Natural Language Processing Journal

A review on knowledge and information extraction from PDF documents and storage approaches

--Manuscript Draft--

Manuscript Number:	NLP-D-24-00048
Full Title:	A review on knowledge and information extraction from PDF documents and storage approaches
Article Type:	Review
Keywords:	Natural Language Processing; Language Models; Information Extraction; Knowledge Base
Corresponding Author:	Salvador Desconsciences Atagong, Master International Centre for Insect Physiology and Ecology Nairobi, State/Region KENYA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	International Centre for Insect Physiology and Ecology
Corresponding Author's Secondary Institution:	
First Author:	Salvador Desconsciences Atagong, PhD
First Author Secondary Information:	
Order of Authors:	Salvador Desconsciences Atagong, PhD
	Henri Tonnang
	John Odindi
Order of Authors Secondary Information:	
Abstract:	Automating the extraction of information from Portable Document Format (PDF) documents represents a valuable information extraction milestone poised to alleviate an important part of manual labour involved and facilitate knowledge discovery across diverse domains. However, the reliability of contemporary solutions designed for this task remains a subject of contention in terms of accuracy, domain adaptability and efforts engaged in providing reliable solutions. This study explores the underpinnings of information extraction from PDF documents by conducting a comprehensive review in the realm of automatic information extraction from PDF documents. The review not only discerns prevailing trends, but also elucidates the limitations entrenched in the existing rule-based, statistical learning and neural network-based methodologies and underscores the critical gaps in the field of automatic information extraction from textual data. To address these challenges, the study introduces a conceptual framework consisting of eight essential components: document manager, document pre-processor, ontology manager, information extractor, annotation engine, question answering tool, knowledge visualizer and data exporter. This comprehensive framework is designed to streamline the information extraction process from PDF documents, leading to improved accuracy, domain adaptability, and ease of use.
Suggested Reviewers:	HITHAM SEDDIG ALHASSAN ALHUSSIAN, PhD Associate Professor, Universiti Teknologi PETRONAS seddig.alhussian@utp.edu.my He co-authored a review that inspired our review methodology.
	Mariana Lara Neves Federal Institute for Risk Assessment, German Centre for the Protection of Laboratory Animals (Bf3R) marianalaraneves@gmail.com She conducted a thorough review on document annotation which helps a lot in our work.

Additional Information:	
Question	Response

January, 05th, 2024

Subject: **Manuscript Submission**"**A review of knowledge and information extraction from PDF documents and storage approaches**"

Dear Editor
Natural Language Processing Journal

I am writing to formally submit my manuscript titled "***A review of knowledge and information extraction from PDF documents and storage approaches***" for your consideration and possible evaluation for publication in the Natural Language Processing Journal. This study was conducted within the context where a substantial portion of scientific knowledge is available exclusively in PDF format through peer-reviewed articles and various scientific reports. Given the exponential growth in the volume of PDF-based knowledge, especially in fields like health and entomology, the need for efficient extraction of key information from this vast database has become increasingly urgent. This endeavor has the potential to significantly contribute to the discovery of new knowledge across multiple domains.

The primary finding of this comprehensive review revealed that approaches for automatically extracting specific information from PDF documents can be classified into three main categories: rule-based, machine-learning-based, and deep-learning-based. However, each of these categories has its limitations, further compounded by the inherent heterogeneity of PDFs, which encompass various types of information (images, tables, text, etc.), and the intricacies of their layout, which can be challenging to decipher due to the absence of a semantic layer within the PDF itself. Additional limitations were attributed to ambiguity resolution and natural language understanding challenges, coupled with the scarcity of high-quality, domain-specific annotated text corpora.

In response to these challenges, we propose a novel conceptual framework that combines rule-based approaches, utilizing common ontologies, with Large Language Models. This innovative approach aims to mitigate the impact of document annotation issues and enhance domain-specific information extraction. We would like to confirm that there are no conflicts of interest among the authors, no financial support or affiliations with other institutions that could influence the research findings. Furthermore, we certify that this manuscript strictly adheres to the journal's prescribed format, as outlined in the instructions for authors. It has not been published elsewhere and is not currently under consideration by another journal. All co-authors have reviewed and approved the manuscript and consent to its submission to the Journal of Data Mining and Knowledge Discovery. We sincerely hope that our manuscript aligns with the standards and objectives of your esteemed journal and can be considered for publication. We greatly appreciate your kind cooperation throughout this process.

Thank you for your time and attention

Sincerely,

Salvador D. Atagong

Declaration on interest statement

We confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

A review on knowledge and information extraction from PDF documents and storage approaches

Salvador Atagong

International Centre of Insect Physiology and Ecology (icipe), Nairobi, Kenya

satagong@icipe.org

Henri E.Z. Tonnang

International Centre of Insect Physiology and Ecology (icipe), Nairobi, Kenya

htonnang@icipe.org

John Odindi

University of KwaZulu Natal, South Africa

odindi@ukzn.ac.za

Corresponding author*:* asalvador@icipe.org

A review on knowledge and information extraction from PDF documents and storage approaches

Abstract

Automating the extraction of information from Portable Document Format (PDF) documents represents a valuable information extraction milestone poised to alleviate an important part of manual labour involved and facilitate knowledge discovery across diverse domains. However, the reliability of contemporary solutions designed for this task remains a subject of contention in terms of accuracy, domain adaptability and efforts engaged in providing reliable solutions. This study explores the underpinnings of information extraction from PDF documents by conducting a comprehensive review in the realm of automatic information extraction from PDF documents. The review not only discerns prevailing trends, but also elucidates the limitations entrenched in the existing rule-based, statistical learning and neural network-based methodologies and underscores the critical gaps in the field of automatic information extraction from textual data. To address these challenges, the study introduces a conceptual framework consisting of eight essential components: document manager, document pre-processor, ontology manager, information extractor, annotation engine, question answering tool, knowledge visualizer and data exporter. This comprehensive framework is designed to streamline the information extraction process from PDF documents, leading to improved accuracy, domain adaptability, and ease of use.

Key words: Natural Language Processing, Language Models, Information Extraction, Knowledge Base.

1 Introduction

Natural language has been employed for centuries to convey information and knowledge, primarily through printed documents such as the Bible, the Koran and several mythologies and civilisation archives. For many years, conserving these physical documents has been challenging due to inherent vulnerabilities that include sensitivity to temperature, paper degradation and fires. However, in the recent past, digital documents have become increasingly popular due to their space-saving, ease of sharing, and enhanced security features. According to Johnson (2021), the Portable Document Format (PDF) is one of the most widely used formats for digital documents, accounting for more than 83% of documents shared over the web (Duf Johnson, 2021). In comparison to physical documents, this prevalence can be attributed to their platform-independence and the ability to preserve original documents formats. According to Abdillah and Leon Andretti, (2013), Nganji, (2015), and Newe and Becker, (2018), PDFs account for a significant portion of scholarly documents, while Bornmann and Mutz, (2015) notes that their creation rate has grown exponentially over years. This growth has meant that the task of collecting and extracting specific information from a large volume of PDF documents has become arduous and time-consuming.

Many studies (e.g. Gupta and Gupta, 2012; Abdollahi et al., 2021; Chen et al., 2022; Guan, Du, and Yang, 2022; Nundloll et al., 2022; Yang, Han and Poon, 2022) have endeavored to address the challenge of automatically extracting specific information from PDF documents. These efforts primarily leverage Natural Language Processing (NLP) algorithms and Optical Character Recognition (OCR) techniques. NLP encompasses a set of computational techniques designed for the automatic analysis and representation of human languages grounded in theoretical foundations as emphasised by Chowdhary,

(2020). These techniques have been extensively adopted to extract information from a wide range of written sources, facilitating the discovery of new and previously undisclosed information in textual data (Chen et al., 2022). According to Chowdhary, (2020) and Abdullah et al., (2023), Information Extraction (IE) in literature has been broken into many sub-tasks namely; 1) Named Entity Recognition (NER) that aims to extract named entities from a given text corpus; 2) Relationship Extraction (RE) that focuses on extracting relationships between the named entity of a given corpus; 3) Question Answering (QA) aimed at answering natural language questions and highly dependent on the two previous sub-tasks; 4) Knowledge Extraction dedicated to building a knowledge base from a text corpus; 5) Event Extraction (EE) aimed at identifying events and all their properties (e.g. organizer, time) from a text corpus and 6) Causality Extraction (CE), that aims to extract cause-effect relations between pairs of labelled nouns from text (Yang et al., 2022).

Furthermore, IE approaches have been generally classified into three categories namely, rule-based approaches, statistical learning-based approaches, and neural network approaches (Abdullah et al., 2023; Mannai et al., 2018). To gain a comprehensive understanding of why and how IE is performed from PDFs and identify the challenges encountered along with potential unaddressed gaps, we structured our review around the following research questions:

1. What motivates information or knowledge extraction from PDF documents?
2. Which techniques or algorithms are used for automated information extraction, and what difficulties are encountered?
3. How are the performance and effectiveness of these techniques or algorithms evaluated?
4. In what ways is the extracted information or knowledge stored and represented?

