Multimedia Tools and Applications (2024) 83:21229–21280
https://doi.org/10.1007/s11042-023-15080-y

Extracting and structuring information from the electronic medical text: state of the art and trendy directions

Mohamed Yassine Landolsi[1] · **Lobna Hlaoua**[1] · **Lotfi Ben Romdhane**[1]

Received: 18 July 2022 / Revised: 19 September 2022 / Accepted: 2 March 2023 /
Published online: 1 August 2023

Abstract

In the medical field, doctors must have comprehensive knowledge by reading and writing narrative documents, and they are responsible for every decision they take for patients. Unfortunately, reading all the necessary information about drugs, diseases, and patients might be time-consuming due to the large number of documents that are increasing every day. Consequently, potential medical errors could be hazardous. Likewise, information extraction can handle this problem using several important tasks to structure the text and extract the relevant and desired information from unstructured text written in natural language. The main principle tasks are named entity recognition and relation extraction. However, to treat the narrative text, we should use natural language processing techniques to extract useful information and features. In our paper, we show and discuss several techniques and useful data used for these tasks. Furthermore, we outline the challenges in information extraction from medical documents. To our knowledge, this is the most comprehensive survey in the literature with a numerical comparison and a suggestion for some uncovered directions.

Keywords Electronic medical records · Information extraction · Medical named entities recognition · Medical relation extraction · Section detection

Lobna Hlaoua and Lotfi Ben Romdhane are contributed equally to this work.

✉ Mohamed Yassine Landolsi
medyassine.landolsi@isitc.u-sousse.tn

Lobna Hlaoua
lobna.hlaoua@essths.u-sousse.tn

Lotfi Ben Romdhane
lotfi.benromdhane@isitc.u-sousse.tn

[1] MARS Research Laboratory, SDM Research Group, ISITCom H-Sousse, University of Sousse, Hammam Sousse, 4011, Tunisia

1 Introduction

For centuries, physicians play an important role in ensuring good health. Indeed, a physician must be well trained and must be able to manage the disease and patient information to find the right treatment and make the right decision. In medicine, different types of information are used for treatment and can be found in narrative documents written by humans. For example, to make a medical prescription, the patient record and medication documents are used [94]. Thanks to the development of information technologies and the Hospital Information System (HIS), medical information is digitized into electronic records named Electronic Medical Record (EMR) [121] or Electronic Health Record (EHR). Digital records could be stored, managed, transmitted, and reproduced efficiently. The widespread adoption of HIS has contributed to billions of records, and they are recognized as valuable resources for large-scale analysis [92].

Nevertheless, the physicians find a big difficulty to manage the large amount of data and read a lot of natural language text to select the important information. In fact, there is many errors happened due to missing the right information and the drug prescription is a main source of them. For the last 18 years, there is more than 237 million errors related to the drug prescription and even some of them caused deaths or costs to the country [93, 94]. In addition, understanding a new disease is a big challenge since it requires a lot of time and information studies to find the right treatment. For example, the coronavirus disease (COVID-19), which is appeared in 2019, caused more than 234 million cases and 4 million deaths in 2021 where there is no cure was discovered at that time [145].

All these facts make the use of the available electronic documents more than a necessity. Hence, extracting useful information will be greatly helpful [150]. Generally, medical text mining and Information Extraction (IE) help in medical decisions and disease risk prediction [63, 88, 164]. Unfortunately, this process is not trivial mainly due to the huge number of documents and consequently requires models able to deal with big data. This is hampered by the unstructured (or semi-structured) nature of such documents [149]. To make IE feasible and efficient, structuring these documents in a more abstract form that is easily readable by machines/algorithms becomes a fundamental step. The main technologies of IE are Named Entity Recogntion (NER) and entity Relation Extraction (IE) [134]. Entity recognition involves recognizing references to different types of entities such as medical problems, tests, allergies, risks, adverse events, drugs, and treatments. In addition, detecting different sections in a document can improve IE tasks by providing more context [110].

There are many real world applications of IE which can be summarized into 3 categories [150]: disease study areas, drug-related study areas, and clinical workflow optimization. Considering the disease study areas, a large portion of studies have focused on IE of diseases and conditions [39, 150]. Mehrabi et al. [100] identify patients with a family history of pancreatic cancer. Carrell et al [18] process clinical notes for women with early-stage breast cancer to identify whether recurrences were diagnosed and the timing of the diagnoses. For drug-related studies, Zheng et al. [166] extract mentions of aspirin use and dosage information from clinical notes. Lupşe and Stoicu-Tivadar [94] check for drug incompatibility with other drugs, diseases, or other characteristics. Wang et al. [148] use unstructured data in EHR to identify an adverse drug reaction. Wei et al. [151] determine patient drug exposure histories. For clinical workflow optimization areas, Rochefort et al. [119] detect adverse events of hospital-acquired pneumonias, central venous catheter-associated bloodstream infections, and in-hospital falls. Hsu et al. [55] conduct quality assessment of radiologic interpretations. Popejoy et al. [111] identify and extract care coordination activities from nursing note and show how they can be quantified in order to support patient management. Garvin et al. [41] extract left ventricular ejection fraction value, which is a heart failure quality measure, from echocardiogram reports.

One of the main challenges in this field is the reasonable selection of processing tools, where a designed method may have poor performance in the medical field due to the complexity of its variable and ambiguous natural language text [134]. There is also the lack of sufficient annotated public datasets, especially for languages other than English, where English text studies are more mature and systematic [163]. Another challenge is the lack of knowledge base dictionaries, where the quality of dictionaries requires increased evaluation and certification of specialized institutions [134]. Also, the patient privacy protection in medical documents by automatically and precisely de-identifying personal and sensitive information is required since the transmission of EMR records is becoming more important and should be in legal way [101]. Finally, a larger scale and more complex structure of medical documents make information extraction harder to process but more beneficial [134].

Several reviews related to IE in the medical field have been published. Meystre et al. [102] have focused on research about IE from clinical narrative from 1995 to 2008. However, they have not discussed the research on IE from biomedical literature. Liu et al. [86] have introduced a novel IE paradigm Open IE (OpenIE) which begins to attract great attention in Biomedical IE (BioIE). Their review focuses mainly on recent advances in learning-based approaches such as Convolutional Neural Network (CNN) [22, 124] and Recurrent Neural Network (RNN) [128]. However, they have focused only on the main deep learning techniques in Bio-OpenIE. Wang et al. [150] have presented research on clinical IE applications from 2009 to 2016 for a discussion in terms of publication venues, data sources, clinical IE tools, methods, applications in different areas. Also, they have gained a more concrete understanding of the underlying reasons for the gap between clinical studies using EHR and studies using clinical IE. However, their research is made before deep learning was adopted as mainstream in the informatics community. Sun et al. [134] have focused on the process of medical document processing and analyzes the key techniques involved including NER and RE for IE. They make an in-depth study on the text mining applications, the open challenges, and research issues for future work. Pomares-Quimbaya et al. [110] have reported the results of a systematic review concerning section identification in narrative medical documents. It was the first understanding review that focuses on this concept, its existing methods, and its growing contribution to the IE tasks. Koleck et al. [65] have summarized the use of natural language processing to analyze symptom information in EHR to indicate diseases. They have found little coverage of deep learning approaches in this application area. Hahn and Oleynik [47] have discussed the contributions of recent publications from 2017 to 2020 in medical IE and foreshadow future directions of research. They have focused on the methodological paradigm shift from standard machine learning techniques to deep learning. Also, they have selected only two of the diseases and drugs' semantic classes and the relation between them. Nasar et al. [105] have focused specifically on NER and RE with a major focus on advances via deep learning approaches. They have presented recent trends in the domain of IE along with open research areas.

The main objective of this paper is to give an up-to-date survey about IE in the medical field which can support disease and drug studies and optimize clinical workflow. In particular, we will make a comprehensive review of recent models for section identification, NER, and RE. Thus, we will focus on both semantic and syntactical structuring by highlighting the importance of section detection to the medical IE. Likewise, we will make a numerical comparison between existing approaches based on the published results. The main goal of this numerical study is to get a high-level comparison between existing approaches and to see the directions in which they succeed (eventually fail). We will cover different types of methods and highlight some ideas which may lead to new and promising ways to develop this field. We will review and discuss the different difficulties of the information nature and solutions used to handle them. In addition, we will discuss the useful resources and datasets used in many published studies since they are a reliable basis for the models' efficiency. We find that recent studies focus on the use of advanced deep learning models and propose solutions to better exploit the context during the analysis. Recently, the graph-based technique appeared to be more suitable especially to deal with the named entities. There are many directions that can be taken such as: further explore the graph based techniques, exploit some information in the document such as the formatting style, easily adapt to new data and provide sufficient high-quality resources, use the syntactical structure of the document, etc.

The rest of this paper is structured as follows. Our paper is organized as follows: Section 2 represents the method used to collect and select papers for our survey; Section 3 is a discussion about the medical data and its nature and problems; Section 4 presents useful data and benchmark datasets in IE; Section 5 represents important metrics used for the IE task; Section 6 classifies and discusses the methods of IE tasks according to the techniques they use and it includes a numerical comparison for each task; Section 7 represents problems related to the nature of named entities and discusses the existing solutions; Section 8 represents the important limitations and challenges in the medical information extraction; and in the Conclusion, we suggest some research directions be further developed.

2 Method

In our survey, we have collected and selected medical information extraction papers to be analyzed and compared. We cover 3 sub-fields of the medical information extraction field: Named Entity Recognition (NER), Relation Extraction (RE), and Section Detection (SD). For that, we have used some search engines such as Google Scholar, PubMed, and Scopus in order to search for a collection of interesting papers for each topic in addition to papers that cover the general field. The queries used to perform this search are generally chosen based on combinations of 3 types of keywords: medical data type, task name, and method.

- **Data type:** we have used keywords such as "medical", "clinical", "biomedical", "healthcare" and "medical" which are combined with the object name such as "text", "data", "EMR", "EHR", "corpus", "information", "unstructured data", "document", "narrative", "record", "report", "tweet", and "notes". For example, we can obtain "medical text", "clinical record", etc.
- **Task name:** these keywords are divided into two parts: the verb and the object. For the verb, we have used words like "recognition", "extraction", "identifying", "detection", "annotation", "linking", "mining", "processing" and "analysis". For the object, we have used "named entity", "entity", "section", "subsection", "heading", "title", "relation", "terms", "concept", "data" and "natural language". For example, we can obtain a combination like "natural language processing", "named entity recognition", "identifying heading", etc.
- **Method:** for some queries, we add keywords to target a specific technology or feature such as "machine learning", "deep learning", "unsupervised", "automatic", "online", "ontology", "rules", "formatting style", "dictionary", "boundary detection", "part of

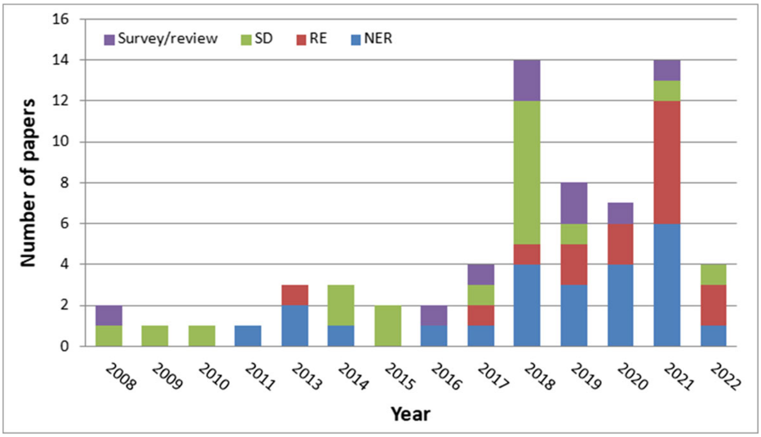

Fig. 1 Statistics about the selected papers for our survey

speech", "fuzzy matching", "contextual embedding", "document layout", "font size", etc.

