SILVA, K., SILVA, T. and NANAYAKKARA, G. 2023. MicroConceptBERT: concept-relation based document information extraction framework. In *Proceedings of the 7th SLAAI (Sri Lanka Association for Artificial Intelligence) International conference on artificial intelligence 2023 (SLAAI-ICAI 2023), 23-24 November 2023, Kelaniya, Sri Lanka*. Piscataway: IEEE [online], article number 10365022. Available from: https://doi.org/10.1109...ICAI59257.2023.10365022

MicroConceptBERT: concept-relation based document information extraction framework.

SILVA, K., SILVA, T. and NANAYAKKARA, G.

2023

microConceptBERT: Concept-Relation based Document Information Extraction Framework

Kanishka Silva
School of Mathematics and Computer Science
University of Wolverhampton
United Kingdom
a.k.silva@wlv.ac.uk

Thushari Silva
Faculty of Information Technology
University of Moratuwa
Sri Lanka
thusharip@uom.lk

Gayani Nanayakkara
School of Computing
Robert Gordon University
United Kingdom
g.nanayakkara@rgu.ac.uk

***Abstract*—Extracting information from documents is a crucial task in natural language processing research. Existing information extraction methodologies often focus on specific domains, such as medicine, education or finance, and are limited by language constraints. However, more comprehensive approaches that transcend document types, languages, contexts, and structures would significantly advance the field proposed in recent research. This study addresses this challenge by introducing microConceptBERT: a concept-relations-based framework for document information extraction, which offers flexibility for various document processing tasks while accounting for hierarchical, semantic, and heuristic features. The proposed framework has been applied to a question-answering task on benchmark datasets: SQUAD 2.0 and DOCVQA. Notably, the F1 evaluation metric attains an outperforming 87.01 performance rate on the SQUAD 2.0 dataset compared to baseline models: BERT-base and BERT-large models.**

Keywords—Concept-Relations, Entity Extraction, Layout Analysis, Ontology, Transformers, Question Answering

I. INTRODUCTION

Information Extraction (IE) deals with extracting information from a bulk of data and revolves around extracting structured information from unstructured or semi-structured textual data. Document information extraction methods encompass a range of tasks, including Named Entity Recognition (NER), Relation Extraction (RE), normalization, and coreference resolution [1]. In recent past, IE methodologies have evolved from rule-based models [2] and 2D image-based document representations [3], [4] to more sophisticated approaches utilizing neural networks such as Recurrent Neural Networks (RNNs) [3]-[5] and transformer models [6], [7].

The diversity in document types has encouraged the development of generic document processing approaches in specific domains, such as financial [8], clinical [9], and legal [10] domains. Additionally, pipeline-based methods for entity-based information extraction have been proposed [11], [12], and end-to-end joint modelling approaches for Neural Relation Extraction have been introduced [13].

Inspired by the human reading process, which involves pre-reading, careful reading, and post-reading stages [14], this study proposes a novel concept-relation-based framework for document information extraction. According to [15], [16], pre-reading involves generating a general cognition of the document content, careful reading to locate detailed information according to a specific purpose and post-reading to complete the comprehension of the document.

In a practical medical research scenario, the need arises to extract information from a complex research paper to answer questions about the effectiveness of a new drug, requiring the extraction of various concepts during the question-answering process. We introduce a novel document information extraction framework that utilises concept-relation mapping to address such challenges to facilitate other domain areas, not limited to the medical domain. To the best of our knowledge, this is the first research integrating a concept-relation mapping-based information extraction framework into an end-to-end process using microConceptBERT models. The primary research questions addressed in this study are as follows:

1) RQ1: How can layout, named entities, and ontology be derived as concept-relation entities from a given document?

2) RQ2: How can extraction tasks be effectively performed using concept-relation mappings?

3) RQ3: What is the comparative performance of the proposed model against baseline models?

The structure of this paper is as follows: Section II provides an overview of the literature study, Section III describes the proposed methodology, Section IV outlines the experiment design for addressing the research questions, Section V presents the results, and finally, Section VI offers concluding remarks and future directions.