Our study further provides a comprehensive overview of recent developments in the field of IE from PDF documents from 2017 to 2023. During this timeframe, significant advancements have been made in the field of NLP, with the introduction of Transformers (Vaswani et al., 2017) and Language Models (Cabot and Navigli, 2021; Devlin et al., 2018). In the review, we aim to not only delineate the current trends in this domain, but also to identify persisting challenges. Furthermore, we propose an innovative approach and pipeline for information extraction from text, which amalgamates language models with common ontologies to fine-tune and oversee the entire extraction process, enhancing its adaptability across diverse domains.

The subsequent sections of the manuscript are organized as follows: Section 2 introduces fundamental concepts in IE, Section 3 presents the methodology we used to select resources from published literature, Section 4 provides a summary and classification of the identified approaches in the literature and presents their inherent limitations, Section 5 is the discussion and introduction of an innovative conceptual framework for information and knowledge extraction, while Section 6 concludes the review.

2 Background

2.1 Natural Language Processing

Information Extraction (IE) can be defined as the automated process of extracting structured information from unstructured text, particularly natural language texts (Yang, Han, and Poon, 2022). It typically involves a series of steps, some of which are universally applied, while others depend on the specific type of information to be retrieved from the text. Among these steps, text pre-processing assumes a pivotal role, aiming to clarify the initial text as comprehensively as possible to facilitate subsequent task-specific processes (Chowdhary, 2020). Text pre-processing generally includes sub-tasks such as tokenization, lemmatization, and stemming. Tokenization involves dividing a character sequence into discrete units,

referred to as tokens, and often entails the removal of certain characters such as punctuation marks. The resulting list of tokens serves as a basis for further analysis (Webster and Kit, 1992). Lemmatization is a process that involves morphological analysis of words, grouping together various inflected forms of a word to treat them as single entity. To lemmatize documents effectively, specifying the part of speech (POS) for each word is necessary. However, due to the the laborious and error-prone nature of POS tagging, stemming methods are commonly preferred in practical applications (Yang et. al, 2022). Stemming is the task of obtaining the root or stem of the derived words. Stemming algorithms are language dependent, meaning that the stem of a word in one language may differ from that in another. For instance, in English, the stem of "eating" is "eat" (Yang et al., 2022).

Various types of information can be extracted from text, the following being the most common categories in literature, or often a combination of two more: 1) Named entities (Cabot and Navigli, 2021; Varshini and Uthra, 2022) (locations, peoples, drugs, disease vectors, events, etc.); 2) Relationship between named entities (Cabot and Navigli, 2021; Devlin et al., 2018); 3) Causal relationships between entities (Jin et al., 2020; Li et al., 2021) and 4) Temporal events (Qiu et al., 2020).

Bird et al., (2009) introduced a comprehensive architecture for IE that combines named entity recognition and relationship extraction. This approach can be considered as the classical or standard approach for extracting information from text, and many methodologies align with it, including those by (Chen et al., (2022) and Nundloll et al., (2022). It comprises five interconnected steps: sentence segmentation, tokenization, part-of-speech tagging, entity detection, and relation detection (see Figure 1).

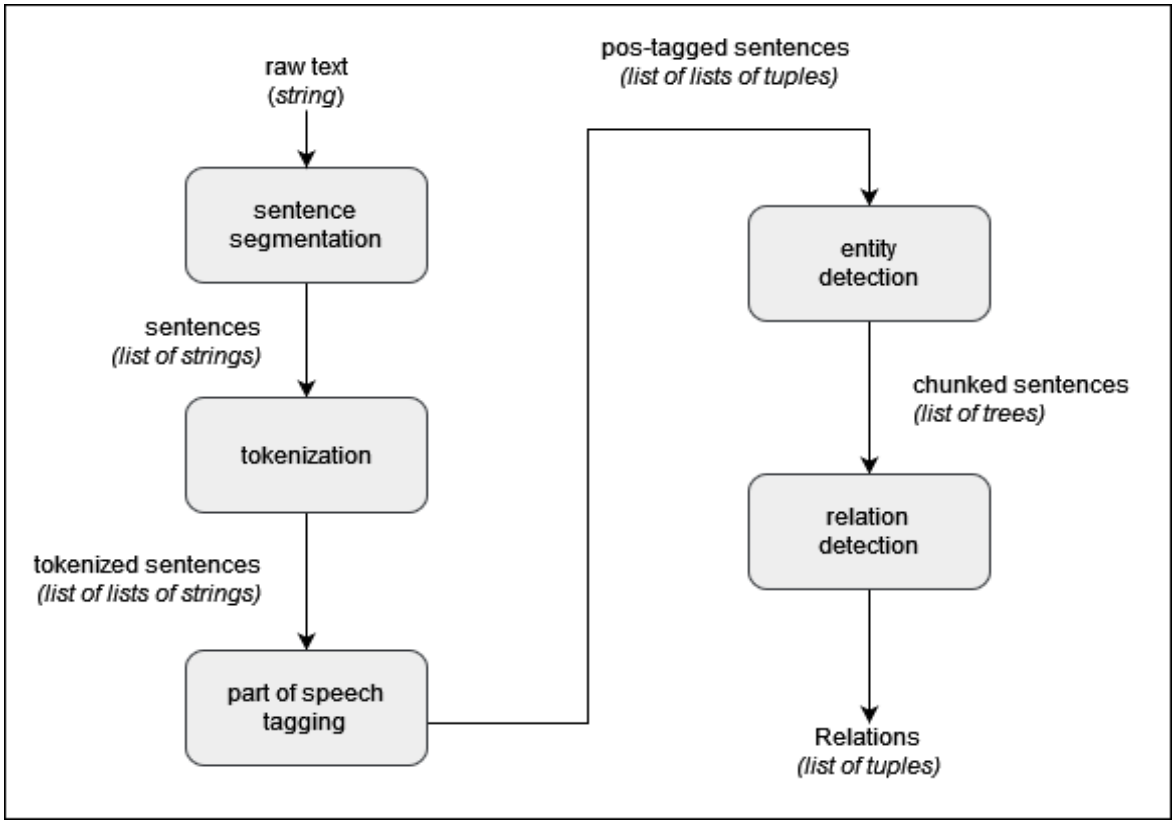

Figure 1. *Information Extraction (IE) architecture visually represents the key steps and components involved in extracting structured information from unstructured text. This architecture serves as a foundational framework for various methodologies in natural language processing and information extraction.*

The primary objective of Named Entity Recognition (NER) is to identify and categorize significant nouns and proper nouns within a given text (Peters et al., 2018), or, in a more general sense, to identify specific types of concepts such as names of places, organizations and people within the text. It is generally regarded as a classification problem, a perspective supported by studies such as Cabot and Navigli, (2021), Nundloll et al., (2022) and Sung et al., (2022). NER approaches can be categorized into three primary categories: rule-based, statistical, and hybrid approaches. In the early stages of NER, rule-based methods dominated the field. These approaches typically comprise three major components:

1) a collection of named entity extraction rules, 2) gazetteers containing various named entity classes, and 3) an extraction engine responsible for applying these rules and lexicons to the text. The rule sets and lexicons were either meticulously crafted by human experts or generated through bootstrapping from a limited set of handcrafted examples (Peters et al., 2018). The effectiveness of rule-based systems typically hinges on the comprehensiveness of the rules and lexicons in use. Introducing deeper knowledge beyond surface-level words and lexicons into a rule-based system necessitates substantial manual efforts, which can be costly (Peters et al., 2018). Statistical and Neural Network NER approaches, in contrast, alleviate the need for extensive manual effort in constructing rule sets and gazetteers. They typically rely on two primary components: labeled training data and a model, which can encompass statistical models, neural networks, or deep neural networks. Labeled training data involves text corpora annotated with named entities, usually created manually with software or document annotation tools (refer to Section 2.2).

Additionally, Relationship Extraction (RE) is an important task in natural language text understanding and heavily depends on Named Entity Recognition (NER) and is usually performed after the latter. However, Cabot and Navigli, (2021) present a sequence-to-sequence approach in their study performing NER and RE concurrently. In performing the NER and RE concurrently, a tuple *t* relationship characterized by equation (1) is usually defined.

$$(1). \quad t = (e_1, e_2, \dots, e_n)$$

where the e_i are entities in a pre-defined relation r within a given document (Bach and Badaskar, 2007). Most studies in literature commonly focus on binary relationships (between two entities only). The Relation Extraction problem is considered a classification problem to classify tuples of entities extracted from the text. Statistical models are usually used at this level, either supervised (like Support Vector Machines) or semi-supervised (like DRIPE) (Bach and Badaskar, 2007). Deep Neural Networks like Convolutional Neural Networks (CNN) or transformer-based models are also used (Bacchi et al., 2022; Banerjee et al., 2019; Cabot and Navigli, 2021).

