is better to have a more generalized solution than limited to the scope of each application area. However, this would not be an easy task due to diversities in document nature, but worth exercising since it provides a novel way of addressing document processing issues. This paper focuses on querying documents tasks, such as in question answering, due to the expandability for several other applications. But, notably, the information extraction area itself is a vast research area employing different methodologies not only for targeting documents. Such examples would be (Corro and Gemulla, 2013) and (Yates et al., 2007) discussing legacy open information extraction approaches. Later, the works of (Zhu et al., 2019), (Jia et al., 2019), (Zhang et al., 2019a) and (Zhang et al., 2019b) explains more advanced approaches in open information extraction, but this is not the scope in this paper as this emphasis on document information extraction aspects.

In (Weikan et al., 2018) explain a reading strategy of the humanized approach of document extraction, where first skim through the test to get a general idea, read the question or query carefully with the equipped general understanding from the document and then search for the answer with the prior knowledge.

This paper is structured to guide through a step-by-step approach of document information extraction while checking on crucial research at a legacy level and recent advancements. These

Proceedings of the Student Research Workshop associated with RANLP-2021, pages 174–179,
held online, Sep 1–3, 2021.
https://doi.org/10.26615/issn.2603-2821.2021_024

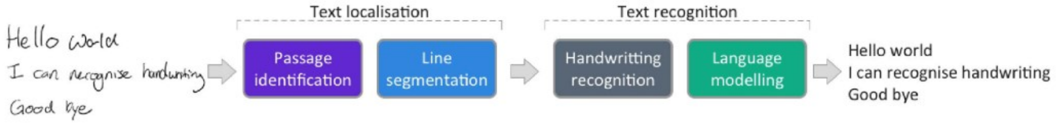

Figure 2: Overview of the Text Recognition System (Chung and Delteil, 2019)

groups could be used as segments in any proposed generic framework for document processing.

2 OCR Handwritten Documents

The paper (Saba et al., 2015) discusses language-independent rule-based classification for handwritten text and the printed text for data entry forms, mainly used in extracting information from form documents. It uses structural and statistical features derived into a rule-based system to determine the origin of the text. It considers baseline characteristics and stroke characteristics to differentiate.

(Chung and Delteil, 2019) propose a methodology for full-page offline text recognition using deep neural networks incorporating object detection neural networks, multi-scale CNN and bidirectional LSTM to perform text localization and then text recognition. In-text localization performs passage identification and line segmentation using IAM datasets and ResNet34 on ImageNet. Then during the text recognition phase, CNN-biLSTM will output NXM in terms of probabilities, and with the language modelling, it has converted to a string of text using the beam search approach. Figure 2 illustrates the overview of the system they proposed.

3 Derive Concept Relations

A Segmenting stream of documents using structural and factual descriptions discuss in (Karpinski and Bela¨ıd, 2016). It explains 4 document levels: records, technical documents, cases, and fundamental documents. This research proposes a novel method for stream segmentation in stream data. In the paper, a logbook for the previous pages they have used to compare descriptors. The system diagram illustrates as follows in Figure 1.

Logical labelling has been discussed in the study (Tao et al., 2014), using 2D Conditional Random Fields to derive the page structure in machine-printed documents. It constructs neighbourhood graphs and pairwise cliques templated while combining local and contextual features obtained from the PDF document. Using an SVM classifier, it derives the baseline without any contextual information considered. Early knowledge-based context analysis considers primary context, the relation between text and text, non-text, and text.

One of the OCR applications is digitizing historical books in libraries where the searchable text is derived. Although the existing OCR applications extract simple structures such as