II. RELATED WORKS

The initial research in document processing centred on Optical Character Recognition (OCR) for handwritten documents such as data entry forms. Various approaches to OCR-based document analysis include probabilistic models [17], [18], rule-based methods [19], and more recent deep neural networks such as multi-scale Convolutional Neural Networks (CNN) [20] and pre-trained language models [21]. Popular OCR tools for end-to-end text processing from visually rich documents include LayoutLMv2 [22], LAMBERT [7], TILT [23], and DocParser [24].

Concept extraction involves identifying relevant concepts from documents by analysing concept-relations, layout, semantic structures, and document entities [25]. Rule-based concept extraction methods have been developed for clinical documents in MedLEE [26], MetaMap [27], cTAKES [28] and MedTagger [29]. The works discussed in [30] employ a logbook method based on structural and factual descriptors, with page structure extraction, utilizing 2D Conditional Random Fields to extract labelled logical structures. Interactive concept extraction methodologies have also been explored in [31]. Additionally, Casualty Extraction (CE) [32], [33] focuses on deriving cause–effect relations from text, representing a specialized form of concept extraction.

Semantic structure extraction is segmenting document regions of interest and providing semantic explanations for each segment, which involves two phases: page segmentation and concept-relation analysis. Various approaches such as rule-based [34], semantic parsing-based [35], and deep learning-based [36], [37] models have been utilized for

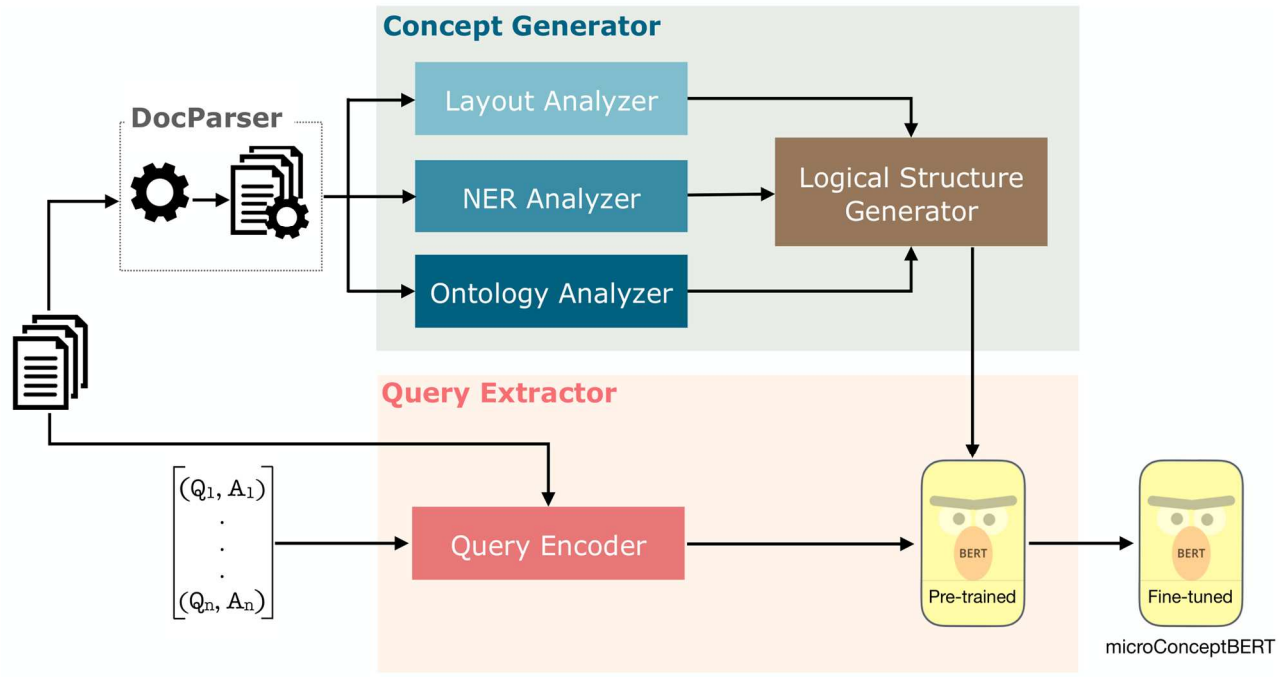

Fig. 1. High-Level Design of the Proposed microConceptBERT Information Extraction Framework

semantic structure extraction. Ontological methods, particularly with vector space-based models, have also proven valuable in deriving semantics, as discussed in [38].