Event Extraction (EE) encompasses deducing specific knowledge about incidents referred to in texts. Event extraction approaches fall into two categories: 1) data-driven approaches through the usage of statistics, machine learning, and linear algebra; and 2) knowledge-driven approaches which extract knowledge through representation and exploitation of expert knowledge, usually by means of pattern-based approaches or ontologies (Hogenboom et al., 2011). However, a third category is composed of the hybridization of the two categories of methods mentioned earlier. This task can be considered as a specific case of a combination of NER and RE, where the relations among concepts are event-specific such as organized in, from and to in English texts.

In 2017, a major advancement was established in the domain of Natural Language Processing by the publication of Google research Transformer model (Vaswani et al., 2017), which is the first sequence model based entirely on the attention mechanism. Based on this advancement, Devlin et al., (2018) proposed a novel language representation model called Bidirectional Encoder Representations from

Transformers (BERT) with significant improvement in state-of-the-art performance in many NPL tasks such as NER, QA, and RE. From this base-model, many derivations were created to further improve the capabilities of language models in natural language understanding (Beltagy et al., 2019a; Lee et al., 2019; Lewis et al., 2020).

2.2 Document annotation

Document annotation is a process whose goal is to add metadata to a text corpus to highlight the relevant information shared in the text. This is extremely critical for providing reliable datasets for training models (Neves and Ševa, 2021). However, the process often highly depends on human efforts. Considerable efforts are underway using software and tools to assist annotation, therefore reducing the annotation time.

In their 2021 study, Neves and Ševa, (2021) conducted an informative review which presented the most common missing features across available tools (e.g., PrettyTags, PDFAnno, Djangology, WebAnno, etc.). Between 2021 and 2023, additional tools were reported with the following considered as the top missing features (Di Martino et al., 2021; Neumann et al., 2021):

- Lack of scientific publications supporting the development of most tools,
- Insufficiency of online available tools,
- Unavailability of the source code of most tools,
- Document-level annotation is not supported by most tools,
- Lack of embedded integration with online scholar databases (e.g., PubMed),
- Lack of ontology-based tools, thus limiting easy domain adaptability,
- Lack of tools that integrates Inter Annotator Agreement (AAI) configuration,
- Inability of most tools to support multi-labeled annotation,
- Insufficient support of team management by most tools.

3 Research Methodology

This review adopted the PRISMA methodology (Liberati et al., 2009). We conducted a systematic search for scientific papers published between January 1, 2017, and March 31, 2023, using publicly accessible databases namely Web of Science (WOS), IEEE Digital Library, and Google Scholar. Our review process encompassed four key steps: preliminary study, publication screening, assessment of eligibility and quality, and synthesis of the included studies (Figure 2).

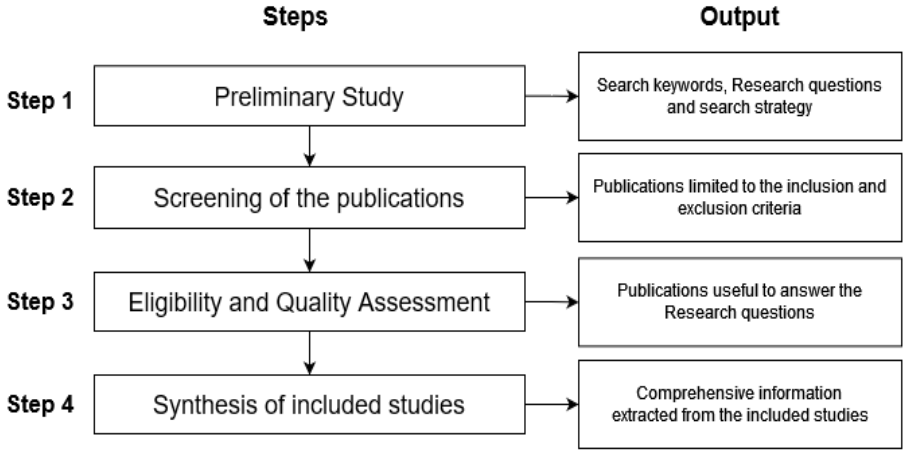

Figure 2*: An overview of the review process*

3.1 Preliminary study

This step sought to identify key terms and fundamental attributes associated with the specific scientific publications to be retrieved. To achieve this, we defined a set of criteria that governed the inclusion or exclusion of publications from the outputs of online databases (see Table 1).

Table 1: *The inclusion and exclusion criteria.*

Inclusion	Exclusion
• Publication year between 2017 and 2023, • Publication is in English language, • PDF means Portable Document Format	• Publication is not in English language, • PDF doesn't mean Portable Document Format (e.g., Probability Density Function)

3.2 Screening of publications

The screening process was to identify more suitable publications that could help us in answering our research questions. We were able to define pertinent keywords and a generic query to refine our searches on selected digital resource databases, thereby facilitating the retrieval of pertinent publications. Additionally, we tailored our generic query to suit the specific requirements of each public database, resulting in a set of distinct search outputs. These outputs were subsequently amalgamated, and any duplicate entries were eliminated, culminating in the creation of a unified and cohesive repository of publications for our comprehensive review.

Initially, we compiled a database consisting of 256 papers (Figure 3). Following our inclusion and exclusion criteria (Table 2), we refined this dataset, reducing it to 213 papers (Figure 4).

Table 2: *Keywords and generic request for peer reviewed journal retrieval form public databases (Web Of Science, IEEE explore).*

Keywords	Generic request
• information extraction • knowledge extraction • nlp • natural language processing • named entity recognition. • named entity extraction. • relation extraction • relationship extraction • event extraction • unstructured document • portable document format • pdf	(Title/Abstract CONTAINS "Information extraction" OR Title/Abstract CONTAINS "knowledge extraction." OR Title/Abstract CONTAINS "NLP" OR Title/Abstract CONTAINS "natural language processing." OR Title/Abstract CONTAINS "named entity recognition." OR Title/Abstract CONTAINS "named entity extraction." OR Title/Abstract CONTAINS "relation extraction." OR Title/Abstract CONTAINS "relationship extraction." OR Title/Abstract CONTAINS "event extraction.") AND (Title/Abstract CONTAINS "Unstructured document" OR Title/Abstract CONTAINS "portable document format." OR Title/Abstract CONTAINS "PDF") AND (Publication Year BETWEEN "2017" AND "2023")

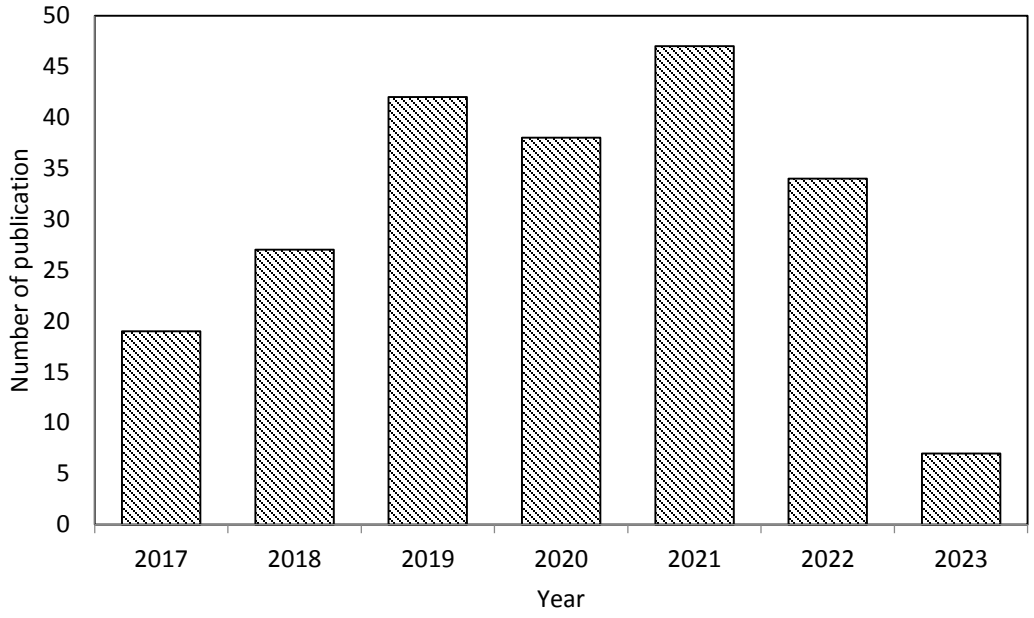

Figure 3*: Initial distribution of fetched publications over years.*

3.3 Eligibility and quality assessment

To further refine our selection, we adopted a rating methodology introduced in Abdullah et al., (2023) to formalize the selection process by quantifying the quality of papers. As shown in Table 3, we established a structured questionnaire to evaluate each publication and retained only papers with a score greater than 3, with a minimum score of 1 on the first question and at least 0.5 on the fourth question (see Equation 2). The rationale for these values is rooted in the criteria detailed in Table 3. Specifically, a score of 1 on the first question indicated that the study's objectives were clear and focused on IE, while a 0.5 score on the fourth question indicated that the study had at least proposed an evaluation of its performance. In addition, a total of at least 3 ensured that the IE methodology was clearly outlined within the study. The final selection was distributed as presented in Figure 4, and the rating results summarized in Figure 5.

