DOI: 10.17148/IJARCCE.2024.13126

A Systematic Survey of Techniques for Document Processing and Natural Language Understanding

S R Suresh[1], Shraddha C [2], Sai kiran [3], Sharmila Chidaravalli [4]

Student, Department of Information Science, Global Academy of Technology, Bangalore, India[1,2,3]

Asst Prof, Department of Information Science, Global cademy of Technology, Bangalore, India[4]

Abstract: This survey paper explores various techniques and methodologies used in the field of document processing and natural language processing. The paper examines different research studies and their contributions in addressing specific issues related to document processing and language understanding. The techniques discussed include template matching, image processing, deep learning algorithms such as YOLOv5 and BERT, optical character recognition (OCR), convolutional neural networks (CNNs), named entity recognition (NER), and machine translation. The survey paper highlights the challenges faced in manual invoice processing and proposes an automatic system based on key fields extraction from invoices. It also addresses the complexities of handling diverse document layouts, including invoices, purchase orders, and newspaper articles, using template-based, rule-based, and OCR techniques. Handwritten text recognition in South Indian languages is explored, considering the cursive and complex structure of handwriting and the unavailability of temporal information. The paper also focuses on the need for annotated datasets and the application of AI approaches in processing unstructured invoice documents. It discusses the utilization of image segmentation, OCR, and NLP for summarizing newspaper articles and efficient processing of unstructured documents using AI techniques. Additionally, the challenges of OCR performance in low-quality images and intelligent handwritten recognition are examined. Furthermore, the paper explores the application of NLP techniques such as named entity recognition, coreference resolution, relation extraction, and knowledge base reasoning for information extraction. It discusses the challenges and applications of NER in finance and biomedicine. The survey also investigates the use of deep learning models like BERT and transformers for semantic keyphrase extraction and presents a comprehensive overview of Indian language speech synthesis techniques. Finally, the paper explores the challenges in text-to-speech training, machine translation, and Indian regional language processing. It discusses the limitations of parallel training data for voice conversion and the lack of linguistic grounding in autoencoder-based voice conversion methods. The survey paper provides a comprehensive overview of the techniques, challenges, and advancements in the field of document processing and language understanding, paving the way for future research and development.

Keywords: Natural Language Processing (NLP), Convolutional neural networks (CNNs), non-native speakers, Optical Character Recognition (OCR), key insights, inclusivity, marginalized communities.

I. INTRODUCTION

Document processing and natural language processing (NLP) play crucial roles in various domains, ranging from business operations to information retrieval and understanding. The ability to efficiently extract key information from documents, understand their content, and derive meaningful insights is of great importance in today's data-driven world. This survey paper aims to provide a comprehensive overview of the techniques, methodologies, and challenges related to document processing and NLP. The field of document processing encompasses a wide range of tasks, including invoice processing, summarization, handwritten text recognition, and unstructured document handling. Manual invoice processing is a labour-intensive and time-consuming task that can be prone to errors. Therefore, the paper explores techniques such as template matching, image processing, deep learning algorithms like YOLOv5, and optical character recognition (OCR) using software like Tesseract OCR to propose an automatic system for key fields extraction from invoices.

Furthermore, the complexity of document layouts poses challenges in real-time situations, such as invoices with diverse structures and purchase orders with varying formats. The paper discusses template-based and rule-based approaches, as well as OCR techniques, to address the complexities of handling different document layouts. It also highlights the significance of techniques like image segmentation, OCR, and NLP in summarizing newspaper articles and efficiently processing unstructured documents. The recognition of handwritten text, especially in South Indian languages, presents unique challenges due to the cursive nature and complex structure of handwriting.