Thus, we can obtain queries like "medical text natural language processing", "clinical document unsupervised identifying section", "medical document named entity recognition contextual embedding", "healthcare tweet entity linking rules", and "unsupervised medical entity recognition linking online medical text", etc. We have executed queries with different combinations of keywords to search for relevant papers which cover many aspects of this field. After executing multiple queries, we have selected papers that are considered to be important, different, or recent in order to take diversified papers covering all the important aspects of the field and recent work in it. As a result, we have obtained 53 methods and 7 survey or review papers which are published between 2008 and 2022. Figure 1 and Table 1 represent statistics about the number of the selected publications for each topic per year. All the selected methods are discussed in our survey according to different criteria.

3 Problem setting

There are different sources of medical information, such as daily activities and clinical staff [22]. Indeed, the rapid development of hospital information technology has resulted in a rapid accumulation of medical data. There a significant amount of this data is in the form of free text written by the author [161]. Note that it is very likely to have long and maybe useless parts in narrative content that are not relevant to the information searched by the reader. A clinical narrative is a report-style free text that is found in the medical document and used for clinical documentation. It is a rich source of information for medical research and analysis, and this source of data is needed to make health care decisions. There are several examples of clinical narratives that vary between full-fledged documents or clinical notes: Discharge Summaries, Emergency Reports, Urology Reports, Letters of Communication, History or Family History, Physical Exam, Medical Dictation, Admission Notes, Nursing Notes, Progress Notes, Operative or Procedure Notes, and Clinic Visit Notes. Thus, there are several sources for these narratives: analytical repositories [83, 90], EMR/EHR

Table 1 Statistics about the selected papers for our survey

year	NER	RE	SD	Survey/review
2008	0	0	1	1
2009	0	0	1	0
2010	0	0	1	0
2011	1	0	0	0
2013	2	1	0	0
2014	1	0	2	0
2015	0	0	2	0
2016	1	0	0	1
2017	1	1	1	1
2018	4	1	7	2
2019	3	2	1	2
2020	4	2	0	1
2021	6	6	1	1
2022	1	2	1	0
Total	24	15	18	8

systems [78, 139], speech recognition or dictation systems [58, 122], external sources provided by competitions [26, 139], or other open data sources [17]. EMR or EHR records are digital representations of medical information which are popularized thanks to the development of information technologies and the HIS. They allow medical staff to record digital information, such as texts, symbols, diagrams, graphs, data, etc. Thus, they allow medical institutions to record the patient's condition, such as diagnostic information, procedures performed, and treatment results. They are therefore sources of clinical information such as demographic data, diagnostic history, medications, laboratory test results, and vital signs. Also, these digital records could be stored, managed, transmitted, and reproduced efficiently. The widespread adoption of HIS has contributed to billions of records [92], and they are recognized as valuable resources for large-scale analysis. It is difficult to discover the hidden knowledge lies in the massive amount of medical texts [161]. Indeed, these unstructured data lack common structural frameworks [135]. The processing complexity of this data is increased by grammatical misuse, misspellings, local dialects, and semantic ambiguities.

Electronic drug prescription helps to control the prescription and monitors the consumption of drugs. To make the appropriate treatment decision, the patient's EHR and information provided by the drug manufacturer are used [94]. The medical prospectus has information to avoid interactions with other medications, diseases, allergies, or the patient's conditions [22]. For example, SmPC is a legal document approved by the European Medicines Agency, used to represent the package insert, and it is also available in electronic form. According to physicians, the most preferred SmPC section titles are: "4.3 Contraindications", "4.1 Therapeutic indications", "4.2 Posology and method of administration", "4.8 Undesirable effects", "5.3 Preclinical safety data", "4.5 Interaction with other medicinal products and other interactions", "5.2 Pharmacokinetic properties", "4.4 Special warnings and precautions of use" and "5.1 Pharmacodynamic properties". Sorted from most to least preferred [48]. Also, the most important sections of the medical prospectus for establishing the appropriate treatment are: Contraindications, Therapeutic indications, and Dosage [94].

Indeed, one of the main sources of errors in medicine is related to the prescription of drugs, where the number of diseases and drugs is gradually increasing, and the doctor must know all the indications and contraindications to prescribing a drug [94]. For this reason, there are several problems especially for new doctors, as it needs a lot of time to read all the unstructured instructions [22, 93, 94]. It is necessary to find out about new diseases that require effective treatments. Also, have to find out what new treatments or drugs the doctors need to access. So, doctors may prescribe the wrong treatments or drugs. In 2006, there were 3900 prescribing errors in Germany [94]. For the last 18 years, the U.S. Institute of Medicine has reported that there are 237 million medication errors per year in England with costs to the country [93]. Also, it reported that there were between 1700 and 22303 deaths per year due to adverse drug reactions. Also, understanding new diseases is a very critical problem in medicine. For example, there are more than 234 million cases and 4 million deaths caused by coronavirus disease (COVID-19) in 2021, as no effective drug for this virus had been discovered at that time [145]. Likewise, the information about symptoms of new diseases needs to be analyzed by efficient tools to inform risk assessment, prevention and treatment strategy development and outcome estimation [145]. Current solutions to avoid these problems are the creation of written procedures, improving training for health care professionals, automation of support or research operations, quality control in medicine, improving communication between physicians, and encouraging cooperation between medical departments [93]. It is a great challenge to effectively use text in medical documents [161]. Automatic document analysis is affected by ambiguity, format diversity, brevity, sloppy writing, redundancy, and complex longitudinal information [153]. Also, obstacles to data mining are diversity, incompleteness, redundancy, and privacy [134].

4 Benchmark datasets and supplemental resources

In this section, we cite the most important benchmark datasets used in the IE field which are manually annotated and considered a gold standard. Most of these datasets are used in shared tasks and used to train and evaluate IE state-of-the-art methods. Table 2 shows some details about these datasets. Generally, the annotation focus on information related to diseases, medicament, and chemical entities. Other datasets are obtained from discharge summaries and clinical reports which are de-identified to hide personal information. Informatics for Integrating Biology & the Bedside (i2b2) or National NLP Clinical Challenges (n2c2) [130, 131, 143] is an open-source clinical data warehousing and analytics research platform, which enables sharing, integration, standardization, and analysis of heterogeneous data from health-care and research. It has an important role in supporting the development of methods for several clinical NLP tasks such as named entity recognition, relation extraction, and so many other tasks. It provides datasets with gold-standard annotations which are used for medical NLP challenges and it has big popularity in the medical IE field. However, most of the available datasets are only aimed at the English language and there is a lack of some other languages [10]. Also, there is a lack of benchmark datasets for section detection tasks where most methods collect and annotate data by themselves [110]. This lack of data is due to the multiplicity of tasks and languages [10, 43], and the sensitive information about the patient's privacy contained in the medical documents [134]. Most of the available annotated datasets are in textual form, especially for the NER and RE tasks. Thus, useful information such as formatting style is not available, which is useful to enrich some tasks with information about the document structure [9, 26, 110]. Many datasets are made from

Table 2 Benchmark datasets used in IE

Dataset	Size	Data type	Information	Statistics	Data source	Annotation method
i2b2 2009 [142] (Shared task)	170 train, 251 test, 679 un-annotated	Discharge summary	Medication-related information: medications, dosages, modes, frequencies, durations, reason for administration.	27589 mentions: –12773 medications –4791 dosages –3552 modes -4342 frequencies –597 duration –1534 reasons	Partners Healthcare	annotation by physician and revision by researcher
i2b2 2010 [143] (Shared task)	394 train, 477 test, 877 un-annotated	Discharge summaries	–Entity types: Medical problem, Treatment and Test. –Assertions: Present, Absent, Possible, Conditional, Hypothetical and Not associated with the patient. –Relations (11): Treatment improves medical problem, Treatment causes medical problem, Test reveals medical problem, etc.	–30518 tests –20268 problems –22060 treatments –14333 relations	Partners Healthcare, Beth Israel Deaconess Medical Center and University of Pittsburgh Medical Center.	partnered with VA Salt Lake City Health Care System.
i2b2 2014 [130, 131] (Shared task)	296 patients: 790 train, 514 test records	longitudinal records for patients	Protected health informations (18): name, geographic, date, phone, e-mail, etc.	28872 mentions	Partners Healthcare	6 annotators: double annotation per patient followed by adjudication.
n2c2 2018 [52] (Shared task)	303 train, 202 test	Discharge summaries	–Entity types: Concepts related to medications, their signature information, and Adverse drug events (allergic reactions, drug interactions, overdoses, and medication errors). – Relation types: Linking concepts to their medication.	–83869 concepts. –59810 relations	MIMIC-III clinical care database.	2 independent annotators while a third annotator resolved conflicts.

Table 2 (continued)

Dataset	Size	Data type	Information	Statistics	Data source	Annotation method
n2c2 2019 [146] (Shared task)	113000 notes: 1642 train, 412 test sentence pairs	Clinical notes	Clinical semantic textual similarity	2054 sentence pairs	n2c2 2018 + 2 electronic health record systems, GE and Epic	2 clinical experts for independent annotation
SemEval 2014 task 7 [113] (Shared task)	199 train, 99 valid, 133 test	Clinical reports: Discharge summaries, echo-cardiogram reports, electrocardiograph reports and radiology reports.	Entity type: Disease/Disorder	19165 mentions	ShARe (clinical notes from MIMIC II)	21 participants.
SemEval 2015 task 14 [36] (Shared task)	298 train, 133 valid, 100 test	Discharge summaries and radiology reports	Disorder attributes: mentions, normalization, negation, subject, uncertainty, course, severity, conditional, generic, body location	19111 mentions	ShARe	double annotation and adjudication by professional trained coders.
SemEval 2016 task 12 [12] (Shared task)	293 train, 147 valid, 151 test	Clinical notes and pathology reports	–Time expression identification –Event expression identification –Temporal relation identification	–7863 time expressions –78854 event expressions –23243 temporal relations	Mayo Clinic cancer patients	manually annotated and revised by the THYME project
SemEval 2017 Task 12 [13] (Shared task	621 train, 296 valid, 299 test	Clinical notes and pathology reports	–Time expression identification –Event expression identification –Temporal relation identification	–18623 time expressions –130293 event expressions –31169 temporal relations	Mayo Clinic cancer patients	manually annotated and revised by the THYME project
SemEval 2018 Task 6 [77] (Shared task)	232 train, 35 valid, 141 test	Clinical notes	Parsing time normalization	27362 time entities	THYME	linguistic students
Semeval 2021 task 10A [76] (Shared task)	10259 train, 5545 valid, 9580 test, 622703 un-annotated instances	Clinical notes	Negation detection	22409 asserted, 2975 negated	SHARP Seed, i2b2 2010, MIMIC II	manual annotation
Semeval 2021 task 10B [76] (Shared task)	278 train, 99 valid, 17 test, 47 un-annotated	Clinical notes	Time expression recognition	22151 time entities	THYME, TimeBank	two independent annotators and an adjudicator

Table 2 (continued)

Dataset	Size	Data type	Information	Statistics	Data source	Annotation method
NCBI-disease 2014 [33]	593 train, 100 valid, 100 test	PubMed abstracts	Entity type: Disease	6892 mentions	MeSH, OMIM	14 annotators, 2 per document: annotation in 3 phases with checking for corpus-wide consistency of annotations.
BC5CDR 2015 [85] (Shared task)	500 train, 500 valid, 500 test	PubMed abstracts	–Entity types: Disease and Chemical. –Relation types: chemical-induced disease.	–4409 disease –5818 chemical –3116 chemical –induced disease relations	Most were selected from an existing CTD-Pfizer collaboration-related dataset.	MeSH annotators were recruited for manual annotation.
CRAFT 2017 [23]	97	PubMed full-text articles	–Entity types: Gene, Chemical, Protein, Organism, Cell and Taxonomy. –Structural annotation : syntax and document structure. –Co-reference annotation.	≈ 140000 concept mentions	PMC Open Access subset	Manual annotation using an annotation model and guidelines by the help of a realist ontologies.
NLM Chem 2021 [56] (Shared task)	80 train, 20 valid, 50 test	PubMed full-texts journal article	–Entity type: Chemical. –Segments: sections, paragraphs, figure caption, titles, etc.	38342 mentions	PMC Open Access dataset	Doubly annotated by ten expert NLM indexers.

the abstract of PubMed articles. Indeed, there are some datasets [24, 56] obtained from full-text articles which can provide richer information rather than just an abstract.