Table 3: *Papers rating scale.*

Code	Criteria	Score	Description
C1	Does the study define clear objectives, and do they meet our research question?	1	yes, the study presents clear objectives and goals, which are clearly related to information extraction from text
		0.5	the study presents its objectives, but the end goal is not information extraction even though it somehow intervein
		0	the study does not clearly define its objectives
C2	Does the study present clear methodology?	1	yes, the methodology of information extraction is clearly defined
		0.5	the methodology is superficial or incomplete
		0	No, the study does not present its methodology
C3	Does the study present limitations?	1	yes, the study presents its limitations in detail
		0.5	the study states its limitations but not in detail
		0	No, the study does not state its limitations

C4	Does the study evaluate its performance?	1	yes, the study is evaluated clearly using common metrics and compared to state-of-the-art methodology results in the field
		0.5	the study is evaluated but no clear metrics are provided nor clear comparison with other methodologies
		0	no metric is provided for study evaluation
C5	Does the study handle the storage aspect?	1	yes, the study presents the storage approach used to structure the saved information in detail
		0.5	the study superficially talks about the restructuring of the extracted information, but no further details are provided
		0	the study does not talk about how the extracted information is stored

(2). $\forall\, i \in [1;5];\ \forall\, x \in P;\ Pk \cup \{x\} \leftrightarrow \left(\sum_{i=1}^{5} S_i > 3\right) and \begin{cases} S_1 = 1 \\ S_4 \geq 0.5 \end{cases}$

Where Si stands for a score on the i^{th} question (Ci from Table 3),
P is the set of initial papers,
P_k is the set of kept papers.

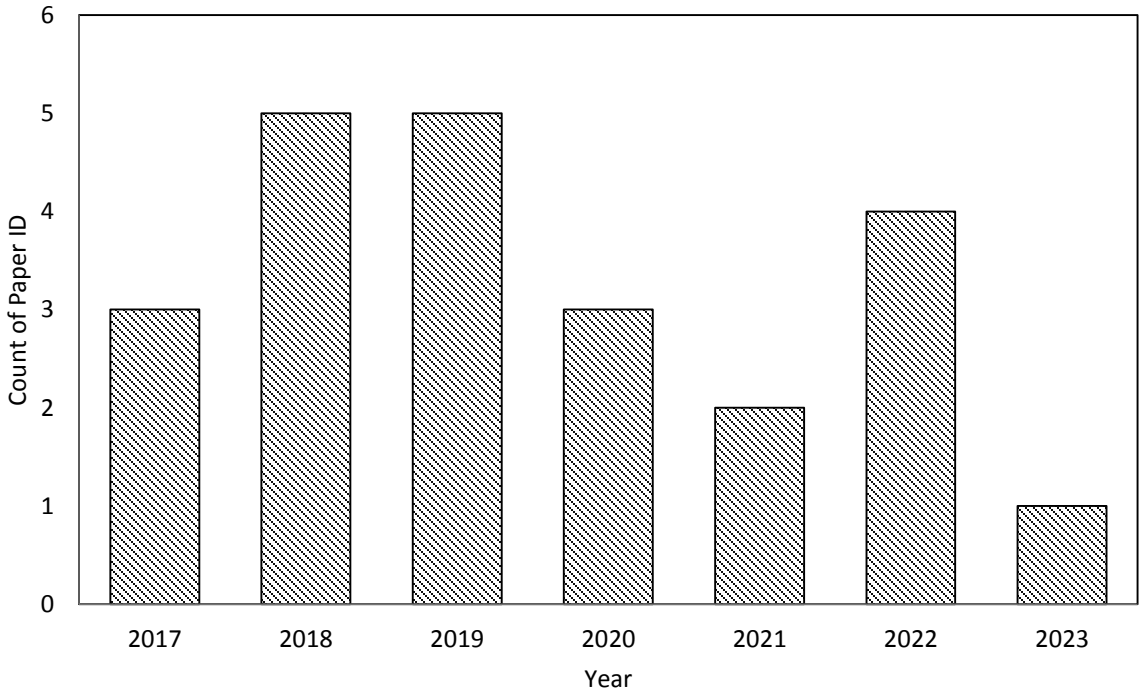

Figure 4: *Final selection distribution of retrieved publications per year.*

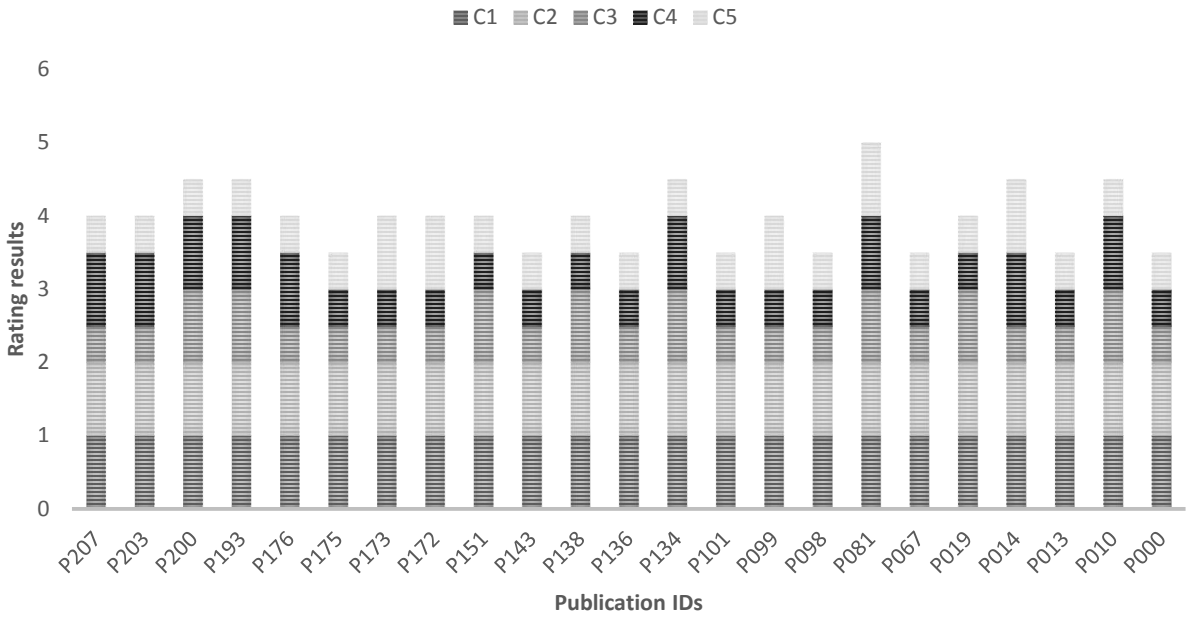

Figure 5: *Rating results of the selected publications, where each bar summarizes the marks reported for the paper on each of our evaluation questions (C1 to C5, see Table 3), with each section of the bar proportional to given mark within the range of possible marks (0, 0.5 and 1).*

4 Results

Table 1 provides a detailed representation of our inclusion and exclusion criteria. These criteria are a predefined set of rules and conditions that guided our selection process, determining studies or publications eligible for inclusion in our analysis. Their significance lies in the validity they bring to our methodology, providing readers with a clear understanding of the basis for our choices. Moreover, these criteria ensured that our selection process remains systematic and consistent, guarding against arbitrary decisions and potential bias. They also promoted research replicability and serve as a quality control mechanism to ensure that the chosen studies meet predefined standards of relevance and reliability, thus enhancing the overall credibility and validity of our study. Table 2 provides a comprehensive overview of the keywords and the generic query we employed during the search process. It serves as a structured reference to determine fundamental elements that underpin our systematic approach to literature retrieval. Specifically, it facilitates the comprehension of the specific terms and queries we employed to gather the dataset for replicability. Furthermore, it demonstrates our commitment to a systematic and rigorous approach to literature retrieval, underscoring the reliability and credibility of our research findings. Papers rating scale (Table 3) is a critical component of our research methodology, designed to quantitatively assess and evaluate the quality and relevance of research papers for inclusion in our study. This rating scale serves as a structured framework for systematically appraising each publication based on specific criteria.

Following our systematic paper selection process, we narrowed down the initial pool to a final set of 23 papers (Figure 3). The process culminated in the organization and presentation of the ratings in Figure 4. These ratings, obtained through a detailed and structured evaluation, offer a comprehensive overview of the quality and relevance of each paper. Through this process, we established a transparent and accessible reference point for the selected papers, streamlining our in-depth analysis. This careful curation of

research materials ensured that our study maintained an important level of rigor and relevance, hence a comprehensive and robust review of the chosen publications.