DOI: 10.17148/IJARCCE.2024.13126

Additionally, the lack of temporal information in handwriting recognition further complicates the task. The paper explores the application of convolutional neural networks (CNNs) and classifier combination methods to address these challenges and improve the accuracy of handwritten text recognition. An important aspect of document processing is the availability of annotated and high-quality datasets. The paper discusses feature extraction techniques such as Glove, Word2Vec, FastText, as well as AI approaches like Bidirectional LSTM (BiLSTM) and BiLSTM-CRF in the context of processing unstructured invoice documents. It also highlights the need for domain-specific and task-specific datasets, as well as data validation approaches to ensure the quality and reliability of the data used in document processing tasks. The survey paper also delves into the realm of NLP and its various applications. Named entity recognition (NER), coreference resolution, relation extraction, and knowledge base reasoning are among the techniques discussed for information extraction. The paper explores the challenges of time-consuming processes and error propagation in NLP tasks and investigates the application of NER in domains such as finance and biomedicine.

Additionally, the paper examines the use of deep learning models like BERT and transformers for semantic keyphrase extraction, enabling the automatic extraction of meaningful keywords from large volumes of text. It also provides a comprehensive overview of speech synthesis techniques in Indian languages, considering the language-specific characteristics and the need for language-specific data. The survey paper further investigates the challenges in text-to-speech training, machine translation, and Indian regional language processing. It addresses issues such as limited availability of parallel training data for voice conversion and the lack of linguistic grounding in autoencoder-based voice conversion methods. These challenges hinder the development of accurate and natural-sounding speech synthesis systems. In conclusion, this survey paper aims to provide a comprehensive overview of the techniques, methodologies, and challenges in document processing and NLP. By exploring various research studies and their contributions, the paper sheds light on the advancements made in these fields and identifies areas for future research and development. The insights gained from this survey will contribute to the improvement and innovation of document processing and NLP techniques, ultimately enhancing our ability to extract valuable information and derive meaningful insights from diverse sources of textual data.

II. DOCUMENT PROCESSING TECHNIQUES

Document processing encompasses a wide spectrum of techniques to unlock valuable insights concealed within documents, document processing entails a multifaceted amalgamation of techniques spanning computer vision, machine learning, and natural language understanding domains to unlock valuable insights embedded within documents. With the exponential growth in digital documents across sectors, the importance of automated document processing continues to amplify.

A pivotal and ubiquitous technique includes template matching in conjunction with optical character recognition for key information extraction from invoices. Segmenting documents into semantic zones defined via rule-based or layout analysis algorithms enables extraction of relevant entities in a contextual manner. Specifically for invoice processing, we propose an end-to-end architecture encompassing YOLOv5 for object detection coupled with Tesseract OCR software for digitization and data capture. Another pervasive challenge warranting attention includes development of robust systems adept at handling heterogeneous documents with varying layouts, structures and formats. Our survey revealed template matching, graph-based document analysis, vision transformers and multimodal fusion approaches to be particularly promising directions. For instance, we analyze a system incorporating textual BERT and graph-convolutional encoders to classify document elements and types in an interdependent yet dynamically adaptive way. Such capabilities will facilitate seamless ingestion of diverse invoices, purchase orders, contracts into automated pipelines. Newspaper articles, which blend text, images and graphics, pose unique complexities for analysis requiring specialized multimodal techniques. We discuss systems leveraging the trinity of image segmentation, Tesseract OCR and CoreNLP for converting images into text and subsequently generating summary highlights using extractive summarization algorithms tuned on news datasets. For handwritten text, especially cursive scripts prevalent in South Asian languages, neural approaches can account for variations in writing styles. Our survey suggests convolutional-recurrent architectures trained in an end-to-end manner as fitting candidates for handwritten text recognition without extensive pre/post-processing steps while also retaining language-specific peculiarities. We further underscore the indispensable role of large, high-quality labeled datasets in supervised document analysis approaches along with robust data validation protocols for ensuring model generalizability across domains. To circumvent laborious manual annotation, promising semi-supervised approaches include style regularization via textual descriptions and layout consistency imposition rather than complete layout markup. Our survey findings suggest avenues for generating synthetic documents with controlled variation as a scalable alternative complementing available human-annotated corpora. In summary, our comprehensive analysis of existing literature surrounding document processing techniques has revealed promising directions but also hitherto unexplored aspects warranting future investigation. The field will immensely benefit from deconstruction of complex document understanding tasks into modular subtasks with corresponding datasets and tailored evaluation to push state-of-the-art.