Also, many methods use supplemental resources such as dictionaries and ontologies. The most popular resources that used in IE are UMLS [15] meta-thesaurus and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [32] ontology. The UMLS meta-thesaurus in its 2021 update can cover 16543671 terms in different languages (11755677 in English) which are associated with 4536653 medical concepts. This large biomedical thesaurus is collected from different sources such as SNOMED-CT, MEDLINE, MeSH, NCBI, etc. Also, it provides 127 semantic types of the concepts such as Disease/Syndrome, Clinical drug, Therapeutic or preventive procedure, laboratory or test result, etc. Furthermore, it provides 54 semantic relations between these semantic types, such as Process of, Result of, Property of, Part of, Associated with, Complicates, Affects, Causes, etc. The SNOMED-CT ontology covers more than 1314668 clinical terms with their description and they are associated with different concepts. There are 19 top-level concepts which are body structure, finding, event, environment/location, linkage concept, observable entity, organism, product, physical force, physical object, procedure, qualifier value, record artifact, situation, social context, special concept, specimen, staging scale and substance. In addition, it provides more than 3092749 relation mentions of 122 types of relations such as “is a”, “finding site”, “associated morphology”, “method”, “causative agent”, etc.

5 Evaluation metrics

The frequently used metrics for the information extraction tasks are accuracy, precision, recall, and f1-score. The accuracy reflects the guilt of a model to return the correct tag. However, the accuracy is not enough for multi-class problems where the precision and the recall should be taken into account to reflect the exactness and coverage, respectively. Additionally, the f1-score is the harmonic mean of precision and recall and balances both their concerns of them in one value. After calculating the scores from all the samples in the testing set, their average can be calculated to get the overall score for each metric which is between 0 and 1, where a higher value means higher performance. For the task of NER, each token is annotated according to the Inside-Outside-Beginning (IOB) [116] style that can be adapted to precise the beginning and the ending of each entity. For example, the tag “B-DIS” indicates that the token defines the beginning of a named entity of type DIS (Disease) while the tag “O” indicates that the token is outside all named entities. To perform this type of evaluation, we have applied the different metrics using the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

with TP (True Positive) is the number of well-annotated tokens by a specific tag, FN (False Negative) is the number of badly annotated tokens by a specific tag, TN (True Negative)

Table 3 The components of the confusion matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted Condition	Predicted condition positive	**True positive**	**False positive**, error I
	Predicted condition negative	**False negative**, error II	**True negative**

is the number of well-tagged tokens of other tags, and FP (False Positive) is the number of badly tagged tokens of other tags. Table 3 clarifies their meaning by representing the components of the confusion matrix.

For the entity-level evaluation, there are different methods to apply these metrics by taking only the named entities, instead of all the tokens, and analyzing them as samples. Thus, there are the exact matching and the partially methods [8]. The exact method is based on strict or exact matching while the partial method is based on type or partial matching. For strict matching, the exact named entity boundary and the type should be matched. For exact matching, the exact boundary must be matched even if it's not the same type. For partial matching, partial boundary detection is allowed, regardless of the type. For type matching, the type must be correctly matched even with a partially detected boundary. Indeed, the calculation of the precision and recall is slightly different with these matching ways. Before calculating the measurements, we need to calculate the number of correct, incorrect, partial, missing, and spurious samples which varies from one matching way to another. The formulas used to calculate the precision and the recall of all entity-level measures are the following:

$$Precision = \frac{COR + (0.5 * PAR)}{COR + INC + PAR + SPU} \tag{5}$$

$$Recall = \frac{COR + (0.5 * PAR)}{COR + INC + PAR + MIS} \tag{6}$$

With COR, INC, PAR, MIS and SPU are the number of correct, incorrect, partial, missing and spurius samples, respectively. Note that PAR is available only for the partial method (i.e. partial or type matching), otherwise it is counted as 0. Also, PAR is only used for the partial matching while INC is counted as 0 for this matching.

6 Information extraction

Information extraction is a step in text mining that is similar to pre-processing in the classical data processing procedure. This step is mainly based on automatic Natural Language Processing (NLP) and machine learning [150]. The main technologies of EMR Information Extraction (IE) are Named Entity Recognition (NER) and Relation Extraction (RE). Indeed, there are three main steps for the clinical IE that start with the extraction of medical problems, tests, and treatments from discharge summaries and progress notes. Then, the classification of relationships between medical concepts [27]. In addition, the main tasks in the recognition of named entities are the identification of clinical events and temporal expressions. In addition, the main task in the biomedical literature is the recognition of bio-entity names [158]. In fact, the Section Detection (SD) has proven to be an important task for the medical information extraction and it is able to enhance the different other tasks in this field [110]. In this section, we discuss some state-of-the-art methods of each task such as NER, RE, and SD. For that, we have categorized these methods according to their used techniques and we have discussed them all with a numerical comparison for each task in order to make a comprehensive review.

6.1 Named entity recognition

NER was introduced in 1995, its general role is to identify types of names and symbols [44], but its role in the medical domain is to identify medical entities that are important for treatment, such as disease names, symptoms, and drug names. It is a task of IE and it consists of recognizing references to different types of entities such as medical problems, tests, allergies, risks, adverse events, drugs, and treatments. Indeed, the complexity of natural language increases with misspellings [73, 115], different language structures [136], detection of acronyms or abbreviations, expansion and disambiguation [69], anaphoric relations [20], etc. There are two main steps for this task which are the identification of the entity boundary followed by the determination of the entity class. The metrics used for evaluating NER methods are precision, recall, and f1-score, which can be computed based on token level or entity level. For entity level, there are two methods, partial matching and exact matching [8]. These methods are affected by the physician's writing style, different writing forms of medical terms, ambiguity in term abbreviations, and compound or modified medical terms, especially in Chinese. NER methods are classified into two main classes: classical approaches and deep learning based approaches. Hereafter, we review state-of-the-art approaches within each category.

6.1.1 Classical approaches

Classical methods use hand-coded rules to exploit supplemental resources where they use fuzzy matching to find entities. These methods can identify entities by satisfying the rule's conditions which are based on text features and maybe a knowledge base. In addition, many methods use classical machine learning models such as Conditional Random Field (CRF) or Support Vector Machine (SVM) to perform a sequence-labeling or a classification task by training them on an annotated training set. These models are usually based on traditional features where the Part of Speech (PoS) is the most used one.

Lei et al. [80] have tested multiple features such as bag-of-characters, word segmentation, PoS, and section information by training multiple machine learning models on manually annotated Chinese clinical documents to predict named entities such as clinical problems, procedures, laboratory tests, and medications. They conclude that Structural SVM (SSVM) and CRF sequence-labeling models outperform others, where SSVM is the highest. In addition, most features are gainful for the Chinese NER systems even with limited improvements. Moreover, the fusion of word segmentation and section information leads to the highest performance, and they complement each other. Also, they found that domain knowledge is important for Chinese word segmentation. However, the NER on English clinical text is more difficult than the Chinese because it contains many more entity mentions and the boundaries of its entities are harder to detect. Furthermore, most errors occur in long entities where they are not often completely detected. In addition, the training set can't cover all concepts, and some errors are caused by unseen samples.

Quimbaya et al. [115] have proposed a combined approach by preprocessing the text of electronic health records to apply exact and fuzzy NER techniques with the help of a knowledge base. In addition, a lemmatized recognition is applied after lemmatizing the target text and the knowledge base. Then, the overlapping entities recognized by these three techniques are combined to decide the final result. As an advantage, the use of the Fuzzy Gazetteer match approach can find more instances of the dictionary concepts and even mistyped instances. Furthermore, the combination of these techniques improves the recall results. However, this method does not take into consideration the context and surrounding words that appear near a candidate named entity.

Ghiasvand and Kate [42] use the Decision Forest classifier to find named entities and their boundaries, while seed terms from UMLS are used for an unsupervised annotation. They have used 3 words before and after the target named entity which is presented by PoS, lemmatize and stemming forms, and UMLS semantic types as features. Initially, training samples are collected using exact matched unambiguous terms from UMLS. Then, new samples are gathered by applying back the trained method to the corpus for self-training. To select samples for the model, they extract noun phrases that have at least one medical word which is included in any UMLS term. Then, another step is added to learn if a word can expand the boundary of its named entity using the same classifier and features as UMLS. As an advantage, this method does not require any manual annotations. In addition, this method can determine the correct boundaries for the detected named entities. Moreover, it performs better than other unsupervised methods and is competitive with supervised methods. However, using data from different sources may reduce the performance. Furthermore, the automatically obtained noun phrases are not always perfect and can lead to errors.

The work of Song et al. [128] consists in evaluating learning models for NER in the biomedical domain. For this, the authors use a model based on CRF, and another one based on RNN. Also, different word representations are tested with these models, such as Word2vec, Global Vector (Glove), and Canonical Correlation Analysis (CCA). In addition, these models have been compared with other state-of-the-art methods. In this work, a dataset is annotated by biomedical categories such as protein, DNA, RNA, cell type, and cell line. Thus, the models trained on the annotations to predict the correct categories for the named entities. Indeed, the results show that the CRF model with Word2vec features outperforms all other models. As an advantage, word features are automatically built thanks to unsupervised learning by Word2vec which exploits the juxtaposition of words to extract their context. Thus, feature construction does not require manual annotation, dictionaries, domain knowledge, or other external resources. In addition, CRF can consider the context and neighbors of the entered word. However, CRF still needs manual annotation for training. Moreover, the models proposed in this work are not optimized enough.