In 2017, the dominant approach in IE was rule-based (Figure 6). IE relied heavily on predefined rules, gazetteers, dictionaries, and domain ontologies to identify and extract information from text. These rule-based methods provided a structured way to process textual data, but often required extensive manual curation and adaptation to specific domains. However, as illustrated in Figure 6, there has been a noticeable and gradual shift in approach over time. This shift corresponds with recent advancements in Neural Network-based approaches for Natural Language Processing (NLP). Notably, the emergence of pre-trained language models has marked a transformative period in NLP. Models such as BERT, BioBERT, and GPT-4 have revolutionized the way information is extracted from text, and the way computers understand natural language texts. These pre-trained language models leverage large-scale data and advanced machine learning techniques to capture complex language patterns and semantics, making them exceptionally powerful for various NLP tasks. Their adaptability and ability to generalize across different domains have contributed to the shift away from traditional rule-based approaches. Instead, researchers are increasingly exploring the use of these neural network-based models to automate and enhance the information extraction process.

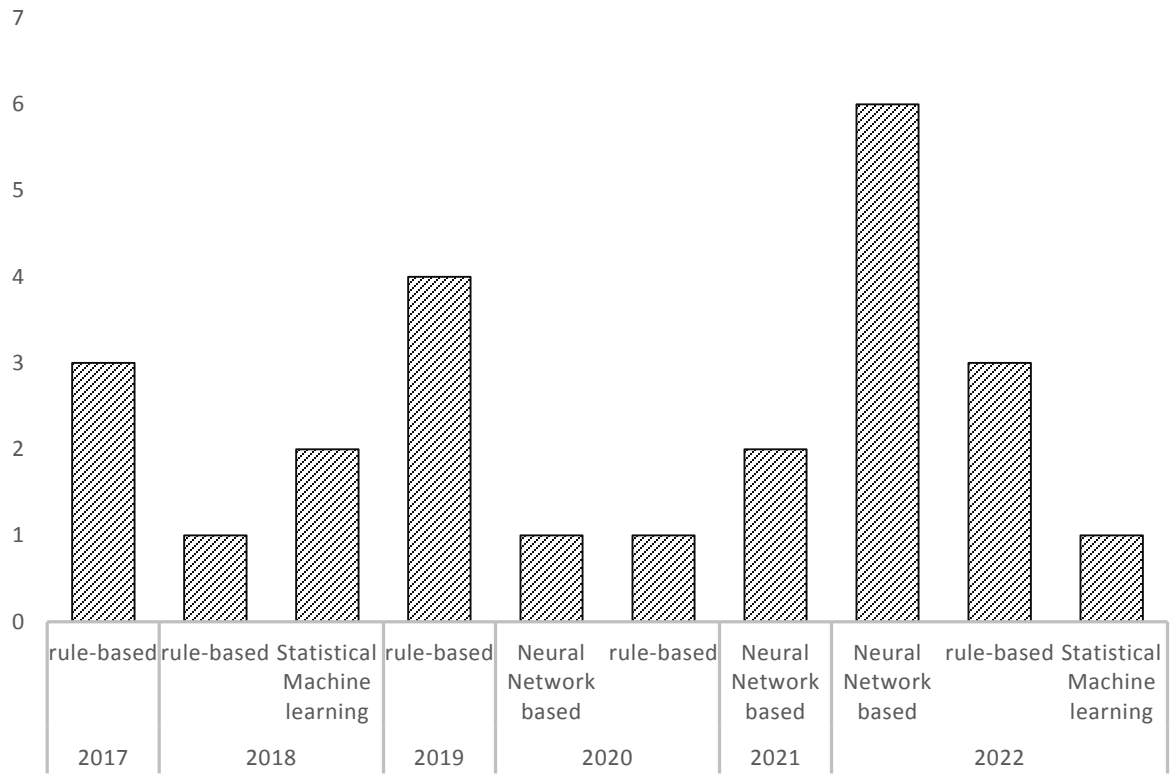

Figure 6*: Studies class distribution over year.*

While a significant portion of the consulted studies did not specify their data storage methods, the prominence of knowledge graphs indicated a growing recognition of their value in managing and organizing extracted data. This insight reflects the evolving landscape of storage approaches within the IE field, where knowledge graphs are emerging as a preferred solution for representing and harnessing structured knowledge from unstructured text (see Table 4 in the annex).

From the analysis of retained papers, we observed that the process of IE from PDF documents can be distilled into two overarching and interconnected steps: PDF document preprocessing and Effective Information Extraction (EIE). These steps are fundamental to the successful extraction of valuable insights from PDF files and represent a structured approach to navigating the challenges posed by this prevalent document format.

The initial step involves preparing and structuring the PDF document for subsequent analysis. PDFs often contain various elements, including text, images, tables, and formatting complexities. PDF document pre-processing encompasses tasks such as text extraction, layout analysis, and the identification of structural elements. During this phase, researchers or algorithms work to disentangle and organize the raw PDF data, making it amenable to further processing. Efficient PDF document preprocessing is crucial as it sets the foundation for accurate EIE.

Once the PDF document is suitably pre-processed, the focus shifts to IE's core task. This step involves identifying and extracting specific information or knowledge from the pre-processed PDF data. IE methods, which can range from traditional rule-based approaches to modern neural network-based techniques, are employed to discern patterns, entities, relationships, and other relevant data within the document. The effectiveness of this stage is pivotal in deriving meaningful insights and structured knowledge from unstructured textual content. Our review showed the dominance of Named Entity Recognition (NER) and Relationship Extraction (RE) over other tasks such as Knowledge Extraction (KE), Event Extraction (EE) and Causality Extraction (CE).

PDF document preprocessing and EIE are intrinsically linked, as the quality of IE is greatly dependent on the quality of the preceding PDF document preprocessing. Together, they represent a comprehensive and systematic approach to harnessing valuable information from PDF documents, a format that continues to play a central role in the dissemination and sharing of knowledge and data across various domains.

In this review (summarized in Table 4 in the Annex Section), we identified several common challenges in IE from PDF documents, notably, 1) insufficient datasets, 2) the issue of ambiguity resolution preventing accurate classification of extracted concepts, 3) the accuracy of PDF documents conversion to usable text format, which is almost never 100% especially for scanned documents, 4) document heterogeneity, referring to a document incorporating many languages, 5) the execution time of the IE process and 6) the model training time for neural-network and statistical machine learning based models.

5 Discussion

This study sought to explore the motivations for the automatic extraction of information from PDF documents and to track its evolution from 2017 to 2023. This exploration involved identifying the methods employed, associated challenges, and existing gaps in literature. Our findings revealed significant emphasis on IE across various domains during the study period considered. According to our review, the motivations for IE from PDF documents can be categorized into four key clusters: 1) Time optimization in critical tasks, such as medical records analytics (Chen et al., 2022; Meystre et al., 2017), 2) Specific IE from large documents volumes (Papadopoulos et al., 2020) or automatic report analysis (Qiu et al., 2020), 3) Knowledge discovery to help decision making and 4) Building a structured databases for targeted information retrieval and analytics. We observed a shift in methodologies over time, with early studies favoring rules-based approaches and recent studies increasing adoption of automatic training approaches, capitalizing on pretrained Language Models (LMs) based on the transformer's architecture (Vaswani et al., 2017). This gradual shift is attributed to the ease of adoption and adaptability (Bai et al., 2022; Cabot and Navigli, 2021; Jehangir et al., 2023; Kabir et al., 2023;

Kalyan, 2024; Nundloll et al., 2022) offered by automatics learning approaches compared to rule-based methods, which are experts dependent, domain-specific, and less flexible (Turner et al., 2022).

However, both approaches face challenges rooted in their underlying philosophies. Rule-based methods, while offering greater confidence in outputs, are complex, time-consuming to develop and lack adaptability (Banerjee et al., 2019; Reátegui and Ratté, 2018; Sonntag and Profitlich, 2019). In contrast, automatic learning-based approaches suffer from a shortage of well-annotated training data, yet they demonstrate greater flexibility and adaptability to various domains. Ambiguity in understanding natural language remains a common challenge for both approaches, and language-specific solutions persist (Le et al., 2023; Nundloll et al., 2022; Papadopoulos et al., 2020). Efforts to reduce data requirements for training language models have been explored, including fine-tuning pretrained models (Bai, Wang, and Zhang, 2022), nevertheless, the need for high-quality domain-specific datasets and human validation of data quality remains paramount (Nundloll et al., 2022; Papadopoulos et al., 2020).

Despite the popularity of PDF documents for information storage, their diverse content elements, including text, images, tables, and forms, pose significant challenges for accurate information extraction. While OCR-based approaches have been prominent, especially for scanned documents, their processing speed may be a bottleneck, particularly when dealing with large document volumes. Enhancing OCR engines to approach the processing speed of direct text extraction-based methods could be a valuable advancement in information extraction.

In terms of data storage, most of the examined studies organized the extracted information into fact triples, comprising subject, predicate, and object. These set of triples are usually loaded in a knowledge graph (Chen et al., 2022; Sung et al., 2022), for easy visualization or graph database for easy querying (Nundloll et al., 2022), where it could be stored using semantic web technologies standard formats such as RFD-XML or OWL. Extracted data could also be stored using JSON, plain-text, XML (Sonntag and Profitlich, 2019). The choice of storage approach depends on the intended downstream processes. We believe knowledge bases and graphs are favoured for their support of automatic reasoning, contributing significantly to knowledge discovery, a primary objective of automatic information extraction.