DOI: 10.17148/IJARCCE.2024.13126

III. NLP TECHNIQUES

Natural language processing spans a vast repertoire of AI techniques empowering machines to comprehend human language, unlock information from text and facilitate intelligent dialogue. We survey key NLP approaches holding immense significance in extracting structured knowledge and insights from documents across diverse domains. Named entity recognition offers the critical capability to identify spans of text corresponding to real-world entities such as person names, location descriptors, medical codes or financial indicators within unstructured documents in an automated manner. Relation extraction builds on this to detect named relations between entities, revealing connections such as vendor-customer or manager-employee relationships. Together, NER and relation extraction convert free-flowing text into structured quarriable knowledge. Another technique called coreference resolution links entity mentions in text that refer to the same real-world entity, resolving ambiguities and enhancing context. Knowledge base construction assimilates entities, relations and disambiguated mentions into structured graph databases to enable inference, question answering and improved information access. Each NLP technique thus complements the others in the collective quest towards deeper language understanding. However, multi-step NLP pipelines often suffer from propagation and aggregation of errors from preliminary text processing into later interpretation tasks. Our survey revealed breakthrough advances in contextual representations from bidirectional transformer language models like BERT that have significantly pushed state-of-the-art in many information extraction tasks using just minimal task-specific fine-tuning rather than extensive task-specific architectures. We specifically highlight applications in financial document processing and biomedical concept extraction leveraging such contextual embeddings.

Additionally, we analyse emerging techniques in abstractive summarization employing deep learning sequence-to-sequence architectures to obtain compressed textual representations condensing the gist yet preserving semantic completeness. We also discuss key phrase extraction techniques utilizing BERT and sentence transformers to extract salient topics and concepts from voluminous documents in an automated manner. This aids in discovering pivotal pieces of information. Considering the linguistic diversity of global communities, advancing multilingual NLP poses unmet challenges yet immense value.

Our survey dives deeper into active areas of machine translation research spanning statistical to cutting-edge neural approaches together with associated obstacles for translation between many language pairs caused by lack of parallel corpora. Speech interfaces introduce additional complexity, also demanding language-specific considerations during recognition and synthesis model development in contrast with English-centric solutions. In conclusion, while NLP has achieved formidable performance on certain benchmark English tasks, mainstream adoption and democratization for regional languages hinges on expanding annotated corpora representing linguistic subtleties together with establishment of representative shared tasks. However, cross-lingual transfer learning and on-device edge computing offer promising pathways to make state-of-the-art NLP accessible beyond languages with abundance of task-specific training data resources.