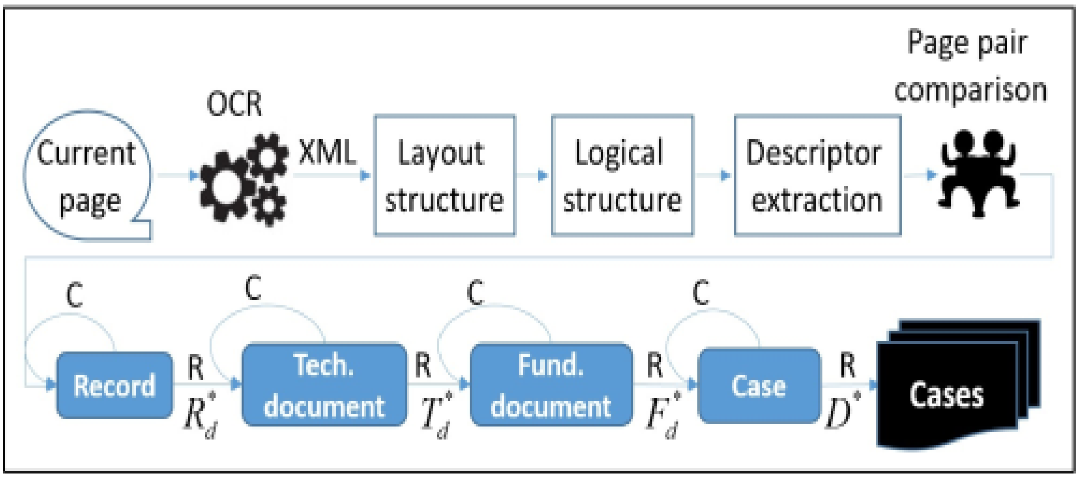

Figure 1: System Diagram of Descriptors (Karpinski and Bela¨ıd, 2016)

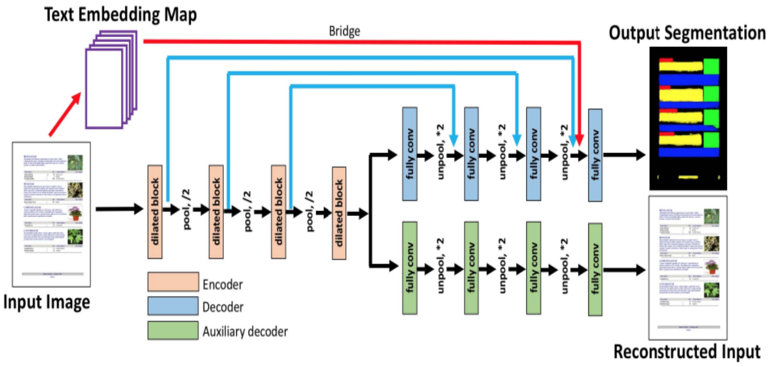

Figure 3: Multimodal Fully Convolutional Neural Network (Yang, et al., 2017)

paragraphs and pages, it is harder to identify complex formatting such as chapters, sections. Mapping such information to a table of content to provide more structural information is proposed by (Nguyen et al., 2017). It introduced an aggregation-based method on two set operators and properties of the table of content entries.

4 Derive Semantic Structure

Document Semantic Structure Extraction is a separate research discipline where document image is segmented into the region of interest and provide a semantic explanation for each (Yang et al., 2017). This paper involves two phases, page segmentation and Concept-Relation analysis. (Yang et al., 2017) present a multimodal fully convolutional network to extract semantic structure in terms of pixel-wise segmentation task.

The paper (Elhadad et al., 2017) proposes an approach to dimensionality reduction of feature vector size for web text document classification using WordNet ontology. This paper has proposed a Vector Space Model for the mining task and used Term Frequency Inverse Document Frequency for the task weighting. The suggested approach is ontology base, and for the evaluation, it uses Principal Component Analysis.

5 Document Entity Recognition

Differential tools use for document entity recognition evaluate in (Marrero et al., 2009). In the entity recognition process, it analyses the features of lexical, semantic, and heuristic. English corpus included around 579 words. Based on entity types of results are evaluated. Finally, this paper proposes a model that eliminates the limitations identified. Also, the corresponding model. is based on the small corpus. The best tools are Supersense - WNSS and the lowest performance from Afner. In (Ritter et al., 2011) uses an NLP pipeline that begins with part of speech tagging and then use chunking to obtain NER. It uses shallow syntax in tweets using annotated tweets and build-tools trained on data. Also, have used features generated from T-POS & T-Chunk in segmenting named entity. The paper (Marrero et al., 2013) discusses the meaning behind Named Entity Recognition and analyses the field from theoretical and practical views considering drawbacks, challenges, and opportunities in the area. It brought the argument to show that this task mainly depends on the development and evaluation of the tools.

NER can apply in many practical scenarios, such as in Twitter post analysis as explained in (Locke and Martin, 2009). It identifies several Named Entities and classifies them among People, Locations and Organizations. Also, it discusses identifying the language-specific features and methods effectively. Further, this paper elaborates on how it can improve to the areas of NER performing on other text sources such as news articles and less structured microblogging texts. Throughout the process of research, it identifies the difference between the microblogging text and the formal process.

The paper (Toral and Munoz, 2006) proposes a framework for named entity detection from Web content associated with semi-structured text data records by exploiting the inherent structure via a transformation process facilitating collective detection. This framework does not require training labels on the data records to learn the sequential classification model. Instead, it makes use of existing named entity repositories such as

DBpedia. It incorporates this external clue via distant supervision by making use of the Generalized Expectation constraint. Finally, a collective detection model supporting the logical inference proposed to consider the consistency among potential named entities and header text.

6 State-of-the-Art Language Models

Several language models have been emerged in more recent research, word embeddings, transformer-based models being key highlights. These model architectures and pre-trained models are used for document-based language modelling successfully.