Xu et al. [155] have proposed an unsupervised rule-based method to detect boundaries and classify medical entities mentioned in a medical Chinese text, and link mentions to their entities. It is based on Word2Vec, string similarity, lexical resources, PoS, and dependency parsing. Initially, the method exploits the PoS and dependency relations and maps the text to concepts in offline and online lexical resources to detect mentions of medical entities. Then, it classifies these mentions and gives a Word2vec representation for each category. Next, the approach selects candidate entities from a knowledge base that are most similar to medical entity mentions based on the characters. In addition, candidates that are not similar to the category representation will be removed. Finally, the method calculates the similarity between the mention and its candidates according to the common characters, the popularity of the entity, and the similarity between the words in the context that have a dependency relationship with the mention and the words in the description of each candidate. In addition, semantic correlation knowledge is added by computing the character similarity between the linking entity descriptions of the context's mentions and the description of each candidate. Thus, the target entity is the candidate with the best similarity. As an advantage, the method is unsupervised and generalizable. Also, it can recognize nested entities and better cover the medical entities. In addition, it outperforms the state-of-the-art methods in terms of performance. Furthermore, the efficiency of online detection to solve the limitation problem of the dictionaries is well proven. However, the linking approach lacks semantic analysis. In addition, the detection may obtain inexact entities in the boundary detection step, and the filtering of non-medical terms may lose medical entities.

Alex et al. [2] have used hand-crafted rules and lexicons made by experts for NER from tokenized and POS-tagged medical text. These rules are based on segmentation with syntactic analysis, PoS, lemmatizing and lexical lookup. This method can recognize named entities reliably and accurately using brain imaging reports from Edinburgh Stroke Study (ESS) data and perform very well on new data. ESS was the first set to be annotated by experts in this domain. Furthermore, this method creates a useful data structure to deal with nested entities. Also, it can outperform machine learning approaches. However, hand-crafted rules are costly and time-consuming, especially when adapted to new and different data.

The method of Chirila et al. [21] consists in extracting the most important information, which are named entities, from the sections of a semi-structured drug package insert. For this purpose, the authors trained the CRF-based Stanford NER Tagger model on leaflets to predict word entities using distributional similarity-based features in addition to other word features. According to the results, the accuracy with drugs of the same type is the highest. But, the method is tested only for the "Therapeutic indications" section and for the Roman language.

Discussion These rule-based techniques are more precise to find entities especially when they are properly hand-crafted and matched correctly to the data. This approach usually aimed to parse sentences and use the PoS feature which is beneficial for NER [162]. Also, these methods focus on entity name variation problem solving by performing a preprocessing step and a fuzzy matching. Additionally, they can apply a post-processing step to filter candidate entities to reduce ambiguity. The machine learning methods, especially based on the CRF model, can better detect the entity boundary based on sequential labeling or boundary expanding techniques. Usually, the used models don't need a manual effort to generate features. However, this approach is valid only for specific datasets [117] and require manual construction of rules, training set, or dictionaries with the assistance of a medical expert where there is a lack of data. Also, it is difficult to cover several variants of medical terminology with a single dictionary. Indeed, the used machine learning models need more improvement and other features need to be exploited.

6.1.2 Deep learning-based approaches

Recently, these are the most used methods for the NER task. They use deep learning models to annotate words in sentences in order to define which words belong to a specific type of entity after training on a big training set. Usually, they segment text into sentences and generate various features for their words to perform a sequential annotation based on a deep learning model such as LSTM which can be combined with other models. Also, they use a special annotation style to determine the beginning and the end of the entity along with its type during the recognition.

Li et al. [84] have improved a deep learning model based on the Deep Belief Network (DBN) to predict if a word belongs to a named entity using its PoS with its Word2vec feature vector. Indeed, using the PoS of the word leads to the best performance. Thus, this method is beneficial for NER. Moreover, Word2vec vectors capture useful semantic information. Moreover, with the improvement, DBN outperforms state-of-the-art methods. But, this requires manual annotation of the training data. Also, the content of the corpus knowledge is not rich enough.

The work of Nayel and ShashrekhaH [106] consists in exploiting dictionary information with the representation vector of the word and its characters to predict the named entity by a Bidirectional Long Short-Term Memory (BiLSTM) model. For this purpose, the authors used a Skip-gram model to build the word representation. Also, they gave an initial representation vector for each character to train a BiLSTM to generate orthographic features from the characters of a word. Noting that all numbers are replaced by "NUM" and characters are changed to lower case. In addition, the method adds dictionary information by using a merged disease vocabulary (MEDIC) as a dictionary. Thus, it represents dictionary information for each word in a binary vector to indicate whether a word is an abbreviation or synonym of a disease or is part of a multi-word disease name. Then, all these 3 types of features were passed to a BiLSTM model to do the learning. Thus, the contextual representation generated by this model is passed to a Conditional Random Field (CRF) layer to classify and select the most appropriate feature to annotate the input word from a sequence of words. As an advantage, Skip-gram learning improves the semantic and syntactic representation of words, especially with a huge biomedical and generic text corpus. Also, the character representation significantly increases the performance. Moreover, the addition of dictionary information improves the result according to f1-score.

Sun and Bhatia [133] is based on sequence tagging by fine-tuning RoBERTa model [87] on Medical Information Mart for Intensive Care III (MIMIC-III) dataset [61] as a NER tagger, and train a gazetteer tagger (w/NER tagger) on clinical datasets. They have prepared a knowledge base for the gazetteer containing medical conditions and drugs from the Unified Medical Language System (UMLS) [15] meta-thesaurus. Thus, the NER tagger is trained using word and character BERT embedding, while the gazetteer tagger is trained separately using gazetteer embedding. The output of these models before their Softmax layers are merged to return the entity type of the word such as medication, treatment, etc. As an advantage, the fusion of the separate taggers leads to better interpretability and flexibility. Also, it's worth noting that even without the gazetteer tagger during inference, the NER tagger can preserve the gains. Furthermore, the fusion model is data-efficient, interpretable, and able to improve NER systems and easily adapt to new entity mentions in gazetteers. In addition, this model can benefit from different clinical NER datasets. Also, the use of name knowledge from the gazetteer leads to an improvement. Furthermore, this method is effective in handling non-stationary gazetteers and limited data. However, this method can be improved by extending to structured knowledge to further improve NER systems and for more interpretability.

Zhao et al. [165] propose a weakly supervised method where they manually prepare some seeding rules and automatically extract all possible rules from the unlabeled text for each of the six rule types, and connect them in a graph using cosine similarity. Note that the rule is represented by the average contextual embedding of its matched candidate entities. Then, propagate the labeling confidence from seeding rules in the graph to obtain new rules. These new rules are applied to the text to obtain a label matrix in order to estimate noisy labels using Linked Hidden Markov Model (LinkedHMM) generative model. Finally, Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) discriminative model is trained with noisy labels using Bidirectional Encoder Representations from Transformers (BERT) [31] as token embedding to label tokens by their entity types. As an advantage, high-quality new labeling rules can be automatically learned from only a few manually constructed rules and an unlabeled text. In addition, with a limited number of manual rules, substantially better performance can be achieved. Furthermore, they defined six useful types of rules for entity recognition. Moreover, the performance is improved using a graph neural network with a new class distance-based loss function to maximize the distance between positive and negative rules. However, some rules are overlapping and are not helpful to improve the recognition. Also, some rules learned from training data can't be applied to the testing set due to the mismatch between these datasets.

Deng et al. [30] combine BiLSTM with CRF to be trained on crawled and manually annotated Chinese TCM patents' abstract texts using character embedding, in order to recognize entity types such as herb names, disease names, symptoms, and therapeutic effects. Firstly, each character embedding vector is an input of a BiLSTM time step. A pre-trained embedding matrix is used to represent each character's one-hot vector and is fine-tuned during the back-propagation, in order to extract sentence features. Finally, the CRF layer learns the potential relationship between sequences and returns the optimal labeling sequence. As an advantage, this method can learn semantic information in the context without feature engineering. Mainly, the use of characters instead of words leads to better performance, while the characters contain a lot of linguistic information and are able to mostly avoid errors caused by poor segmentation. Furthermore, BiLSTM can easily learn about contextual relationships in the text to provide more comprehensive contextual information. Likewise, the CRF layer complements the BiLSTM by optimizing the recognition comprehensively from the sentence level. However, the used data scale affects the learning model and cannot well support its requirements. Also, this method is restricted by the entity labeling granularity where some entities are nested within other entities.

Zhou et al. [168] have pre-trained two deep contextualized language models on the clinical corpus from the PubMed Central (PMC): Clinical Embeddings from Language Model (C-ELMo) for word-level features and C-FLAIR clinical contextual string embeddings for character-level features. Then, each of the two embeddings is concatenated with Glove embedding and passed to the BiLSTM-CRF model to extract entity types. As an advantage, the models gain dramatic improvements compared to domain-generic language models and static word embeddings. In addition, these models can support different applications. Results show that C-Flair can handle entities that do not appear in word-level vocabulary. As well, C-ELMo can better capture the relationship between the word-level contextual features. Also, word-level models may be more robust than purely character-level models. However, it is hard for the two models to correctly recognize complex phrase-level entities.

Li et al. [82] have proposed a span-based method that exploits BERT, BiLSTM, and attention-guided GCN (AGGCN) to recognize entity fragments from text spans and predict the relations between them which can be "Succession", "Overlapping" or "Other". For a given sentence, the words are represented by BERT embedding followed by a BiLSTM layer. Hence, this representation is enhanced by an AGGCN after transforming the sentence into a graph using dependency parsing which provides syntax information. Then, text spans are generated [91] and represented by the original and the enhanced representations. Two Multi-Layer Perceptrons (MLP) are used to recognize entity fragments from these spans and to predict the relations between them. As an advantage, this graph-based method is able to enumerate and represent well the text spans to recognize the discontinuous and overlapping entities which can't be done by sequential methods. Also, it is more suitable for modern end-to-end learning since it doesn't need feature engineering or manually-designed transitions for parsing [25]. The authors show that BERT is the most effective word embedding which effectively captures the contextual information. In addition, BiLSTM is able to enhance any word embedding and especially the non-contextualized ones by helping them capture the context. Moreover, word embedding such as Word2vec or BERT is more effective than ELMo character-level embedding by capturing the whole meanings of words. Indeed, AAGCN gives higher precision results compared to the BiLSTM sequence labeling method. However, the BiLSTM sequence tagger is more accurate and gives higher precision in entity boundary detection where it uses label dependence.

The method of Sui et al. [132] is based on the GCN model to consider the problem as a node classification after transforming the sentence into a graph. Thus, the interactions among the words, entity triggers, and the whole sentence semantics are used to recognize the entity. For that, a dictionary of entity triggers is prepared manually. As an advantage, by using the entity triggers, the model is able to recognize the entity by its context to treat the ambiguity and variation problems. Also, this method is cost-effective and efficient and is able to cast the problem into a graph node classification by using a trigger-based graph neural network rather than a sequence tagging to well recognize nested and discontinuous entities. However, this method requires a manual effort by experts to annotate a lot of words in order to prepare the entity triggers.

Discussion Various features can be well used by these methods and can be deeply exploited in a sequential way which is very appropriate for the NER task. The most used models especially in the recent methods are BiLSTM and CRF which are usually combined. BiLSTM is able to extract high-quality features considering the contextual dependency in double directions where CRF performs an optimized sequence tagging by these features. The IOB annotation style is the most appropriate to identify the boundary of the entities. The contextual word embedding is highly used and it performs better when it is trained on medical text. Few recent methods [82, 132] have been based on graph tagging which is better to handle nested and discontinuous entities. However, a big high-quality training set is required to cover all cases in new data which is the problem of most methods. Indeed, by the IOB annotation style and in a sequential manner, the models are unable to handle the nested entities. In fact, the use of GCN to treat sentences transformed into graphs [132] is not well explored while it should be more suitable for this task.