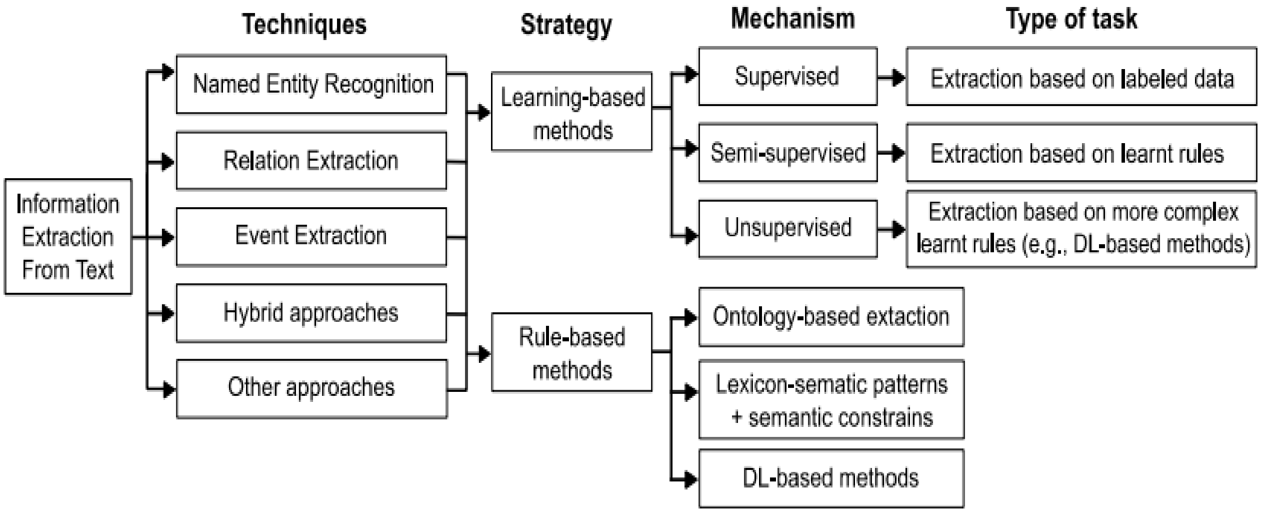

Fig. 1 IE hierarchies according to techniques, methods, mechanisms of operation, and types of tasks

DOI: 10.17148/IJARCCE.2024.13126

TABLE I Table Analysis

SL NO	Paper Title	Techniques	Addressed Issue
1	End to End Invoice Processing Application Based on Key Fields Extraction	Template matching, image processing, deep learning (specifically YOLOv5), and optical character recognition (OCR) using the Tesseract OCR software	Manual invoice processing and proposes an automatic system to extract key information from invoices
2	Exploring the Landscape of Automatic Text Summarization	Template-based or rule-based, optical character recognition (OCR)	Handling complex document layouts in real-time situations such as invoices and purchase orders
3	Handwritten text recognition in south Indian languages	Convolutional Neural Networks (CNNs) and classifier combination methods	Cursive and complex structure of handwriting, the unavailability of temporal information in handwriting recognition
4	Multi-Layout Unstructured Invoice Documents Dataset	Feature extraction techniques such as Glove, Word2Vec, FastText, as well as AI approaches like BiLSTM and BiLSTM-CRF	Lack of annotated and high-quality datasets, poor-quality images, domain-related datasets, and a lack of data validation approaches to evaluate data quality
5	Summarizing Newspaper Articles using Optical Character Recognition and Natural Language Processing	image segmentation, optical character recognition (OCR), and natural language processing (NLP)	Complexity of newspaper layouts, the quality of scanned pages, and the fluency of language in the articles
6	Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence	template-based or rule-based, named entity recognition (NER), optical character recognition (OCR), robotics process automation (RPA),	Handling complex document layouts in real-time situations, processing multiple layouts of unstructured documents
7	Improving Optical Character Recognition performance for low quality images	Optical Character Recognition (OCR), applying sharpening and blurring filters, and using a clustering method to extract text from colorful backgrounds	Low image resolution, image quality, and the presence of colorful backgrounds, which can negatively impact OCR performance
8	Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout	Convolutional Neural Network (CNN) and Support Vector Machine (SVM)	Open databases, handwriting variations, and freestyle writing
9	Natural Language Processing for Information Extraction	Named Entity Recognition, Coreference Resolution, Named Entity Linking, Relation Extraction, and Knowledge Base reasoning	Time-consuming and the propagation of errors from low-level tasks to high-level tasks in NLP
10	Decoding Knowledge Transfer for Neural Text-to-Speech Training	Multi-teacher knowledge distillation (MT-KD) for neural text-to-speech (TTS) training	Exposure bias problem in autoregressive models used for text-to-speech (TTS) training
11	Multi-Layout Invoice Document Dataset (MIDD): A Dataset for Named Entity Recognition	Named Entity Recognition, rule based approach, deep learning models like Recurrent Neural Networks (RNNs) or Transformer-based models like BERT	Poor-quality document images, obsolete formats, domain-specific and task-specific datasets, and lack of labelling
12	A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data	BERT (Bidirectional Encoder Representation Transformers) and Sentence Transformer for semantic keyphrase extraction	Time-consuming and labor intensive process of keyphrase extraction