The paper (Vasvani et al.,2017) presents an attention-based novel architecture named. Transformers. It mainly uses in natural language processing for tasks such as text summarization and translations. The main advantage in transformers is it does not require sequential data to insert in an orderly manner. Because of that, it allows parallelization compared to Recurrent Neural Networks and reduces the overall training time due to that. Pretrained models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) are using this model. Transformers follows encoder-decoder architecture where the encoder contains several encoding layers, and the decoder consists of several decoder layers. It was designed based on how long sentences memorize in neural machine translation. The input sentence is first embedded and then passed into a stack of encoders. Then the output from the last encoder pass to each decoder layer. In terms of the internal architecture, there are two main parts in the encoder, namely, self-attention and feed-forward neural network. And the decoder follows similar architecture to the encoder. There is an Encoder-Decoder Attention in between these two, which follows the same pattern as multiheaded self-attention. It takes sequential input for the decoder. The output will pass to another linear layer with SoftMax activations following the same concepts as forwarding propagation and loss functions such as cross-entropy to update trainable parameters.

ELMo (Peters et al., 2018) provides deep contextualized word representation using a bi-directional LSTM model to get word representations. It analyses the words by considering the context used. As it uses character-based representation, it creates dynamic vectors from the given input text. In ELMo, it learns both word and linguistic context. The pre-trained model can utilize for any downstream required NLP task, such as question answering, coreference resolution, textual entailment, named entity extraction, semantic role labelling and sentiment analysis. Multiple word embeddings produce for every single word considering different scenarios. Higher-level layers use to capture context-dependent aspects. And the model aspects such as syntax are captured by lower-level layers. ELMo (Peters et al., 2018) performs Language Modelling by training to predict the next word for the given word sequence. This contextual embedding is by grouping hidden states by concatenation.

The encoder part in BERT (Devlin et al., 2019) is similar to the architecture in transformers as it consists of a feed-forward neural network and self-attention heads. These two segments act as the training parameters to differentiate within different BERT (Devlin et al., 2018) models. The various applications of BERT are question-answer modelling, classification tasks such as spam detection, answer generation from a given passage, sentence tagging such as entity recognition.

7 Question Answering from Documents

Since any generic document information extraction framework should be able to provide answers for the given queries, we must consider the design aspects of question answering systems, where it requires the composing of several components, Named Entity Recognition, Semantic Analysis of the question, Named Entity Disambiguation, query expansion and execution. The study (Jayalakshmi and Sheshasaayee, 2017) has proposed, Web And semantic knowledge Driven (WAD) approach to increase the accuracy of the document selection. This approach first evaluates the user query, entity linking method and the query expansion technique. Finally, ontological information uses to rank the answers.

The paper (Veena et al., 2017) suggests a graph-based Question Answering system for reading comprehension tests, where it picks the sentence in the given passage which gives the best answer. It uses combined techniques of morphological analysis, coreference resolution and synonym checks to achieve higher accuracy.

8 Conclusion

Most of the research is firmly dependent upon the dataset/corpus and limited to the addressed language. The suggested model is usually trained and implemented within the given environment, whereas less performance for complex unstructured documents. For some research, a precompiled training dataset requires building the model so that when applying within a different environment, an additional effort need for preparing such datasets. Also, the current state of art referring to this study is usually subjective for the problem area such as academic, business, biomedical.

As per future trends regarding this research area, most recent studies are moving towards extracting information by considering the conceptual factors rather than digitizing the text and then grabbing the semantics. Current approaches are limited to given scope, document type, structure, language. Applying this research in real-time applications will not provide aid for day-to-day applications. In the existing approaches, a large corpus of documents uses for training and fitting the model. A requirement for a generalized framework capable of extracting information from any document exists due to that. The key idea is to mimic how a human understands an unknown text document through the content and the relation between the sections. Such a generalized framework will employ the benefits from existing approaches, but not solely dependent, as the main framework components could be replaceable by any other future approaches by providing a baseline. This review will provide a solid background analysis before approaching this generic framework.

References

Alan Ritter, Sam Clark, Mausam and Oren Etzioni. (2011). Named Entity Recognition in Tweets: An Experimental Study. EMNLP. (pp. 1524–1534) https://aclanthology.org/D11-1141.

Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. 2007. TextRunner: open information extraction on the web. In Proceedings of *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-Demonstrations '07).* Association for Computational Linguistics, USA, 25–26. https://doi.org/10.3115/1614164.1614177.

Antonio Toral and Rafael Munoz (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. *Workshop On New Text Wikis And Blogs And Other Dynamic Text Sources*. https://aclanthology.org/W06-2809

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the *31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Brian Locke, and James Martin (2009). Named entity recognition: adapting to microblogging. *In Computer Science Undergraduate Contributions* (p. 29).