6.1.3 Numerical Comparison

Tables 4 and 5 show the available results of classical and deep learning-based named entity recognition methods, respectively. The A, P, R, and F symbols refer to Precision, Recall, and F1-score, respectively. Many types of features have been exploited which can be categorized into 4 types [137]: word-level information (e.g. word embedding and orthographic information), syntactical information (e.g. PoS), lexical and semantic information (e.g. UMLS) and discourse information (e.g. section). For classical approaches, the word-level information is not well exploited where only TF-IDF and Bag of Word (BoW) were used by some work. In contrast, the most used feature for deep learning based approaches is the word embedding and especially the domain specific and contextualized ones such as BERT. The semantic information is the most used feature for classical methods which is usually generated by UMLS. However, the supplemental resources can't be useful for all types of data. For example, the F1-score result of the method of Ghiasvand and Kate [42] is decreased by $\approx 18\%$ when it is applied on another dataset. In addition, the PoS information is widely used and it is relevant to this task since the named entities are mostly overlapping with noun phrases

Table 4 Classical NER methods

Pub.	Dataset	Dataset size	Method	Features	P	R	F
Zhang and Elhadad [162]	i2b2 2010	170 train, 477 test	unsupervised annotation and boundary detection: noun phrase chunker + IDF filter + Cosine similarity	UMLS + TF-IDF	49.73%	54%	51.06%
Tang et al. [137]	i2b2 2010	349 train, 477 test	SSVM	clustering-based word embedding + section + UMLS + PoS + BoW + orthographic information	87.38%	84.31%	85.82%
Ghiasvand and Kate [42]	i2b2 2010	349 train, 477 test	unsupervised Decision Forest classifier for annotation and boundary detection	UMLS + PoS + lemma + stem	52.66%	68.86%	59.43%
Ghiasvand and Kate [42]	SemEval 2014 Task 7	199 train, 133 test	unsupervised Decision Forest classifier for annotation and boundary detection	UMLS + PoS + lemma + stem	88.1%	69%	77.3%
Popovski et al. [112]	FoodBase corpus annotated by food concepts	1000 recipes	Pre-processing + morphological analysis + rule engine	computational linguistics + semantic information which describe concepts	97.80%	94.37%	96.05%

Table 5 Deep learning based NER methods

Pub.	Dataset	Dataset size	Method	Features	P	R	F
Wu et al. [154]	i2b2 2010	349 train, 477 test	LSTM	Word embedding	85.33%	86.56%	85.94%
Lee et al. [79]	i2b2 2010	170 train, 256 test	BERT model	pre-trained and fine-tuned BioBERT	-	-	86.46%
Zhou et al. [168]	i2b2 2010	349 train, 477 test	LSTM-CRF	pre-trained contextualized embeddings + Glove embedding	-	-	87.45%
Sun and Bhatia [133]	i2b2 2010	170 train, 256 test	BiLSTM-CRF and gazetteer model combined on a shared tagger	word/character embedding RoBERTa-mimic + Gazetteer embedding	-	-	87.41%
Lee et al. [79]	NCBI-disease 2014	593 train, 100 valid, 100 test	BERT model	pre-trained and fine-tuned BioBERT	-	-	89.36%
Zhou et al. [168]	NCBI-disease 2014	593 train, 100 valid, 100 test	LSTM-CRF	pre-trained contextualized embeddings + Glove embedding	-	-	87.88%
Zhao et al. [165]	NCBI-disease 2014	593 train, 100 valid, 100 test	propagate seeding rules in graph + LinkedHMM + BiLSTM-CRF	contextual embedding + BERT	89.9%	73.2%	80.2%
Kim et al. [64]	n2c2 2019 for family history extraction	99 train, 117 test	voting ensemble of BiLSTM models + heuristic rules + OGD + ConText	UMLS + PoS + dependency-based embeddings + static embedding + context-dependent embedding + lexical	84.83%	87.24%	86.02%
Yang et al. [160]	n2c2 2019 for family history extraction	99 train, 117 test	majority voting of LSTM-CRF models with BERT fine-tuning	Fasttext embedding + pre-trained BERT	79.69%	79.20%	79.44%
Lee et al. [79]	MACCROBAT 2018 case reports	160 train, 20 valid, 20 test	BERT model	pre-trained and fine-tuned BioBERT	-	-	64.38%
Zhou et al. [168]	MACCROBAT 2018 case reports	160 train, 20 valid, 20 test	LSTM-CRF	pre-trained contextualized embeddings + Glove embedding	-	-	65.75%

Table 5 (continued)

Pub.	Dataset	Dataset size	Method	Features	P	R	F
Sun and Bhatia [133]	i2b2 2009	170 train, 256 test	BiLSTM-CRF and gazetteer model combined on a shared tagger	word/character embedding RoBERTa-mimic + Gazetteer embedding	-	-	92.35%
Sui et al. [132]	BC5CDR 2015	500 train, 500 valid, 500 test	LSTM + GCN + CRF	sentence embedding + entity triggers	94.19%	92.73%	93.45%
Zhao et al. [165]	BC5CDR 2015	500 train, 500 valid, 500 test	propagate seeding rules in graph + LinkedHMM + BiLSTM-CRF	contextual embedding + BERT	88.2%	84.6%	86.3%
Zhao et al. [165]	LaptopReview 2016 laptop aspect terms	3048 train, 800 test sentences	propagate seeding rules in graph + LinkedHMM + BiLSTM-CRF	contextual embedding + BERT	82.4%	64.2%	72.2%
Vunikili et al. [144]	CANTEMIST-NER sub-task with tumor morphology mentions	501 train, 500 valid, 5232 test	transfer learning to fine-tune the BERT model	BERT contextual embeddings pre-trained on general domain Spanish text	72.7%	74.1%	73.4%
Devlin et al. [30]	crawled TCM patents' abstract texts annotated with herb names, disease names, symptoms, and therapeutic effects.	1600 copies: 60% train, 20% valid, 20% test characters	BiLSTM-CRF	pre-trained and fine-tuned character embedding	94.63%	94.47%	94.48%
Sun and Bhatia [133]	DCN [14] de-identified clinical notes with medications and medical conditions annotations	1500	BiLSTM-CRF and gazetteer model combined on a shared tagger	word/character embedding RoBERTa-mimic + Gazetteer embedding	-	-	84.59%

[162]. Indeed, some classical approaches [42, 162] are based on unsupervised techniques to prepare the training set, but they give lower results which explains the difficulty of automatically constructing a high-quality training data. The method of Popovski et al. [112] based on rules gives the highest results, but it should be not suitable for other data types. Most of deep learning methods are based on LSTM, CRF and BERT. However, there is a recent method based on GCN [132] which gives higher results compared to BiLSTM-CRF on the same dataset. Thus, the GCN models could be more suitable for this task since it can handle all forms of entities such as discontinuous and nested entities.

6.2 Relation extraction

In RE technology, the relationship represents the semantic relation between two named entities appearing in the same sentences. Indeed, understanding the semantic relations between entities is required for many applications in IE such as semantic knowledge base construction to infer the relationships between entities, and the development of question answering systems for text summarization or concepts taxonomy construction. Semantic relations can be classified into two families according to their types, paradigmatic relations and syntagmatic relations [11]. Mainly, paradigmatic relations are operating on concepts of the same class. Usually, concepts are organized as a tree, where these relations are represented hierarchically by vertical links. These relations include the relation of synonymy, antonymy, and hypernymy while the syntagmatic relations link two or more medical entities. They represent semantic links in an expression and link multiple linguistic units. They can be found by analyzing the syntactic forms in text and by a predicate. According to Uzuner et al. [143], we can also categorize these relations depending on the type of relationships, such as disease relationships, disease-medical examination relationships, and disease-treatment relationships. Generally, the cause-effect or causal RE has received ongoing attention in many medical fields [157] which can assist doctors, by supporting the construction of a knowledge graph, to quickly find causality and customize treatment plans. There are many examples of causal relations such as diseases-cause-symptoms, diseases-bring-complications, and treatments-improve-conditions. The most commonly used techniques in this task are rule-based methods and machine learning-based methods [159]. In our work, we have categorized relation extraction methods into classical approaches and deep learning based methods.

6.2.1 Classical approaches

Recently, most of the recent classical methods are based on machine learning models such as SVM usually by using different types of word embedding and high-level features generated by supplemental NLP tools [54, 98, 114]. These methods aim to train models to classify the relationship between two entities that appear in the same sentence or consecutive sentences. Commonly rule-based strategies include dependency parsing [40], concept co-occurrence detection [60] and pattern matching [81]. Usually, rules are defined manually by domain experts [74] or even automatically generated by using machine learning techniques [129] on annotated data and they rely on a set of patterns, procedures, or heuristic algorithms to directly identify candidate relations. Additionally, this technique can use supplemental knowledge resources and apply some matching and mapping techniques. The co-occurrence-based methods are the most straightforward techniques. The key idea is when there are 2 entities mentioned together with higher frequency, there is a stronger chance to be related together.

Eriksson et al. [37] have proposed a rule-based method to extract Adverse Drug Event (ADE) relations. They have built a dictionary of ADE based on the “Undesirable effects” sections in drug Summary of Product Characteristics (SPC) documents. They have used NER tagger for matching and post-coordination rules for ADE compound terms construction. Then, they have applied some post-processing rules and filters. As advantage, the synonym and anatomical location collapse enhance the method. The sections are exploited to reduce the manual effort of dictionary constructing. However, the absence of the lexeme from the dictionary causes missed annotations. Also, this method requires temporal knowledge of administered drugs for some relation types.

The method of Zhou and Fu [167] is based on rules to extract gene-disease association relationships. They have calculated co-occurence frequency weights for entities using textual resources and key words dictionary. Thus, each entity is represented by a vector based on weighted IDF. Then, the cosine similarity is calculated between each pair of entities to get the strength of the relation. As advantage, this method is easy to be conducted exploits the combination of text mining and graphic model. However, it targets only the gene-disease association.

Ben Abdessalem Karaa et al. [10] extract relations between drug and disease entities such as “cure”, “no cure”, “prevent”, “side effect”, and “other relations”. This method uses a combination of features for each sentence with the help of UMLS meta-thesaurus to provide semantic annotations by configuring the MetaMap system to extract concepts and semantic types. By combining the NLP technique and UMLS knowledge, many features can be extracted such as frequency, lexical, morphological, syntactic, and semantic features. Thus, these features are passed to an SVM model to predict the relation associated with the sentence using MEDLINE 2001 [120] as a standard training set. As an advantage, this method can extract correct and adequate features while they are relevant to discover interesting relationships between concepts. Furthermore, it outperforms other methods for all types of relations especially in terms of f1-score and the “cure” relation. Note that this performs better in a multidimensional context and is suitable for semantic relations in natural language texts. However, the lack of training data for the “no cure” relation leads to low performance for all comparison methods. In addition, this method requires an annotated training set.

The work of Tran et al. [141] is aimed at acquiring knowledge from COVID-19 scientific papers. For that, an enormous number of relations between entities are extracted by combining several methods. ReVerb [38] is based on verb-based relation phrases. OLLIE [99] extracts relations mediated by the verbs, nouns, adjectives, and more. ClausIE [28] is a clause-based approach. Relink [127] extracts relations from connected phrases. OpenIE [3] finds the maximally simple relations after breaking a long sentence into short and coherent clauses. Thus, the extracted relations are tagged by biomedical entities recognized by SciSpacy models [107] trained on a different corpus. Finally, the extracted relations are clustered and scored for their informativeness over the corpus to construct the retrieval system. As an advantage, higher extraction coverage could be obtained by combining several methods. In addition, various specialized entity information can be obtained by covering different sets of biomedical entities. Also, acquiring knowledge can be rapid and efficient across a large number of scientific papers. However, some wrong results are caused by the complexity of the biomedical text where there are long sentences, and conjunctions and nested clauses are commonly used.