13	A survey on speech synthesis techniques in Indian languages	Concatenative synthesis, formant synthesis, articulatory synthesis, syllable-based synthesis, HMM-based synthesis, statistical parametric synthesis, polyglot synthesis, multilingual synthesis, and waveform concatenation using deep learningz	Language-specific characteristics and the need for language-specific data
14	COMPREHENSIVE OVERVIEW OF NAMED ENTITY RECOGNITION: MODELS, DOMAIN-SPECIFIC APPLICATIONS AND CHALLENGES	Named Entity Recognition (NER), modern deep learning techniques such as BERT and transformers, Optical Character Recognition (OCR)	NER in finance and biomedicine
15	An Effective Neural Machine Translation for English to Hindi Language	Neural Machine Translation (NMT) and statistical machine translation techniques	Machine translation(E->H), Lack of Resources, Word Order Differences
16	Seamless Integration of Common Framework Indian Language TTSes in Various Applications	Syllable-based concatenative speech synthesis	Usability, Integration with applications, accessibility for visually challenged individuals
17	PARALLEL TEXT-TO-SPEECH WITH PITCH PREDICTION	Text-to-speech synthesis, Parallel/non-autoregressive neural network architecture	Slow speed of autoregressive models, Lower quality of non-autoregressive models
18	Expressive TTS Training with Frame and Style Reconstruction Loss	Tacotron architecture	Manual selection of style tokens during run-time inference
19	A comprehensive survey on Indian regional language processing	machine translation, Named Entity Recognition (NER), Sentiment Analysis, and Parts-Of-Speech (POS) tagging	Availability of resources and datasets
20	Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data	TTS-VC transfer learning (TTL-VC), sequence-to-sequence encoder-decoder architecture, Tacotron-2 as the TTS framework	limited availability of parallel training data for voice conversion, lack of linguistic grounding in autoencoder-based voice conversion methods

IV. CONCLUSION

Through this extensive survey, we aimed to explore and synthesize the landscape of techniques utilized in the realms of document processing and natural language understanding. As researchers, we set out to comprehend developments in extracting information from unstructured documents and textual data – a capability holding profound significance in the modern digital era. Our analysis revealed promising interdisciplinary techniques blending computer vision, NLP and machine learning to unlock insights concealed within invoices, contracts, reports, articles and manuals. From optical character recognition to contextual language models, the field has attained remarkable milestones. Template matching defines the contours of documents while learning algorithms interpret the content through continuously advancing neural approaches.

However, challenges endure in deciphering intricate details from diversified, complex document formats at industrial scales. Cursive handwriting recognition for regional languages warrants deeper investigation. Variability in invoice layouts necessitates adaptive systems adept at generalized understanding. Further augmentation of labelled datasets can catalyse progress by accounting for niche domains and writing peculiarities. As we stride towards more immersive human-machine interaction, language-centric interfaces will continue to gain prominence, elevated by multilingual and multi-modal capabilities. NLP infused within business workflows has untapped potential for revolutionizing document search, analytics and process automation. The key opportunity lies in transforming academic advancements into real-world solutions through focused engineering efforts. Our study has sought to comprehend innovations in academia and industry while crystalizing directions for channelizing further efforts. We sincerely hope the research trajectories and techniques analysed in our survey paper stimulate advancements in alleviating manual document handling bottlenecks. The journey has yielded optimistic outlooks nested in prevailing challenges. With dedicated exertion, the promise of automated insights from humanity's overwhelming legacy of written treasures can inch closer to realization

REFERENCES

[1]. R. Devika1, Subramaniyaswamy Vairavasundaram1, C. Sakthi Jay Mahenthar1, Vijaykumar Varadarajan2, And Ketan Kotecha3, "A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data", IEEE 2021, DOI: 0.1109/ACCESS.2021.3133651

[2] Sonit Singh, "Natural Language Processing for Information Extraction", Department of Computing, Faculty of Science and Engineering, Macquarie University, Australia 2018, DOI: arXiv:1807.02383v1 [cs.CL]

[3] Mingyang Zhang, Member IEEE, Yi Zhou, Student Member, IEEE, Li Zhao, and Haizhou li, Fellow, IEEE, "Transfer Learning from Speech Synthesis to Voice Conversion With Non-Parallel Training Data", IEEE 2021, DOI: 10.1109/TASLP.2021.3066047