Named Entity Recognition (NER) plays a pivotal role in extracting entity information like people, locations, and organizations from text [39]. Prominent machine learning methodologies, including Support Vector Machines (SVM), Conditional Random Fields (CRF), Maximum Entropy Markov Models (MEMM), Hidden Markov Models (HMM), and Decision Tree Classifier (DTC) [40] have been employed for NER. Notable techniques include T-POS and T-Chunk-based segmentation [41], as well as semi-structured logical inferences [42], for applications spanning from tweets [39], [41] to web content [42], [43]. Relation Extraction (RE), the identification of relationships in lengthy text, has been addressed using techniques like Multi-Instance Learning (MIL) [44], [45], graph-based neural approaches [46], [47], and Localized Context Pooling [48].

Document processing can be projected as a question-answering (QA) problem involving components like NER, semantic analysis [49], query expansion [49] and execution. A range of techniques have been employed in QA systems, including graph-based [50], embedding-based [51], [52], multi-column convolutional neural networks (MCCNNs) [53], attention mechanism [54], [55].

Despite many successful document information extraction models, spanning from task-specific models to end-to-end pipelines, a unified framework that integrates different components for concept-relation-based information extraction is required in this area of research. Hence, this study addresses this gap by proposing a comprehensive concept-relation-based framework for document information extraction.

III. METHODOLOGY

The proposed framework comprises two main components: the Concept Generator and Query Extractor, as depicted in Fig. 1. Each of these components comprises replaceable sub-modules. The concept-relation extraction model was implemented as microConceptBERT models inspired by the BERT [56] and ELMO [57] models.

A. Concept-Relation Generator

Within the Concept-Relation Generator, three sub-modules operate in parallel: the Layout Analyzer, Ontology Generator, and NER Analyzer. The Generator concatenates their outputs to establish a comprehensive Concept-Relation mapping, as illustrated in Fig. 2.

1) Layout Analyzer

The Layout Analyzer extracts structure-related insights, encompassing word embeddings, layout embeddings, and positional embeddings. Inspired by LayoutLM [58], microConceptBERT layout models have been used to utilise these different embedding types. Positional embeddings capture inter-word relationships, while layout embeddings encapsulate word positional and layout distribution, incorporating hierarchical structure.

2) NER Analyzer

The NER Analyzer component employs the microConceptBERT Named Entity model to derive entity concepts from the given document. Multi-level Named Entity Recognition (NER) embeddings are employed to address contextual ambiguities. The level 1 NER model encodes the word sequences with high-level named entity parses such as ORG, PER, and NUM, while level 2 NER specifies the NER distribution span. Within the tier-3 NER embeddings layer, entity value types are considered. These NER embeddings, encompassing three levels, are combined and normalized within microConceptBERT NER, generating an entity concept-relation map of the document. This map is utilized for query validation and answer localization.

3) Ontology Analyzer

The Knowledge Map encapsulates contextual information hierarchically through different ontologies. Commencing from the high-level definition of the document, including