The method of Mahendran and McInnes [96] is based on rules and exploits the co-location information to extract ADE relations. It uses a breadth-first search algorithm to find the closest occurrence of a drug entity to the referring non-drug entity. As advantage, co-location information is sufficient to identify most drug relationships. In fact, left traversal mechanism to find drug entities gives better results for some entity types. However, it is not suitable for other types where the drug entity is not on the left side.

Discussion Indeed, sentence structure analysis can be used to explore patterns where it improves the performance and the extraction of implicit causal relations. Also, useful features can be extracted for the RE task which can be well incorporated with machine learning models. However, there is a big lack of training sets for this task and most machine learning-based methods can't provide sufficient data for the relationship types and this greatly affects the classical machine learning models. Furthermore, preparing the rules and the supplemental resources is labor-intensive and needs a domain expert for manual effort. Moreover, this approach severely restricts the generalizability and portability of RE for other types of data. Also, some difficult cases can be found in sentences and phrases which prevent rules to be correctly applied.

6.2.2 Deep learning-based approaches

The principle of these methods is to use advanced models based on deep learning such as Long Short-Term Memory (LSTM) and Graph Convolutional Network (GCN). By this approach, useful information from syntactic structures could be extracted rather than applying manually constructed patterns for that [45]. Thus, different types of features can be generated with the help of language models, dictionaries, or rules by which a model deeply learns how to identify the relations in sentences after training on a huge amount of training examples. Usually, these methods perform a sequential analysis of words based on the LSTM model combined with other models.

Tran and Kavuluru [140] propose a novel distant supervision approach to extract medical treatment predication relations in PubMed abstracts by training a BiLSTM model after leveraging MeSH sub-headings and preparing training sentences with entities using NLM's MetaMap [6] and UMLS Semantic Network. Indeed, this method uses a variant of BiLSTM with a modified noise-resistant loss function, where the input is word embeddings and learnable position vectors. As an advantage, the position vectors can enhance the RE performance. Furthermore, the automatically generated training data is of reasonable quality without the costs of human involvement. In addition, MeSH sub-headings are precisely utilized which leads to better filtering of treatment and drug entities. However, this method focuses only on treatment predictions. Furthermore, this method may have difficulties dealing with trivially negative cases as it is not trained on them. Also, linguistic phrasing is understandably difficult when there is a weak connecting word. Moreover, false negatives may occur when a unique concept entity is mentioned several times in the same sentence.

The work of Yang et al. [159] consists of developing a series of RE models based on 3 transformer architectures, namely BERT [31], RoBERTa [87], and XLNet [156] to identify relation like "drug-Adverse events" and "Drug-Reason". This method uses already annotated entities to select candidate entity pairs for classification. Some rules are used as a strategy to generate these candidates, where the 2 entities must be a valid combination according to the annotation guidelines. Thus, the strategy selects only 2 entities in the same sentence or 2 consecutive sentences as a candidate entity pair. Another strategy is to use the same rules but by using cross-sentence distance to select the number of consecutive sentences. Hence, it applies a separate model for each group of candidates which have different distance values. The results show that clinical pre-trained transformers consistently achieved better performance. In addition, XLNet and RoBERTa achieved the best performance for 2 different datasets. Furthermore, the binary classification strategy consistently outperformed the multi-class classification strategy. Moreover, adding positional information to entities as features is critical to learning useful representations. However, there is no significant difference between the 2 generating candidates' strategies, but there is still a lack of an efficient method to solve both missing samples and sample distribution bias issues. Furthermore, there is a quite small number of training examples for some relation's categories which significantly lead to a training difficulty and decent performance. Indeed, this work only focused on the RE task, while this task is highly dependent on the NER result.

Shi et al. [126] train an end-to-end deep learning method to identify people's pandemic concerns, and extract "co-occurrence" and "cause-effect" relations between them in tweets. The authors consider 8 types of concerns which are finance, government, disease, medicine, person, location, food, and date and time. This method uses BiLSTM-CRF to detect concern entities combined with the Bidirectional GCN (BiGCN) model to extract relations, where a hidden state of BiLSTM is shared with BiGCN. Accordingly, each tweet is represented by sequential features using BERT embeddings and regional features using the Concern Graph (CG) module. In the CG, each node represents a concern associated with its score and type. The concern score is calculated by sentiment polarity and retweet count of the tweet. To represent the regional features, 3 vectors merge for each concern word to represent PoS and syntactic dependency relation, concern score and type, and relation features. For that, an automated deep learning-based framework [125] was used to detect and construct a concern knowledge graph to get the concern types and relations. Likewise, the same framework is used to annotate the training set. As an advantage, state sharing enhances the influences from concerns to improve the performance of the RE. Note that this relation can reveal people's thoughts behind the expressed concerns or identify the cause of public concerns. Furthermore, the regional features from CG improve the concern identification effectiveness and lead to a high noise tolerance. In addition to contextual information, this method captures specific features of entities by a designated CG to perform better on tweets. However, this method is not the best for high-quality and manually annotated datasets.

Kim et al. [64] propose a sequence labeling hybrid method to recognize family members and observation entities in EHR text notes and extract relations between them in addition to living status. A rule-based system is used to select family member entities by matching relevant noun terms with the help of PoS. Then, a number of BiLSTM models trained on dependency-based embeddings [67] as static embedding and Embeddings from Language Models (ELMo) [109] as context-dependent embedding. In addition, MetaMap [6] maps semantic types from UMLS and aligns them with entities to choose relevant ones. the family members and observations recognized by these BiLSTM models are ranked and have been voted based on the models' f1-scores. The heuristic rules are used to normalize the family member entities by a simple dictionary-based mapping, and determine the family side by looking at cue words considering the degree of relatives. Thus, two Online Gradient Descent (OGD) [16] models are trained on lexical features based on the identified entities to determine living status and observations associated with family members. Hence, alive and healthy scores are assigned for living status phrases using cue words. Likewise, the negation attribute is assigned to observations using ConText algorithm [19] with customized trigger terms. As an advantage, the voting ensemble of BiLSTM models contributes in terms of diversity to achieve better performance and provides efficient and convenient integration of individual LSTM models which are not deterministic. In addition, this method substantially benefited from a combination of 2 datasets. Integrating heuristics and advanced IE models lead to a high level of performance. The performance is improved especially on RE and benefited by the large training set and the pre-trained embeddings. However, choosing the voting ensemble threshold can achieve the best performance for one task but not the highest accuracy for other tasks. Also, some positive relations which rely on 2 entities in different sentences can be missed by using a carriage return character to filter examples.

Mahendran et al. [97] have trained a BERT model to predict the chemical relations in each sentence by combining the point-wise mutual information which is represented by a GCN model. Each time, a pair of candidate entities is selected from the sentence to predict a relation between them, while the other entities are masked. As advantage, a global association information is used with the local contextual information. However, BERT model is not trained separately and can't use a pre-trained model to make embeddings. Moreover, there is no features to represent word nodes for GCN.

drissiya El-allaly et al. [34] have used a GCN model to extract Adverse Drug Events (ADR) relations by jointly learning N-level sequence labelling, i.e. without treating each candidate relation independently. These authors combine the BERT representation with weighted GCN, where the dependency tree of the target sentence is used as input for GCN where each edge is scored. Furthermore, this method applies a multi-head attention to exchange boundary knowledge across levels. As advantage, the weighted dependency tree is able to capture rich syntactic features and determine the most influential edges. Also, this method leverages the contextual and structural information by combining BERT with GCN. In addition, it can deal with complex relations that include discontinuous, overlapping and nested entities. It is able to deal with complex relations by jointly learn N-level sequence labelling to capture greater interaction between relations. However, there is other important features which are not exploited such as PoS, relevant side information and global information. Also, the recognized named entities are not exploited.

Discussion This approach is widely used and can deal with the feature sparsity problem by transforming features into low-dimensional dense vectors. Deep learning techniques have exhibited superior performances compared to the traditional methods [7, 72] and they can better handle the input features. Recently, the GCN models [34, 97] are explored which are more suitable for this task to exploit relations between words, such as the dependency relations. However, the quality and the quantity of data, which is usually manually annotated by medical experts, have a big impact on these methods since they need sufficient examples to cover all cases. Also, it is hard to adapt the model to another type of data where it should be specified on the target type.

6.2.3 Numerical comparison

Tables 6 and 7 show the results of classical and deep learning-based RE methods, respectively. The P, R, and F symbols refer to Precision, Recall, and f1-score, respectively. For the classical approaches, most of the methods are based on SVM and CRF models while others are based on rules which consider the co-occurrence, statistical and other information. However, Usually, these methods use supplemental resources to get knowledge about the relations. Recently, this type of methods is rarely used since most of the methods are deep learning oriented. Generally, these deep learning based methods give higher results and they are usually based on LSTM, CRF and BERT. In fact, these models give highest results when they are enhanced by rules. For example, Kim et al. [64] outperform the method of Yang et al. [160] on the same dataset by +7% in terms of F1-score. Likewise, Yang et al. [159] generate candidate by rules for a BERT model and their method outperform

Table 6 Classical RE methods

Pub.	Dataset	Dataset size	Method	Features	P	R	F
Mahendran and McInnes [96]	n2c2 2018	303 train, 202 test	breadth-first search algorithm	co-location information	88%	83%	86%
Eriksson et al. [37]	Patient notes with ADE mentions	200 notes	NER NER gazetteer + post-coordination rules + post-processing rules	section + dictionary	89%	75%	81%
Wang et al. [147]	MEDLINE 2001 abstracts from biomedical journals annotated by cure and side effect relations	4600+ abstracts: 75% train, 25% test sentences	pattern to extract candidate pairs + generate the degree of correlation	UMLS + lexical + network embedding	91.75%	86.55%	89.025%
Ben et al. [10]	MEDLINE 2001 abstracts from biomedical journals annotated by cure and side effect relations	4600+ abstracts: 75% train, 25% test sentences	SVM model	UMLS + frequency + lexical + morphological + syntactic + semantic	86.51%	91.56%	88.78%

Table 7 Deep learning based RE methods

Pub.	Dataset	Dataset size	Method	Features	P	R	F
Yang et al. [160]	n2c2 2019 for family history extraction	99 train, 117 test	majority voting of LSTM-CRF models with BERT fine-tuning	Fasttext embedding + pre-trained BERT	69.95%	61.84%	65.44%
Kim et al. [64]	n2c2 2019 for family history extraction	99 train, 117 test	voting ensemble of BiLSTM models + heuristic rules + OGD + ConText	UMLS + PoS + dependency-based embeddings + static embedding + context-dependent embedding + lexical	73.27%	71.70%	72.48%
drissiya et al. [34]	n2c2 2018	303 train, 202 test	GCN + N-level sequence labelling	weighted dependency graph + biomedical BERT	96.63%	94.86%	95.74%
Wei et al. [152]	n2c2 2018	303 train, 202 test	multi-class classification with BERT model	BERT fine-tuned on MIMIC-III	98.38%	90.15%	94.09%
Mahendran and McInnes [96]	n2c2 2018	303 train, 202 test	binary classification with BERT model	fine-tuned BERT	93%	96%	94%
Yang et al. [159]	n2c2 2018	303 train, 202 test	rules to generate candidates + binary classification	Clinical BERT + RoBERTa + XLNet	97.01%	95.12%	96.06%
Hasan et al. [49]	i2b2 2010	170 train, 256 test	BiLSTM	Word2vec, relative distances, PoS, Concept embedding, dependency tree	-	-	88.08%
Wei et al. [152]	i2b2 2010	170 train, 256 test	multi-class classification with BERT model	BERT fine-tuned on MIMIC-III	76.24%	77.34%	76.79%
Yang et al. [159]	MADE1.02018 [57] fully de-identified longitudinal EHR notes	876 train, 213 test	rules to generate candidates + binary classification	Clinical BERT + RoBERTa + XLNet	91.26%	87.99%	89.59%
Shi et al. [126]	COVID-19 Twitter dataset with concern categories	1418 train, 355 test	BiLSTM-CRF to detect entities + BiGCN with shared BiLSTM hidden state	BERT embeddings + PoS + syntactic dependency relation + concern score and type + sentiment polarity and retweet count	54.5%	63%	56.7%

the BERT model proposed by Mahendran and McInnes [96] which does not use rules. Thus, hybrid methods can perform better for this task. It is worth noting that syntactical information, such as dependency tree and PoS, is useful and contributes to reach higher results [49, 64]. Recently, some methods [34] are based on GCN by transforming sentence into tree, or graph, based on dependency relations. These methods give promising results and should be further explored.