Dongxu Zhang, Subhabrata Mukherjee, Colin Lockard, Xin Dong and A. McCallum (2019). Integrating Web-Scale OpenIE Extractions and Knowledge Bases with Entity Neighborhood Encoder.

Dongxu Zhang, Subhabrata Mukherjee, Colin Lockard, Xin Dong and A. McCallum (2019). OpenKI: Integrating Open Information Extraction and Knowledge Bases with Relation Inference. NAACL. https://doi.org/10.18653/V1/N19-1083

G. Veena, S. Athulya, S. Shaji and D. Gupta, "A graph-based relation extraction method for question answering system," 2017 *International Conference on Advances in Computing, Communications and Informatics (ICACCI),* 2017, pp. 944-949, https://doi.org/ 10.1109/ICACCI.2017.8125963.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" In Proceedings of the *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), https://doi.org/10.18653/v1%2FN19-1423

Jonathan Chung and Thomas Delteil, "A Computationally Efficient Pipeline Approach to Full Page Offline Handwritten Text Recognition," *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, pp. 35-40, https://doi.org/10.1109/ICDARW.2019.40078.

JungHyen An and Y. B. Park (2018). Methodology for Automatic Ontology Generation Using Database Schema Information. *Mobile Information Systems*. 2018. 1-13. https:/doi.org/10.1155/2018/1359174.

Luciano Del Corro and Rainer Gemulla. 2013. ClausIE: clause-based open information extraction. In Proceedings of the *22nd international conference on World Wide Web (WWW '13)*. Association for Computing Machinery, New York, NY, USA, 355–366. https://doi.org/10.1145/2488388.2488420

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer, (2018). Deep Contextualized Word Representations,*NAAC*,.https://doi.org/10.18653/v1/N18-1202

Mohamed K. Elhadad, Khaled M. Badran and Gouda I. Salama, "A novel approach for ontology-based dimensionality reduction for web text document classification," *IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS),* 2017, pp. 373-378, https://doi.org/10.1109/ICIS.2017.7960021.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, Juan Miguel Gómez-Berbís, Named Entity Recognition: Fallacies, challenges and opportunities, *Computer Standards & Interfaces*,Volume 35, Issue 5, 2013, Pages 482-489, ISSN 0920-5489, https://doi.org/10.1016/j.csi.2012.09.004.

Mónica Marrero, Sonia Sanchez-Cuadrado, Jorge Morato and George Andreadakis (2009). Evaluation of named entity extraction systems. In *Advances in Computational Linguistics Research in Computing Science* (pp. 47-58).

Romain Karpinski and Abdel Belaïd, "Combination of Structural and Factual Descriptors for Document Stream Segmentation," 2016 *12th IAPR Workshop on Document Analysis Systems (DAS),* 2016, pp. 221-226, https://doi.org/10.1109/DAS.2016.21.

S. Jayalakshmi and Ananthi Sheshasaayee, "Automated question answering system using ontology and semantic role," 2017 *International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2017, pp. 528-532, https://doi.org/10.1109/ICIMIA.2017.7975515.

Shengbin Jia, E. Shijia and Yang Xiang (2019). Neural Open Relation Extraction via an Overlap-aware Sequence Tagging Scheme.

Tanzila Saba, Abdulaziz S. Almazyad and Amjad Rehman "Language independent rule based classification of printed & handwritten text," 2015 *IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2015, pp. 1-4, https://doi.org/10.1109/EAIS.2015.7368806.

Thi-Tuyet-Hai Nguyen, Antoine Doucet and Mickael Coustaty, "Enhancing Table of Contents Extraction by System Aggregation," 2017 *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 242-247, https://doi.org/10.1109/ICDAR.2017.48.

Weikang Li, Wei Li, Yunfang Wu. A Unified Model for Document-Based Question Answering Based on Human-Like Reading Strategy. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, February 2-7, 2018. pages 604-611, AAAI Press, 2018.

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley and Daniel Kifer "Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks," 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2017, pp. 4342-4351, https://doi.org/10.1109/CVPR.2017.462.

Xin Tao, Zhi Tang, Canhui Xu and Yongtao Wang, "Logical Labeling of Fixed Layout PDF Documents Using Multiple Contexts," 2014 *11th IAPR International Workshop on Document Analysis Systems, 2014*, pp. 360-364, https://doi.org/10.1109/DAS.2014.54.

Zhu, Q., Ren, X., Shang, J., Zhang, Y., El-Kishky, A., Xu, F.F., & Han, J. (2019). Integrating Local Context and Global Cohesiveness for Open Information Extraction. Proceedings of the *Twelfth ACM International Conference on Web Search and Data Mining.*