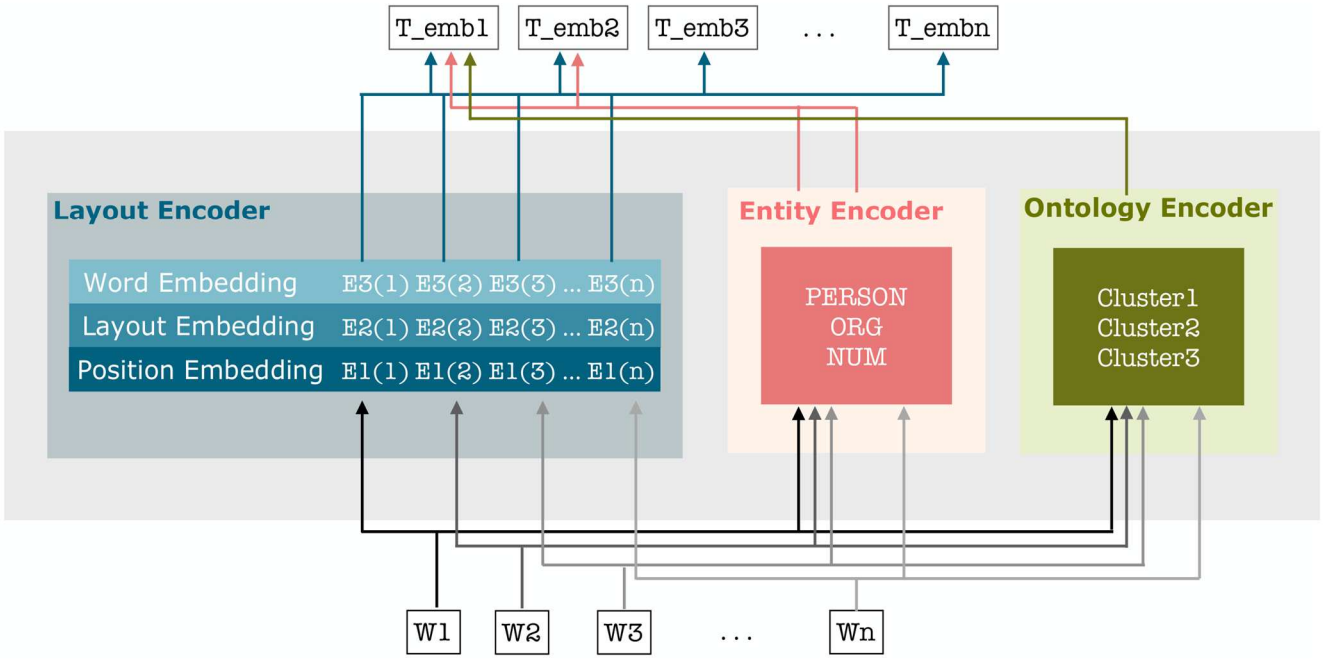

Fig. 2. Concept-Relation Generator

topics, titles, sub-titles and corresponding regions, a level-1 Knowledge Map is formulated to establish each word with its relatedness to each entity. In the tier-2 Knowledge Map, complex tabular structural data is extracted and encoded within embeddings. Further depth is reached with the tier-3 Knowledge Map, which represents form structures in a binary key-value pair format.

B. *Query Extractor*

The query extractor model demonstrates how the pre-trained model can be fine-tuned for the question-answering task, serving as a fundamental information extraction approach. In this context, the model is semi-supervisedly trained on the DocVQA dataset [59].

IV. Experiments

The experimental evaluation of the proposed framework focuses on question-answering tasks, with a comparative analysis against baseline models, namely BERT-base and BERT-large models on the datasets SQUAD 2.0 [60] and DocVQA [59]. The evaluation results on the DocVQA dataset are measured using the Average Normalized Levenshtein Similarity (ANLS) (Eq. 1) and accuracy metrics. ANLS incorporates a slight penalty to account for Optical Character Recognition (OCR) errors. A threshold of 0.5 is applied to determine whether an incorrect or correct answer is provided. The calculation of the ANLS metric is outlined in Eq. 2, where the model's output (o_{qi}) is compared to the ground truth answer (a_{ij}) for each question (i) and its respective ground-truth answers (j), with N representing the total number of questions and M indicating the total number of ground-truth answers for each question. The experiment results on the SQUAD 2.0 dataset were reported using F1 scores.

$$\text{ANLS} = \frac{1}{N}\sum_{i=0}^{N}\left(\max_{j} s(a_{ij}, o_{qi})\right) \quad (1)$$

$$s(a_{ij}, o_{q_i}) = \begin{cases} \left(1 - \text{NL}(a_{ij}, o_{q_i})\right) & \text{if } \text{NL}(a_{ij}, o_{q_i}) < \tau \\ 0 & \text{if } \text{NL}(a_{ij}, o_{q_i}) > \tau \end{cases} \quad (2)$$