6.3 Section detection

The purpose of this task is to structure medical documents by identifying their sections which is useful for many medical IE tasks [110]. For example, the most relevant content of EHR can be found in the core medical content which makes filtering the headers and footers suitable for many tasks [29]. Also, identifying sections in discharge summaries such as "history of illness" and other sections is beneficial to extracting information about clinical problems, procedures, laboratory tests, and medications [80]. Furthermore, the "Therapeutic indications" section which can be found in medical prospectuses is useful to extract information such as drug-treated condition, a medicine name, a drug type, etc [21]. Also, the most important sections of these documents which contain relevant information for a drug prescription are: Contraindications, Therapeutic indications, and Dosage [94]. Generally, section detection aims to improve the performance of medical information extraction tasks that deal with natural language such as entity recognition [80], abbreviation resolution [169], cohort retrieval [35] and temporal RE [71]. This task was used to provide more context for other tasks, support cohort selection by information retrieval, and identify patients with risk factors [110]. In addition, some tasks can be improved such as co-NER and reference resolution by adding the section as a feature, distinguishing sensitive terms in de-identification, considering the order of events by identifying temporal sections, document quality assessment, and selecting supporting educational resources by extracting relevant concepts [110]. By definition, a section is a segment of text that groups together consecutive clauses, sentences, or phrases. It can share the description of a patient dimension, patient interaction or clinical outcome, etc. The unstructured text has sections that are explicitly or implicitly defined by the author. The explicit sections are defined by titles whereas the implicit ones are defined without titles [83]. In addition, the section has a granularity level where it can be a section or a sub-section. Indeed, all methods are limited to a low level of granularity and most of them detect only the top sections [110]. For some types of documents, the author is free and even a precisely defined template may be ignored during writing, which leads to less uniform titles and even sections without titles [138]. Also, section titles and orders in a document may differ from one source to another. Indeed, the lack of benchmark annotated data for training and evaluation represents a major obstacle to this task compared to others. The methods of section detection can be classified into classical approaches and deep learning based approaches.

6.3.1 Classical approaches

For this type of method, there are two used techniques: machine learning-based and rule-based methods. Some methods use classical machine learning models such as CRF and SVM to categorize the text into sections. Usually, these methods use syntactic and lexical features and perform a sequence labeling or classification after segmenting the text mostly into sentences. Most methods are based on rules which depend on supplemental resources where they usually match titles to define the beginning and the ending of sections. Also, some methods use rules to enhance machine learning models usually for the training set construction.

Jancsary et al. [58] have trained CRF to recognize (sub)sections in report dictations giving lexical, syntactic categories, BoW, semantic type, and relative position features for each word. The training data is constructed by aligning the corrected and formatted medical reports with the text from automatic speech recognition while the annotations are generated by mapping (sub)headings to the (sub)section labels using regular heading grammar. As an advantage, this method can detect various structural elements even without explicitly dictated clues. Furthermore, it can automatically assign meaningful types for (sub)sections even in the absence of headings. In addition, it is still effective under ideal conditions and can deal with the errors of real-life dictation. However, manual correction is required to solve the errors of the automatically generated annotations which impact the segmentation results.

Apostolova et al. [4] have constructed a training set by hand-crafted rules to train SVM to classify each medical report sentence into a semantic section using multiples features such as orthography, boundary, cosine vector distance to sections, and exact header matching. As an advantage, a high-confidence training set is created automatically. Also, the classification of semantically related sections is significantly improved by boundary and formatting features. Furthermore, the segmentation problem could be solved when the NLP techniques are applied. Moreover, using an SVM classifier outperforms a rules-based approach. However, it is hard to classify a section when its sentences are often interleaved with other sections.

Beel et al. [9] have proved that style information, specifically font size, is very useful for detecting titles in scientific PDF documents in many cases. The authors used a tool to extract formatting style information from a PDF file such as font size and text position. Then, they used a simple heuristic rule to select the three largest font sizes on the first page. Thus, identifying the texts that have these sizes as titles. This method outperformed an approach based on SVM, which uses only text, in accuracy and even in runtime. Moreover, this technique is independent of the text language because it only considers the font size. However, this method depends on the font size and requires the existence of formatting information.

The method of Haug et al. [50] consists in annotating each section in a medical document by its main concept. This approach is based on Tree Augmented Naive Bayesian Networks (TAN BN) to associate topics with sections as their semantic features. For this purpose, this method was trained using features generated by extracting N-grams from the text of section titles, in combination with the document type. The identification of the section topic improves the accuracy and avoids errors when extracting specific information. Thus, this task can reduce the natural language processing effort and prepares the document for more targeted IE. However, n-grams have limitations with complex and large documents. In addition, this Bayesian model does not consider the consistent sequencing of section topics.

The method of Deléger and Névéol [29] classifies each line in French clinical documents into its specific high-level sections such as header, content, and footer. Thus, a statistical CRF model is trained based on some information about the line taking into account the first token in surrounding lines, the first two tokens in the current line, the first token is in uppercase, the relative position of the line, the number of tokens, presence of preceding empty lines, digits and e-mail addresses. As an advantage, the performance is very high, especially for content and header lines. It is well noted that the headers and footers are very present in the document and should be identified to focus on the core medical content.

However, the granularity level of sections is very high while there are more useful sections within the content that are not identified.

Ni et al. [108] have classified medical document sections into pre-defined section types. These authors applied two advanced machine learning techniques: One is based on supervised learning and the other on unsupervised learning. For the supervised technique, a heuristic model pre-trained on old annotated documents is used to select some new candidate documents that will be annotated by people and will be used for learning. For the unsupervised technique, a mapping method was used to find and annotate sequences of words, which represent section titles, by their corresponding section types using a knowledge base. A maximum entropy Markov model was used for section classification. The chosen model is faster in learning and allows richer features. The techniques used can reduce the cost of annotation and allow a quick adaptation of new documents for section classification. In addition, both techniques can achieve high accuracy. However, the supervised technique requires more annotation costs than the other technique. In addition, the performance of the unsupervised technique is highly dependent on the quality of the knowledge base.

Dai et al. [26] have proposed a token-based sequential labeling method with the CRF model for the section heading recognition using a set of word features such as affix, orthographic, lexicon, semantic, and especially the layout features. To construct training data, they have employed section heading strings from terminology to make candidate annotations. Then, three experts are used to manually correct the annotations of the top most section headings. As an advantage, this was the first work that treats section detection as a token-based sequential labeling task and outperforms sentence-based formulation and dictionary-based approaches. This method has an integrated solution that avoids the development of heuristics rules to isolate heading from content. Also, layout features improve the results and can recognize section headings that are not appearing in the training set. However, it is difficult to recognize rare or nonstandard topmost section headings. In addition, subsections are not taken into consideration. Furthermore, some section headings are not the topmost in some records. Also, the absence of layout information can decrease the recall.

The approach of Edinger et al. [35] is to identify sections in medical documents and use them in queries for information retrieval. To do this, the authors prepared a list of variations of all section titles for each document type. Variations in terminology, punctuation, and spelling were selected to identify the most common section titles using a set of documents for each type. To identify titles in a document, a simple exact search is applied to find them by their variations. Then, the headings are annotated and the document text has been segmented according to these headings. As an advantage, the use of sections in the query instead of searching the whole document increased the accuracy of the search. Thus, this method can avoid the retrieval of irrelevant documents. However, it has a smaller recall than the other method that can retrieve more relevant documents. In addition, the exact search is accurate enough for section detection.

Lupşe and Stoicu-Tivadar [94] have proposed a method that supports prescribing by extracting and structuring information from medical records. The principle of this approach is to detect sections of the text and unify their titles using regular expressions and a set of section titles. Then, it removes empty words and applies the Stemming algorithm to the sections to root the words without touching medical terms. Thus, this method can suggest drugs that match the patient's disease, are not contraindicated and do not conflict with other diseases, treatments, or allergies of the patient. Indeed, this approach reduces medical errors in drug prescriptions and structures the necessary drug information. However, some medical terms are still modified by Stemming. Also, this method is tested only with the Roman language.

Zhang et al. [161] have tried to effectively use temporal information in the text of electronic medical documents to structure them and help medical researchers to examine clinical knowledge and to facilitate computer-aided analysis. This method is based on rules to perform a few successive steps. These steps consist of correcting pronunciation errors, dividing texts according to grammatical rules, describing medical facts and events, and finalizing by processing temporal expressions. However, these texts have little temporal information. In addition, the method gives the same weighting to different words.

Lohr et al. [89] have trained a logistic regression model on manually annotated German clinical discharge summaries, short summaries, and transfer letters to automatically identify sections using BoW statistics as features for each sentence. As an advantage, this method achieves promising results in terms of f1-score. Furthermore, these authors have chosen a set of feasible and relevant categories for annotation. In addition, a sentence was chosen as an annotation unit while it has an appropriate granularity. However, the method does not perform well for categories that barely appear in the corpus.

Lupşe and Stoicu-Tivadar [93] have made a method that consists of homogenizing the sections of drug package inserts by standardizing the section names. At first, the method collects all section names from all drug package inserts and prepares unique and common reference names that represent different kinds of sections. Then, machine learning is used to find the appropriate reference for each section name. Through this method, access to drug information has been improved for better processing. Moreover, this technique can be used in clinical decision applications to provide the necessary data to physicians. Thus, it helps especially new young doctors or those who start a new specialty. Neural network leads to the highest results in the extraction of relevant information and outperforms cosine similarity according to the f1-score metric. Moreover, this model can be generalized to any language or domain. However, this model is appropriate only for records where sections are defined by headings.