IE offers significant potential for enhancing data discovery across various domains. However, existing solutions often exhibit domain-specificality and limited adaptability. To address these challenges, we propose an integrated framework that leverages the strengths of both rule-based and automatic learning-based approaches (see Figure 7). This hybrid approach aims to reduce the reliance on extensive training datasets. Furthermore, we advocate for the integration of language models and common ontologies (C.B. and Mahesh, 2023), facilitating cross-domain adaptability and mitigating the need for large training datasets. Our envisioned framework comprises eight modules:

1. *Document manager*: Enables users to build a document database by uploading files or querying online libraries like PubMed, Web of Science and Google Scholar.
2. *Document pre-processor*: Utilizes NLP algorithms and OCR engines to convert documents in the database into a more usable format, such as text, images, and CSV files.
3. *Ontology manager:* Allows users to configure ontologies for the IE system, facilitating the alignment of IE with the question answering, and document annotation.
4. *Information extractor*: The core module combines rule-based and language model-driven approaches, potentially utilizing models like BERT or REBEL.
5. *Annotation engine:* Enhances the accuracy of the Information Extractor by enabling users to create custom annotation datasets based on common ontology entities and relationships.

6. *Question answering (QA) tool*: Provides a user-friendly interface for stating natural language questions, offering automatic term recommendations for precise data retrieval from the structured data.
7. *Knowledge visualizer:* Facilitates exploration of the structured data through a knowledge graph representation.
8. *Data exporter:* Presents QA results in tabular form and allows users to download data aligned with the underlying domain ontology vocabulary.

In essence, our framework combines document annotation and language models with ontologies as a bridge to enhance domain adaptability and customization. We believe this approach holds significant potential for improving IE system performance. Recent advancements in NLP, particularly the integration of language models and common ontologies, offer promising avenues to enhance the accuracy and efficiency of IE systems across various domains.

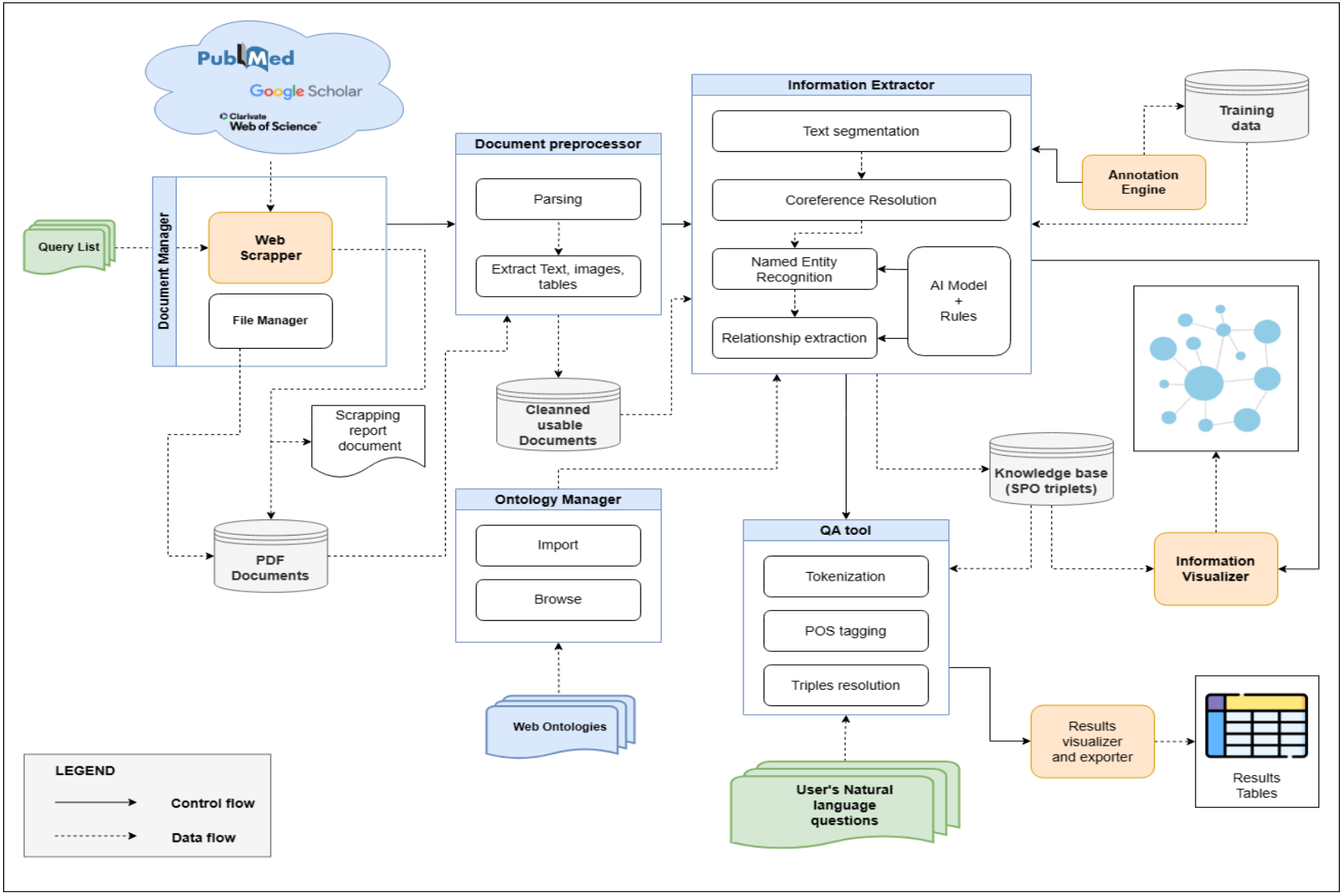

Figure 7: *An overview of the proposed data extraction methodology combining common ontologies and language models to improve the flexibility of Information extraction system.*

6 Conclusion

Information Extraction has garnered significant attention due to the prevalence of unstructured data in natural language text. While impressive solutions have been developed across various domains, several challenges persist in achieving highly reliable IE systems, particularly when extracting information from complex PDF documents. These challenges include the intricate and time-intensive nature of building rule-based systems, the scarcity of well-annotated datasets for automatic learning approaches, and the complexity of handling text ambiguity. To address these challenges and promote adaptability across domains, we have introduced a conceptual hybrid framework that integrates the advantages of these two main categories, with a focus on leveraging common ontologies. Our future endeavours will involve implementing and evaluating the proposed framework.

7 Annex

Table 4: *Presents previous studies in information extraction (IE) from 2017 to 2023. It is a consolidated overview of the landscape of IE research during the specified timeframe. It serves as a reference point for researchers, enabling them to navigate the extensive body of literature, identify relevant studies, and gain insights into the evolution of IE approaches and techniques over the years.*

Author	Domain	Objectives	Method	Class	Storage	Evaluation	Challenges
(Beck-Fernandez et al., 2017)	Social network data-analysis	Automatically extract and compare memes from online forums	Concepts extraction Identification of relations Ontological& structural similarity (WordNet)	rule-based	Not specified	Precision Recall F1-score	The number of concepts and relations extracted is corpus dependent
(Meystre et al., 2017)	Biomedicine	describe a new CHF16 treatment performance measure information extraction system	rules-based approach based on UIMA framework, text pre-processing based on OpenNLP, rule-based patient-level classification	rule-based	Not specified	Precision Recall F1 measure	limited sample size, Insufficient document annotation quality
(Afzal et al., 2017)	Biomedicine	Biomedical information extraction for EHRS documents	classification and regression tree (CART), based prediction model for survival analysis	rule-based	Not specified	Precision Recall F1-score	not all types of documents were handled, lexicon incompleteness
(Reátegui and Ratté, 2018)	Health	Comparison of 2 biomedical knowledge extraction systems (MetaMap, cTAKES)	Test the 2 systems on the same dataset (i2b2 obesity challenge dataset); Compute and compare metrics (recall, precision, F1 score	rule-based	Not specified	Precision Recall F1 score	Not specified
(Rajbabu et al., 2018)	Industry	Industrial information extraction from text documents using ontologies	Sentence classification, Word classification	Statistical Machine learning	Not specified	Precision Recall	process execution time, insufficient training data