V. Results and Discussion

The results of the evaluation on the DocVQA dataset, comparing the proposed framework with BERT-base and BERT-large models, are summarized in Table I. Average Normalized Levenshtein Similarity (ANLS) and accuracy are reported for both the validation and test sets. It is observed that the proposed framework achieves slightly higher accuracy and ANLS scores on the DocVQA dataset compared to the BERT-base model but slightly lower scores compared to the BERT-large model. This could be because the BERT-large model benefits from training and fine-tuning on a more extensive and diverse dataset. Although the results on the DocVQA dataset do not outperform the BERT-large model, it denotes the robustness of the model when dealing with real-world documents, as the ANLS metric accounts for Optical Character Recognition (OCR) errors. Altogether, these results highlight the model's applicability for practical scenarios.

Furthermore, Table II summarizes the accuracy comparison of the SQUAD 2.0 dataset among the BERT-base model, the BERT-large model, and the proposed framework. The proposed framework demonstrates outstanding performance, achieving an F1 score of 87.01, outperforming both the BERT-base and BERT-large models (83.01 and 86.10, respectively). This result highlights the effectiveness of the proposed framework in addressing question-answering tasks across diverse document complexities. Altogether, findings contribute to the relevance of the framework and applicability in real-world document information extraction scenarios.

In conclusion, the results indicate the significance of the 'microConceptBERT' framework in the document information extraction tasks, specifically its exceptional performance on the SQUAD 2.0 dataset to indicate its real-world applicability. The proposed framework offers a context-aware, concept-

based approach facilitating document understanding in textual document information extraction.

TABLE I. ANLS AND ACCURACY COMPARISON ON DOCVQA DATASET

Model	ANLS		Accuracy	
	Validation	*Test*	*Validation*	*Test*
BERT-base	0.56	0.57	45.60	47.60
BERT-large	0.59	0.61	49.28	51.08
microConceptBERT (Proposed)	0.56	0.58	49.31	52.78

TABLE II. F1 SCORES ON SQUAD 2.0 DATASET

Model	F1
BERT-base	83.01
BERT-large	86.10
microConceptBERT (Proposed)	87.01

VI. CONCLUSION

In this study, we have introduced a novel approach to derive concept-relations from documents by considering the textual content and incorporating layout, entity, and knowledge map concepts. The primary focus of this research was to demonstrate the proposed framework's application in the context of question-answering tasks. The evaluation was conducted using the SQUAD 2.0 and DocVQA datasets, and performance was assessed through metrics such as F1 score, accuracy, and Average Normalized Levenshtein Similarity (ANLS), providing scores of 87.01, 52.78 and 0.58, respectively. These results validate the effectiveness of the proposed solution in addressing the research questions stated in deriving a generic model fitting for any question-answering task based on documents. This framework significantly exhibits its potential for further technological enhancements in each constituent component. Additionally, it can be extended to other document-processing tasks, including summarization. This framework empowers a comprehensive and adaptable solution in the domain of document information extraction as a comprehensive and adaptable solution to integrate future methodologies and tools.

REFERENCES

[1] G. Simoes, H. Galhardas, and L. Coheur, "Information extraction tasks: a survey," 2009. [Online]. Available: https://www.inesc-id.pt/ficheiros/publicacoes/5519.pdf

[2] A. Simon, J. Pret, and A. P. Johnson, "A fast algorithm for bottom-up document layout analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 3, pp. 273–277, 1997. [Online]. Available: https://doi.org/10.1109/34.584106

[3] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Hohne, and J. B. Faddoul, "Chargrid: Towards understanding 2d documents," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 4459–4469. [Online]. Available: https://aclanthology.org/D18-1476/

[4] X. Zhao, Z. Wu, and X. Wang, "CUTIE: learning to understand documents with convolutional universal text information extractor," CoRR, vol. abs/1903.12363, 2019. [Online]. Available: http://arxiv.org/abs/1903.12363

[5] T. I. Denk and C. Reisswig, "Bertgrid: Contextualized embedding for 2d document representation and understanding," CoRR, vol. abs/1909.04948, 2019. [Online]. Available: http://arxiv.org/abs/1909.04948