[4] B.S. Harish1, R. Katuri Rangan1, "A comprehensive survey on Indian regional language processing", Springer Nature Switzerland AG 2020, DOI: 10.1007/s42452-020-2983-x

[5] Rui Liu, Member, IEEE, Berrak Sisman, Member, IEEE, Guanglai Gao, Haizhou Li, Fellow, IEEE, "Expressive TTS Training with Frame and Style Reconstruction Loss", IEEE 2020, DOI: 10.1109/TASLP.2021.3076369

[6] Adrian Łancucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction", NVDIA Corporation, IEEE 2021, DOI: 10.1109/ICASSP39728.2021.9413889

[7] Pranaw Kumar1 , Gandhi Annamalai2 , Sajini T3 , Anand Konjengbam4 , Praveen M2 , G R Kasthuri2 , Anil Prabhakar2 , Shuo Qiao2 , Sushanta K Pani1 , Vishal Maral1 , Binil Kumar S L3 , Arun Gopi3 , Neethu E A3 , Bira Chandra Singh1 , Ranbir Singh4 , Hema A Murthy2, "Seamless Integration of Common Framework Indian Language TTSes in Various Applications", IEEE 2021

[8] Saikiran Gogineni1, G. Suryanarayana2, Sravan Kumar Surendran3, "An Effective Neural Machine Translation for English to Hindi Language", IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

[9] Rui Liu , Member, IEEE, Berrak Sisman , Member, IEEE, Guanglai Gao, and Haizhou Li , Fellow, IEEE, "Decoding Knowledge Transfer for Neural Text-to-Speech Training", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 30, 2022

[10] Soumya Priyadarsini Panda1, Ajit Kumar Nayak2, Satyananda Champati Rai3, "A survey on speech synthesis techniques in Indian languages", Springer-Verlag GmbH Germany, part of Springer Nature 2020, DOI: 10.1007/s00530-020-00659-4

[11] Dipali Baviskar1, Swati Ahirrao1, Ketan Kotecha2, "Multi-Layout Invoice Document Dataset (MIDD): A Dataset for Named Entity Recognition", 2021, DOI: 10.3390/ data6070078

[12] Kalyani Pakhale, "Comprehensive overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges", 2023, DOI: arXiv:2309.14084v1

[13] Amani Ali Ahmed Alia,b , Suresha Mallaiahb , "Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout", Elsevier 2021, DOI: 10.1016/j.jksuci.2021.01.012

[14] Matteo Brisinello1, Ratko Grbic1, Matija Pul2, Tihomir Andelic3, 1University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology, Kneza Trpimira 2b, Osijek, Croatia, "Improving Optical Character Recognition performance for low quality images", Elvesier 2020

[15] Dipali Baviskar1, Swati Ahirrao1, Vidyasagar Potdar2, Ketan Kotecha3, "Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions", IEEE 2021, DOI: 10.1109/ACCESS.2021.3072900

[16] Halil Arslan, "End to End Invoice Processing Application Based on Key Fields Extraction", IEEE 2022, DOI: 10.1109/ACCESS.2022.3192828

[17] Bilal Khan1, Zohaib Ali Shah1, Muhhamad Usman2, (Senior Member, IEEE), Inayat Khan2, Badam Niazi3, "Exploring the Landscape of Automatic Text Summarization: A Comprehensive Survey", IEEE 2023, DOI: 10.1109/ACCESS.2023.3322188

[18] A. T. Anju1*, Binu P. Chacko2 and K. P. Mohammad Basheer3, "Review of offline handwritten text recognition in south Indian languages", 2021, DOI: 10.26637/MJM0901/0132

[19] Dipali Baviskar1, Swati Ahirrao1, Ketan Kotecha2, "Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches", IEEE 2021, DOI: 10.1109/ACCESS.2021.3096739

[20] Shashank Sanjay Tomar, "Summarizing Newspaper Articles using Optical Character Recognition and Natural Language Processing", 2022