Author	Domain	Objectives	Method	Class	Storage	Evaluation	Challenges
(Wang et al., 2018)	Geoscience	build a workflow for information extraction and knowledge discovery from textual geoscience data in Chinese	Chinese text segmentation using CRF; Term frequency computation; Knowledge graph construction using chord and bigrams	Statistical Machine Learning	Knowledge graphs	Precision Recall F1-score	Documents heterogeneity (mix Chinese English)
(Banerjee et al., 2019)	Defense & Security	Retrieve information from a huge amount of textual data via natural language queries	NLTK library	rule-based	Tree structure and factoïds	Accuracy Precision	Dependence vis a vis Stanford POS tagger and ne_chunker, not 100% reliable on annotations. Incorrect judgment of verb tenses. Unsupported non-numeric date formats
(Sonntag and Profitlich, 2019)	Health	build an integrated faceted search tool, accompanied by the visualization of results of automatic information extraction from textual documents	Customization of an existing pipeline (UIMA) specialized on breast cancer and Apache SOLR	rule-based	XML, JSON	No standard measure used, study focused on experimentation or use-case evaluation approach	Insufficient user control over the information extraction process. Poor quality when many patient files (> 100, 000) are processed.
(Abulaish et al., 2019)	Biomedicine	identify and extract disease symptoms and their associations from biomedical text documents retrieved from the PubMed database	Document Crawler; Document Pre-processing; Dependency Processing; Abbreviation Extraction; Disease Symptoms Minin; Feasibility Analysis, to filter realistic symptoms; InformationVisualizer	rule-based	knowledge graphs	Precision Recall F1-score	Insufficiency of the symptoms captured by MetaMap. Ambiguity in symptoms classification
(Kang et al., 2019)	Manufacturing	Information extraction from text documents via an ambiguity resolution method that utilizes domain ontology as the mechanism to incorporate the domain context	Extract manufacturing concepts and relations; Identify unresolved ambiguities; Resolve ambiguities considering domain context (ontology)	rule-based	Not specified	No standard measure used, instead, experimentation shows that with ontology-based ambiguity resolution IE result are more accurate	limited to manufacturing domain and ontology dependent

Author	Domain	Objectives	Method	Class	Storage	Evaluation	Challenges
(Papadopoulos et al., 2020)	Open Information Extraction	Extract information from scientific corpora, eliminating the ambiguity and, redundancy of SPO (Subject Predicate Object) triples from OIE (Open Information Extraction).	In-place coreference resolution (Gardner et al., 2018) ; Extractive text summarization (Beltagy et al., 2019b); Parralel triple extraction (Gardner et al., 2018); Entity enrichment &graph representation (Neumann et al., 2019)	Neural Network based	Knowledge graphs	Precision Recall F1-score	Sensitivity: the approach doesn't measure to what extent valid triples are overlooked.
(Qiu et al., 2020)	Geoscience	Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports	Document processing (sentence splitting, tokenization, part-of-speech tagging,); Creation of gazetteers (spatial, temporal); Contextual information extraction (Cunningham et al., 2000); spatiotemporal RE; Spatiotemporal topic relevance network model	rule-based	Bi-grams network	Precision Recall F1-score	The ontologies used aren't wide enough to capture all domain terms, but their development could improve system accuracy
(Abdollahi et al., 2021)	Biomedicine	Extract meaningful information from medical discharge notes, to perform document classification	Sentence tokenization of notes. Terms alignment using MetaMap and UMLS. Feature extraction form documents. Classification using recurrent neural networks.	Rule-based + Neural Network based	Not specified	Accuracy F1-measure	Not specified
(Fei et al., 2021)	Biomedicine	Contextualized language models for biomedical information extraction	pre-training, finetuning, post-training	Neural network based	Not specified	Precision Recall F1-Score	Training time

Author	Domain	Objectives	Method	Class	Storage	Evaluation	Challenges
(Nundloll et al., 2022)	Ecology & conservation science	Automate the extraction of floristic information from the Journal of Botanic, an important historical archive on the biodiversity of Lake District in the UK	Not Specified but Spacy NLP framework and Prodigy annotation tools used	Neural Network based	No SQL database MongoDB, GraphDB	Accuracy	Ambiguity between plant species names and Author names. Errors during the conversion of scanned document to text affects model accuracy
(Chen et al., 2022)	Biomedicine	Improve context awareness in biomedical relation extraction from text documents using machine Reading Comprehension	BERT encoder layer. Knowledge enhanced attention layer. Prediction layer	Neural Network based	Not specified	Precision Recall F1-score	Not specified
(Zhu and Cole, 2022)	Chemistry	Accurately extract and annotates information from scientific PDF documents	Read PDF create text blocks sing PDFMiner. Text block features assignment. Automatic publisher template assignment. Apply template-defined rules and grammar. Metadata extraction. Section extraction. Reference extraction.	rule-based	JSON & plain text	Precision Recall F1-score	cannot work accurately on scanned PDF
(Guan et al., 2022)	Welding Manufacturing	enrich the relationship database of the welding manufacturing domain, by extracting them from domain documents	BiLSTMC Attention, CR-CNN models	Neural Network based	Knowledge graph	Precision Recall F1-score	Not specified
(Gupta et al., 2022)	Material science	Build a more accurate model for information extraction form scientific publication on material science	MatSciBERT an extension on sciBERT through finetuning	Neural network based	Not specified	Micro-F1 Macro-F1	Not specified

Author	Domain	Objectives	Method	Class	Storage	Evaluation	Challenges
(Erdengasileng et al., 2022)	Biomedical sciences	Evaluate different approach of IE in biomedical field	Pretrained NLP models evaluation (BERT, BioBERT etc.), Data augmentation evaluation, Ensemble modeling evaluation.	Neural network based	Not specified	F1-score	Not specified
(Yan et al., 2022)	Chemistry and Materials science	Automatic extraction of organic and inorganic chemical substance from text	ON-LSTM used. instead – BERT envisaged as a future perspective	Neural Network based	Knowledge graph	F1-score	Data insufficiency
(Turner et al., 2022)	Healthcare	extraction of outcome measures from the EHR5 documents in psychiatry	rule-based approach implemented using spaCy + Prodigy	rule-based	Not specified	Accuracy F1-score	Low flexibility of the rule-based model
(Tang et al., 2022)	Civil engineering	identification of construction activities, material, building component, product features, measurement unit, and additional information from work descriptions	Hidden Markov Model (HMM)	Statistical Machine Learning	Not specified	F1-score	Data insufficiency for training and testing

8 References

Abdillah, Leon Andretti. (PDF) PDF articles metadata harvester 2013. https://www.researchgate.net/publication/235326466_PDF_articles_metadata_harvester (accessed June 9, 2023).

Abdollahi M, Gao X, Mei Y, Ghosh S, Li J, Narag M. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. Artif Intell Med 2021;120. https://doi.org/10.1016/J.ARTMED.2021.102167.

Abdullah MHA, Aziz N, Abdulkadir SJ, Alhussian HSA, Talpur N. Systematic Literature Review of Information Extraction From Textual Data: Recent Methods, Applications, Trends, and Challenges. IEEE Access 2023;11:10535–62. https://doi.org/10.1109/ACCESS.2023.3240898.

Abulaish M, Parwez MA, Jahiruddin. DiseaSE: A biomedical text analytics system for disease symptom extraction and characterization. J Biomed Inform 2019;100:103324. https://doi.org/10.1016/J.JBI.2019.103324.

Afzal M, Hussain M, Khan WA, Ali T, Jamshed A, Lee S. Smart Extraction and Analysis System for Clinical Research. Telemed J E Health 2017;23:404–20. https://doi.org/10.1089/TMJ.2016.0157.

Bacchi S, Gluck S, Koblar S, Jannes J, Kleinig T. Automated information extraction from free-text medical documents for stroke key performance indicators: a pilot study. Intern Med J 2022;52:315–7. https://doi.org/10.1111/IMJ.15678.

Bach N, Badaskar S. A Review of Relation Extraction 2007.

Bai W, Wang J, Zhang X. YNU-HPCC at SemEval-2022 Task 4: Finetuning Pretrained Language Models for Patronizing and Condescending Language Detection. SemEval 2022 - 16th International Workshop on Semantic Evaluation, Proceedings of the Workshop 2022:454–8. https://doi.org/10.18653/V1/2022.SEMEVAL-1.61.

Banerjee PS, Chakraborty B, Tripathi D, Gupta H, Kumar SS. A Information Retrieval Based on Question and Answering and NER for Unstructured Information Without Using SQL. Wirel Pers Commun 2019;108:1909–31. https://doi.org/10.1007/S11277-019-06501-Z/METRICS.

Beck-Fernandez H, Nettleton DF, Recalde L, Saez-Trumper D, Barahona-Peñaranda A. A system for extracting and comparing memes in online forums. Expert Syst Appl 2017;82:231–51. https://doi.org/10.1016/J.ESWA.2017.04.010.

Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference 2019a:3615–20. https://doi.org/10.18653/V1/D19-1371.

Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference 2019b:3615–20. https://doi.org/10.18653/v1/d19-1371.

Bird S, Klein E, Loper E. LIVRO: cookbook Natural Language Processing with Python. J Endod 2009;28:330–2.

Cabot PLH, Navigli R. REBEL: Relation Extraction By End-to-end Language generation. Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021 2021:2370–81. https://doi.org/10.18653/V1/2021.FINDINGS-EMNLP.204.

C.B. A, Mahesh K. Ontology-based data interestingness: A state-of-the-art review. Natural Language Processing Journal 2023;4:100021. https://doi.org/10.1016/J.NLP.2023.100021.