[6] B. P. Majumder, N. Potti, S. Tata, J. B. Wendt, Q. Zhao, and M. Najork, "Representation learning for information extraction from form-like documents," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 6495–6504. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.580

[7] L. Garncarek, R. Powalski, T. Stanislawek, B. Topolski, P. Halama, M. Turski, and F. Gralinski, "LAMBERT: layout-aware language modeling for information extraction," in 16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I, ser. Lecture Notes in Computer Science, J. Llados, D. Lopresti, and S. Uchida, ´ Eds., vol. 12821. Springer, 2021, pp. 532–547. [Online]. Available: https://doi.org/10.1007/978-3-030-86549-8 34

[8] S. Zheng, W. Cao, W. Xu, and J. Bian, "Doc2edag: An end-to-end document-level framework for chinese financial event extraction," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 337–346. [Online]. Available: https://doi.org/10.18653/v1/D19-1032

[9] S. Zheng, J. J. Lu, N. Ghasemzadeh, S. S. Hayek, A. A. Quyyumi, F. Wang et al., "Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies," JMIR medical informatics, vol. 5, no. 2, p. e7235, 2017.

[10] K. Kowsrihawat and P. Vateekul, "An information extraction framework for legal documents: A case study of thai supreme court verdicts," in 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE, 2015, pp. 275–280.

[11] K. Shaalan, "A survey of arabic named entity recognition and classification," Comput. Linguistics, vol. 40, no. 2, pp. 469–510, 2014. [Online]. Available: https://doi.org/10.1162/COLI a 00178

[12] Y. S. Chan and D. Roth, "Exploiting syntactico-semantic structures for relation extraction," in The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. The Association for Computer Linguistics, 2011, pp. 551–560. [Online]. Available: https://aclanthology.org/P11-1056/

[13] M. Zhang, Y. Zhang, and G. Fu, "End-to-end neural relation extraction with global optimization," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, M. Palmer, R. Hwa, and S. Riedel, Eds. Association for Computational Linguistics, 2017, pp. 1730–1740. [Online]. Available: https://doi.org/10.18653/v1/d17-1182

[14] S. Cui, X. Cong, B. Yu, T. Liu, Y. Wang, and J. Shi, "Document-level event extraction via human-like reading process," in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. IEEE, 2022, pp. 6337–6341. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747721

[15] A. Saricoban, "Reading strategies of successful readers through the three phase approach," The Reading Matrix, vol. 2, no. 3, 2002.

[16] E. Toprak and G. ALMACIOGLU, "Three reading phases and their applications in the teaching of english as a foreign language in reading classes with young learners," Journal of language and Linguistic Studies, vol. 5, no. 1, 2009.

[17] F. Cesarini, E. Francesconi, M. Gori, and G. Soda, "Analysis and understanding of multi-class invoices," Int. J. Document Anal. Recognit., vol. 6, no. 2, pp. 102–114, 2003. [Online]. Available: https://doi.org/10.1007/s10032-002-0084-6

[18] M. Rusinol, T. Benkhelfallah, and V. P. D'Andecy, "Field extraction ˜ from administrative documents by incremental structural templates," in 12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013. IEEE Computer Society, 2013, pp. 1100–1104. [Online]. Available: https://doi.org/10.1109/ICDAR.2013.223

[19] T. Saba, A. S. Almazyad, and A. Rehman, "Language independent rule based classification of printed & handwritten text," in 2015 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2015, pp. 1–4.

[20] J. Chung and T. Delteil, “A computationally efficient pipeline approach to full page offline handwritten text recognition,” in 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 5. Los Alamitos, CA, USA: IEEE Computer Society, sep 2019, pp. 35–40. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICDARW.2019.40078

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” CoRR, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[22] M. Wei, Y. He, and Q. Zhang, “Robust layout-aware IE for visually rich documents with pre-trained language models,” in Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020, pp. 2367–2376. [Online]. Available: https://doi.org/10.1145/3397271.3401442

[23] R. Powalski, L. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Palka, “Going full-tilt boogie on document understanding with text-image-layout transformer,” in 16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II, ser. Lecture Notes in Computer Science, J. Llados, D. Lopresti, and S. Uchida, ´ Eds., vol. 12822. Springer, 2021, pp. 732–747. [Online]. Available: https://doi.org/10.1007/978-3-030-86331-9 47

[24] M. Dhouib, G. Bettaieb, and A. Shabou, “Docparser: End-to-end ocr-free information extraction from visually rich documents,” in Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San Jose, CA, USA, August 21-26, 2023, Proceedings, Part ´ V, ser. Lecture Notes in Computer Science, G. A. Fink, R. Jain, K. Kise, and R. Zanibbi, Eds., vol. 14191. Springer, 2023, pp. 155–172. [Online]. Available: https://doi.org/10.1007/978-3-031-41734-4 10

[25] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, and H. Liu, “Clinical concept extraction: A methodology review,” Journal of Biomedical Informatics, vol. 109, p. 103526, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420301544

[26] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, “A General Natural-language Text Processor for Clinical Radiology,” Journal of the American Medical Informatics Association, vol. 1, no. 2, pp. 161–174, 03 1994. [Online]. Available: https://doi.org/10.1136/jamia.1994.95236146

[27] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances,” Journal of the American Medical Informatics Association, vol. 17, no. 3, pp. 229–236, 05 2010. [Online]. Available: https://doi.org/10.1136/jamia.2009.002733

[28] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” Journal of the American Medical Informatics Association, vol. 17, no. 5, pp. 507–513, 09 2010. [Online]. Available: https://doi.org/10.1136/jamia.2009.001560

[29] H. Liu, S. J. Bielinski, S. Sohn, S. Murphy, K. B. Wagholikar, S. R. Jonnalagadda, K. E. Ravikumar, S. T. Wu, I. J. Kullo, and C. G. Chute, “An information extraction framework for cohort identification using electronic health records,” AMIA Jt Summits Transl Sci Proc, vol. 2013, pp. 149–153, Mar. 2013.

[30] R. Karpinski and A. Belaıd, “Combination of structural and factual descriptors for document stream segmentation,” in 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 2016, pp. 221–226.

[31] S. Zheng, J. J. Lu, N. Ghasemzadeh, S. S. Hayek, A. A. Quyyumi, and F. Wang, “Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies,” JMIR Med Inform, vol. 5, no. 2, p. e12, May 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28487265

[32] A. Balashankar, S. Chakraborty, S. Fraiberger, and L. Subramanian, “Identifying predictive causal factors from news streams,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 2338–2348. [Online]. Available: https://doi.org/10.18653/v1/D19-1238

[33] J. Qiu, L. Xu, J. Zhai, and L. Luo, “Extracting causal relations from emergency cases based on conditional random fields,” in KnowledgeBased and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference KES-2017, Marseille, France, 6-8 September 2017, ser. Procedia Computer Science, C. Zanni-Merk, C. S. Frydman, C. Toro, Y. Hicks, R. J. Howlett, and L. C. Jain, Eds., vol. 112. Elsevier, 2017, pp. 1623–1632. [Online]. Available: https://doi.org/10.1016/j.procs.2017.08.252

[34] L.-T. Wu, J.-R. Lin, S. Leng, J.-L. Li, and Z.-Z. Hu, “Rulebased information extraction for mechanical-electrical-plumbing-specific semantic web,” Automation in Construction, vol. 135, p. 104108, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580521005598

[35] A. Fader, L. Zettlemoyer, and O. Etzioni, “Open question answering over curated and extracted knowledge bases,” in The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, New York, NY, USA - August 24 - 27, 2014, S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, and R. Ghani, Eds. ACM, 2014, pp. 1156–1165. [Online]. Available: https://doi.org/10.1145/2623330.2623677

[36] F. Meng, S. Yang, J. Wang, L. Xia, and H. Liu, “Creating Knowledge Graph of Electric Power Equipment Faults Based on BERT-BiLSTMCRF Model,” Journal of Electrical Engineering & Technology, vol. 17, no. 4, pp. 2507–2516, Jul. 2022.

[37] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, and C. L. Giles, “Learning to extract semantic structure from documents using multimodal fully convolutional neural networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4342– 4351.