Chen J, Hu B, Peng W, Chen Q, Tang B. Biomedical relation extraction via knowledge-enhanced reading comprehension. BMC Bioinformatics 2022;23:1–19. https://doi.org/10.1186/S12859-021-04534-5/FIGURES/4.

Chowdhary KR. Natural Language Processing. Fundamentals of Artificial Intelligence 2020:603–49. https://doi.org/10.1007/978-81-322-3972-7_19.

Cunningham H, Maynard D, Tablan V. JAPE: a Java Annotation Patterns Engine 2000.

Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2018;1:4171–86.

Erdengasileng A, Han Q, Zhao T, Tian S, Sui X, Li K, et al. Pre-trained models, data augmentation, and ensemble learning for biomedical information extraction and document classification. Database 2022;2022. https://doi.org/10.1093/DATABASE/BAAC066.

Fei H, Ren Y, Zhang Y, Ji D, Liang X. Enriching contextualized language model from knowledge graph for biomedical information extraction. Brief Bioinform 2021;22. https://doi.org/10.1093/BIB/BBAA110.

Gardner M, Grus J, Neumann M, Tafjord O, Dasigi P, Liu NF, et al. AllenNLP: A Deep Semantic Natural Language Processing Platform 2018:1–6. https://doi.org/10.18653/V1/W18-2501.

Guan K, Du L, Yang X. Relationship Extraction and Processing for Knowledge Graph of Welding Manufacturing. IEEE Access 2022;10:103089–98. https://doi.org/10.1109/ACCESS.2022.3209066.

Gupta P, Gupta V. A Survey of Text Question Answering Techniques. Int J Comput Appl 2012;53:1–8. https://doi.org/10.5120/8406-2030.

Gupta T, Zaki M, Krishnan NMA, Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. Npj Computational Materials 2022 8:1 2022;8:1–11. https://doi.org/10.1038/s41524-022-00784-w.

Hogenboom F, Frasincar F, Kaymak U, Jong FD. An Overview of Event Extraction from Text 2011.

Jehangir B, Radhakrishnan S, Agarwal R. A survey on Named Entity Recognition-datasets, tools, and methodologies. Natural Language Processing Journal 2023;3:100017. https://doi.org/10.1016/j.nlp.2023.100017.

Jin X, Wang X, Luo X, Huang S, Gu S. Inter-sentence and Implicit Causality Extraction from Chinese Corpus. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2020;12084 LNAI:739–51. https://doi.org/10.1007/978-3-030-47426-3_57/FIGURES/3.

Kabir MdA, Phillips T, Luo X, Al Hasan M. ASPER: Attention-based approach to extract syntactic patterns denoting semantic relations in sentential context. Natural Language Processing Journal 2023;3:100011. https://doi.org/10.1016/J.NLP.2023.100011.

Kalyan KS. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal 2024;6:100048. https://doi.org/10.1016/j.nlp.2023.100048.

Kang S, Patil L, Rangarajan A, Moitra A, Robinson D, Jia T, et al. Ontology-Based Ambiguity Resolution of Manufacturing Text for Formal Rule Extraction. J Comput Inf Sci Eng 2019;19. https://doi.org/10.1115/1.4042104/422093.

Le L, Demartini G, Zuccon G, Zhao G, Zhang X. Active learning with feature matching for clinical named entity recognition. Natural Language Processing Journal 2023;4:100015. https://doi.org/10.1016/J.NLP.2023.100015.

Lee J, Yoon W, Kim Sungdong, Kim D, Kim Sunkyu, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36:1234–40. https://doi.org/10.1093/bioinformatics/btz682.

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of the Annual Meeting of the Association for Computational Linguistics 2020:7871–80. https://doi.org/10.18653/V1/2020.ACL-MAIN.703.

Li Z, Li Q, Zou X, Ren J. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. Neurocomputing 2021;423:207–19. https://doi.org/10.1016/J.NEUCOM.2020.08.078.

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. Ann Intern Med 2009;151. https://doi.org/10.7326/0003-4819-151-4-200908180-00136.

Mannai M, Karâa WBA, Ghezala HH Ben. Information extraction approaches: A survey. Advances in Intelligent Systems and Computing 2018;625:289–97. https://doi.org/10.1007/978-981-10-5508-9_28/COVER.

Di Martino B, Marulli F, Graziano M, Lupi P. PrettyTags: An Open-Source Tool for Easy and Customizable Textual MultiLevel Semantic Annotations. Lecture Notes in Networks and Systems 2021;278:636–45. https://doi.org/10.1007/978-3-030-79725-6_64/COVER.

Meystre SM, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. J Am Med Inform Assoc 2017;24:e40–6. https://doi.org/10.1093/JAMIA/OCW097.

Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. BioNLP 2019 - SIGBioMed Workshop on Biomedical Natural Language Processing, Proceedings of the 18th BioNLP Workshop and Shared Task 2019:319–27. https://doi.org/10.18653/v1/W19-5034.

Neumann M, Shen Z, Skjonsberg S. PAWLS: PDF Annotation With Labels and Structure. ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the System Demonstrations 2021:258–64. https://doi.org/10.18653/v1/2021.acl-demo.31.

Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Brief Bioinform 2021;22:146–63. https://doi.org/10.1093/BIB/BBZ130.

Newe A, Becker L. Three-Dimensional Portable Document Format (3D PDF) in Clinical Communication and Biomedical Sciences: Systematic Review of Applications, Tools, and Protocols. JMIR Med Inform 2018;6. https://doi.org/10.2196/10295.

Nganji JT. The Portable Document Format (PDF) accessibility practice of four journal publishers. Libr Inf Sci Res 2015;37:254–62. https://doi.org/10.1016/J.LISR.2015.02.002.

Nundloll V, Smail R, Stevens C, Blair G. Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. Heliyon 2022;8:e10710. https://doi.org/10.1016/J.HELIYON.2022.E10710.

Papadopoulos D, Papadakis N, Litke A. A Methodology for Open Information Extraction and Representation from Large Scientific Corpora: The CORD-19 Data Exploration Use Case. Applied Sciences 2020, Vol 10, Page 5630 2020;10:5630. https://doi.org/10.3390/APP10165630.

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2018;1:2227–37. https://doi.org/10.18653/V1/N18-1202.

Qiu Q, Xie Z, Wu L, Tao L. Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. Earth Sci Inform 2020;13:1393–410. https://doi.org/10.1007/S12145-020-00527-9/METRICS.

Rajbabu K, Srinivas H, Sudha S. Industrial information extraction through multi-phase classification using ontology for unstructured documents. Comput Ind 2018;100:137–47. https://doi.org/10.1016/J.COMPIND.2018.04.007.

Reátegui R, Ratté S. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. BMC Med Inform Decis Mak 2018;18:13–9. https://doi.org/10.1186/S12911-018-0654-2/TABLES/3.

Sonntag D, Profitlich HJ. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. Artif Intell Med 2019;93:13–28. https://doi.org/10.1016/J.ARTMED.2018.08.003.

Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics 2022;38:4837–9. https://doi.org/10.1093/bioinformatics/btac598.

Tang S, Liu H, Almatared M, Abudayyeh O, Lei Z, Fong A. Towards Automated Construction Quantity Take-Off: An Integrated Approach to Information Extraction from Work Descriptions. Buildings 2022, Vol 12, Page 354 2022;12:354. https://doi.org/10.3390/BUILDINGS12030354.

Turner RJ, Coenen F, Roelofs F, Hagoort K, Härmä A, Grünwald PD, et al. Information extraction from free text for aiding transdiagnostic psychiatry: constructing NLP pipelines tailored to clinicians' needs. BMC Psychiatry 2022;22:1–11. https://doi.org/10.1186/S12888-022-04058-Z/TABLES/3.

Varshini KS, Uthra RA. Extraction of Meaningful Information from Unstructured Clinical Notes Using Web Scraping. Https://DoiOrg/101142/S021812662350041X 2022. https://doi.org/10.1142/S021812662350041X.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Adv Neural Inf Process Syst 2017;2017-December:5999–6009.

Wang C, Ma X, Chen Jianguo, Chen Jingwen. Information extraction and knowledge graph construction from geoscience literature. Comput Geosci 2018;112:112–20. https://doi.org/10.1016/J.CAGEO.2017.12.007.

Webster JJ, Kit C. TOKENIZATION AS THE INITIAL PHASE IN NLP. Actes de COLING-92 Nantes 1992.

Yan R, Jiang X, Wang W, Dang D, Su Y. Materials information extraction via automatically generated corpus. Scientific Data 2022 9:1 2022;9:1–12. https://doi.org/10.1038/s41597-022-01492-2.

Yang J, Han SC, Poon J. A survey on extraction of causal relations from natural language text. Knowl Inf Syst 2022;64:1161–86. https://doi.org/10.1007/s10115-022-01665-w.

Zhu M, Cole JM. PDFDataExtractor: A Tool for Reading Scientific Text and Interpreting Metadata from the Typeset Literature in the Portable Document Format. J Chem Inf Model 2022;62:1633–43. https://doi.org/10.1021/ACS.JCIM.1C01198/SUPPL_FILE/CI1C01198_SI_001.ZIP.