[38] M. K. Elhadad, K. Badran, and G. I. Salama, “A novel approach for ontology-based dimensionality reduction for web text document classification,” in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, pp. 373–378.

[39] B. Locke, “Named entity recognition : Adapting to microblogging,” in Computer Science Undergraduate Contributions, 2009.

[40] Z. Nasar, S. W. Jaffry, and M. K. Malik, “Named entity recognition and relation extraction: State-of-the-art,” ACM Comput. Surv., vol. 54, no. 1, feb 2021. [Online]. Available: https://doi.org/10.1145/3445965

[41] A. Ritter, S. Clark, Mausam, and O. Etzioni, “Named entity recognition in tweets: An experimental study,” in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 1524–1534. [Online]. Available: https://aclanthology.org/D11-1141

[42] A. Toral and R. Munoz, “A proposal to automatically build ˜ and maintain gazetteers for named entity recognition by using Wikipedia,” in Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources, 2006. [Online]. Available: https://aclanthology.org/W06-2809

[43] Q. Wang, Y. Fang, A. Ravula, F. Feng, X. Quan, and D. Liu, “Webformer: The web-page transformer for structure information extraction,” in WWW ’22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022, F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Medini, Eds. ACM, 2022, pp. 3124–3133. [Online]. Available: ´ https://doi.org/10.1145/3485447.3512032

[44] S. Riedel, L. Yao, and A. McCallum, “Modeling relations and their mentions without labeled text,” in Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), September 2010.

[45] P. Verga, E. Strubell, and A. McCallum, “Simultaneously self-attending to all mentions for full-abstract biological relation extraction,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 872–884. [Online]. Available: https://aclanthology.org/N18-1080

[46] P. Gupta, S. Rajaram, H. Schutze, and T. Runkler, “Neural relation ¨ extraction within and across sentence boundaries,” in Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, ser. AAAI’19/IAAI’19/EAAI’19. AAAI Press,

2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33016513

[47] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W.-t. Yih, "Crosssentence n-ary relation extraction with graph LSTMs," Transactions of the Association for Computational Linguistics, vol. 5, pp. 101–115, 2017. [Online]. Available: https://aclanthology.org/Q17-1008

[48] W. Zhou, K. Huang, T. Ma, and J. Huang, "Document-level relation extraction with adaptive thresholding and localized context pooling," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 16, pp. 14 612–14 620, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17717

[49] S. Jayalakshmi and A. Sheshasaayee, "Automated question answering system using ontology and semantic role," in 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017, pp. 528–532.

[50] G. Veena, S. Athulya, S. Shaji, and D. Gupta, "A graph-based relation extraction method for question answering system," in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 944–949.

[51] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 615–620. [Online]. Available: https://doi.org/10.3115/v1/d14-1067

[52] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I, ser. Lecture Notes in Computer Science, T. Calders, F. Esposito, E. Hullermeier, and R. Meo, Eds., vol. 8724. Springer, 2014, pp. 165– ¨ 180. [Online]. Available: https://doi.org/10.1007/978-3-662-44848-911

[53] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question answering over Freebase with multi-column convolutional neural networks," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 260–269. [Online]. Available: https://aclanthology.org/P15-1026

[54] Y. Hao, Y. Zhang, K. Liu, S. He, Z. Liu, H. Wu, and J. Zhao, "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, R. Barzilay and M. Kan, Eds. Association for Computational Linguistics, 2017, pp. 221–231. [Online]. Available: https://doi.org/10.18653/v1/P17-1021

[55] Y. Zhang, K. Liu, S. He, G. Ji, Z. Liu, H. Wu, and J. Zhao, "Question answering over knowledge base with neural attention combining global knowledge information," CoRR, vol. abs/1606.00979, 2016. [Online]. Available: http://arxiv.org/abs/1606.00979

[56] J. Devlin, M.-W. Chang, K. Lee, and K. 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). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[57] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[58] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1192–1200. [Online]. Available: https://doi.org/10.1145/3394486.3403172

[59] M. Mathew, D. Karatzas, and C. V. Jawahar, "Docvqa: A dataset for vqa on document images," in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 2199–2208.

[60] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264