Discussion These methods can identify sections using rules which are aimed to precisely detect titles and are appropriately made for the target type of documents. These rules usually take into account the title variation or the context of words to detect the title variants. Also, few rule-based methods use interesting information such as the font size which is very important to detect titles where it is not dependent on the language. The use of the machine learning technique improves well the results by incorporating different features. This technique can detect implicit sections since it usually does not depend on headings. However, the used machine learning models are not optimized enough for this task, where most methods just apply pre-built models and use traditional word embedding. Also, the lack of training data for some types of sections is a big drawback for these models. The rules depend a lot on supplemental resources which contain pre-defined titles to be matched. Also, they lose their value when they do not find the format they target, especially in the absence of titles.

6.3.2 Deep learning-based approaches

This category uses deep learning models such as LSTM and CNN coupled with different word embeddings techniques to annotate words by their sections. Almost all these methods apply a sequence tagging task to annotate sequences of words. A set of annotated training data is needed to train these models where rules can be used to prepare them.

Sadoughi et al. [122] have applied section detection on clinical dictations in real-time. They used automatic speech recognition to transform the speech into plain text. Also, a unidirectional LSTM model, which tracks short and long-term dependencies, is run on the text to annotate its section boundaries using Word2vec vectors to represent the input words. To do this, regular expressions were applied to a set of reports to annotate the headings. Each time, a post-processing task is applied to each section to transform the text into a written report. As an advantage, the post-processing task can become faster with the processing of a complete section each time, instead of re-executing after each dictated word. Thus, the post-processing of the previous section happens in parallel with the dictation of the current section without disturbing the user. Moreover, the post-processor can benefit from the full context of the section during the transformation. Thus, real-time section detection ensures that the medical report is directly usable for other processes after dictation. However, the detection of a section depends only on the words dictated so far without seeing the whole document. This prevents it from exploiting all the information in the document to provide a better quality result.

The method of Chirila et al. [22] consists in supporting the prescription of drugs by structuring and categorizing the text into sections. For this, a machine learning model was trained to associate each part of the text with its appropriate section. According to the results of this method, the accuracy of the CNN-based model is superior especially with uniform name sections. Moreover, this method was applied to the Roman language where there is no dataset in this language with fully structured information. However, the execution time of CNN increases significantly when compared with a model based on Naive Bayesian classification. Moreover, this method is applied only to the Roman language.

Goenaga et al. [43] have tested rules and machine learning-based methods for section identification on Spanish Electronic Discharge Summaries. They have found that the machine learning-based method gives the best results. This method is based on transfer learning using the FLAIR model [1] and generates character embeddings for a sequence of tokens to annotate them by a BiLSTM-CRF model. Indeed, the rules have lower results, especially when an incorrectly marked section affects the surrounding sections even by carefully designing rules which is a time-consuming process. In contrast, the FLAIR method can identify sections even with variations or where headings are absent while it can learn from the headings and the vocabulary inside the sections. Also, training the method on data with more variability is useful to keep obtaining higher efficiency on different types of data. However, the results degrade when testing the trained model on different data and the degradation may be drastically in some cases. Furthermore, errors can be caused by high variability with the lack of training. In addition, implicit and mixed sections are the cause of several errors.

Nair et al. [104] have proposed a method to classify the sentences of i2b2 2010 clinical notes into different major SOAP sections using the BiLSTM model with the fusion of Glove, Cui2Vec, and ClinicalBERT embeddings. As an advantage, the contextual embeddings and the transfer learning provide an efficient solution to this task. Also, the authors have found that 500 sentences per section are a sufficient starting point to achieve high performance. However, they have considered only 4 sections and ignored other sections and sub-sections. Moreover, they have not considered the context sensitivity of clinical sentences.

Discussion Deep learning models can learn well from a training set even with a lower amount of data. Different word embeddings are well used with these models to identify the sections based on the semantic information. Recently, most models tend to use BiLSTM which can deeply sequentially exploit the features in double directions, this is useful to find the sequences which define the beginning and the end of sections. Thus, they can well identify the implicit and explicit sections and are not dependent on the titles. However, these methods are usually targeted for a few types of sections and usually, the granularity level is too limited. Also, it is tiring to adapt these methods to another type of data and the high variability of data negatively affects the effectiveness of the model. Also, these models need to be further developed while many types of features are not explored which provide more useful information instead of just words.

6.3.3 Numerical comparison

Tables 8 and 9 show the results of classical and deep learning-based section detection methods, respectively. The A, P, R, and F symbols refer to Accuracy, Precision, Recall, and F1-score, respectively. This task have no benchmark datasets that can be used for better evaluation and comparison. Also, most of methods use only accuracy metric and not aimed at the same type of sections or granularity level. However, we make an approximate comparison based on the available information. Most of methods are based on manually prepared list of titles as a supplemental resource to match them by rules. Usually, they are useful only for specific pre-defined types of sections and data. For example, the method accuracy of Lupşe and Stoicu-Tivadar [94] is significantly decreased by $\approx 22\%$ when the source of data is changed. In fact, the methods which depend on titles can’t detect implicit sections. Some methods based on machine learning [43, 104] are able to categorize the text into sections without the need of section titles. Most of them use static embedding while others use contextual embedding. The best approaches in terms of F1-score are based on deep learning models such as BiLSTM and CRF [26, 43]. Thus, treating this task as sequence labeling is able to detect implicit sections with higher performance. However, these methods predict only some pre-defined sections and they require a big training set. Furthermore, all these section identification methods are aimed to detect only the top-most sections where the granularity-level does not exceed 2 and not all sections are covered. A higher granularity level can be useful to provide more specific context.

7 Issues related to the entity nature

In this section, we highlight some solutions proposed by state-of-the-art methods to solve problems related to the nature of the medical NER. Thus, we have cited and classified some methods into four categories based on the main problems: Ambiguity, Boundary detection, Name variation, and Composed entities. In the following, we have described each of these 4 various problems and highlighted proposed solutions of some cited methods used to deal with each problem including their limits. Finally, we conclude with a general discussion about these solutions. Table 10 shows a summary of these methods.

7.1 Ambiguity

The ambiguity is when a medical named entity can belong to more than one class depending on the context. For example, the entity “distress” may be a disease or a physiological process according to its context. Thus, we need to recognize the entity by exploring its context. As a special case, the abbreviations are likely to be ambiguous, for example, “EGFR” can be the abbreviation of “Epidermal Growth Factor Receptor” or “Estimated Glomerular

Table 8 Classical section detection methods

Pub.	Dataset	Dataset size	Segment	Method	Features	A	P	R	F
Dai et al. [26]	i2b2 2014 shared task (track 2) discharge summaries, procedural notes, emails between the primary physician and the consultant	521 train, 269 valid, 514 test	Sentence	employ heading strings from terminology to make train set with little manual correction + CRF	terminology + layout + affix + orthographic + lexicon occurrence + semantic	-	96.04%	92.4%	94.19%
Lupṡe et al. [94]	Pagina Farmacistilor prospectuses	1630	Text	regular expressions	set of section titles	90%	-	-	-
Lupṡe et al. [94]	HelpNet prospectuses	3002	Text	regular expressions	set of section titles	66%	-	-	-
Lupṡe et al. [94]	CSID prospectuses	3814	Text	regular expressions	set of section titles	88%	-	-	-
Lee and Choi [78]	Korean discharge summaries of rheumatism patients	50 train, 15 valid, 30 test	Chunk	collect patterns + multiple pattern matchings	text	-	84.1%	88.2%	86%
Lohr et al. [90]	German discharge summaries	1106	Sentence	logistic regression model	BoW statistics	-	82%	84%	82%

Table 9 Deep learning based section detection methods

Pub.	Dataset	Dataset size	Segment	Method	Features	A	P	R	F
Goenaga et al. [43]	clinical reports of long-term hospital discharges	100 train, 100 dev, 100 test	Token	BiLSTM-CRF + FLAIR model fine-tuning	FLAIR character embeddings	-	93.40%	92.55%	93.03%
Nair et al. [104]	i2b2 2010	427 notes: 80% train, 10% valid, 10% test sentences	Sentence	BiLSTM to classify sentences	Glove + Cui2Vec + ClinicalBERT	-	-	-	88.68%
Chirila et al. [22]	Pagina Farmacistilor + HelpNet + CSID prospectuses	8147 prospectuses: 70% train, 30% test sentences	Sentence	uniform section names + CNN	list of title synonymes + word embedding	86.55%	-	-	-
Sadoughi et al. [122]	Medical reports with their parallel ASR hypotheses	9073 train, 575 valid, 597 test	Token	regular expressions to annotate headings + unidirectional LSTM	Word2vec	-	84.4%	70.3%	76.7%

Table 10 Some methods that addresses the issues related to the nature of named entities

Pub.	Ambiguity	Boundary	Name variation	Composed entities
Lei et al. [80]	+ Word segment and section information.	+ Medical dictionary to segment words. – Most of errors are in long entities.	NI	NI
Quimbaya et al. [115]	– Ignore the context and surrounding words.	NI	+ Edit distance, exact and lemmatized matching by a knowledge base.	NI
Xu et al. [155]	+ Category Word2vec, PoS and dependence relations, and semantic correlation knowledge. – Filtering may miss some medical entities.	+ Medical native noun phrases. + Based on knowledge base. – May obtain some inexact entities.	NI	+ All medical native noun phrases.
Ghiasvand and Kate [42]	+ Exact matching of unambiguous words from UMLS.	+ Boundary expansion model trained on UMLS words. + Classify all possible noun phrases. – Noun phrase extraction not always perfect. – There is some non noun phrase entities.	+ Lemma and stem forms as features.	+ Complete parsing to extract all noun phrases. – Automatic noun phrase extraction is not always perfect. – Some entities not belong to noun phrases.
Zhou et al. [168]	+ Word and character embeddings. – Capture the contextual relation on word-level.	– Can’t treat complex entities in phrase-level.	+ Character representation can capture out-of-vocabulary words.	NI

Table 10 (continued)

Pub.	Ambiguity	Boundary	Name variation	Composed entities
Deng et al. [30]	+ Learn contextual semantic information without feature engineering. + BiLSTM can learn the contextual dependences. + CRF can improve the annotation in phrase-level.	+ Ensures the integrity and the accuracy of the entity by bidirectional storage of textual information. + IOB annotation format. + Avoid segmentation errors by character embeddings. – Nested entities results in unclear boundaries.	+ Character embedding.	– Limited by the entity annotation granularity.
Zhao et al. [165]	+ Extract lexical, contextual and syntactic clues. + Fine-tune BERT with BiLSTM-CRF. + Rules contextual embedding using ELMO model.	+ Extract noun phrases in sentence by PoS patterns. + IOB annotation format. – annotation not appropriate for nested entities.	+ Clues-based rules. – Rules not appropriate for other domains.	NI
Li et al. [82]	+ Word2vec is improved by BiLSTM to capture contextual information. + BERT is better and can capture the context without BiLSTM.	+ Relation classification between pair of spans is able to recognize discontinuous entities.	+ ELMo character-level embedding. – Word-level embedding is needed to capture the whole meaning of words.	+ Enumerates and represents all text spans and apply a relation classification.
Sui et al. [132]	+ Interactions between the words, entity triggers and the whole sentence semantics.	NI	+ Entity triggers to recognize entity by cue words. – Manual effort is required to prepare entity triggers.	+ Cast the problem into a graph node classification task.

The abbreviation “NI” in this table means Not Included. The signs “+” and “-” mean advantage and inconvenient, respectively.

Filtration Rate”. Thus, an abbreviation mostly may have different meanings depending on the context. As common solutions, the studies try to enrich context information by word or character embedding, knowledge base, and word position in the text such as section, surrounding words, PoS, etc. Also, they try to capture the contextual dependency and relation. Lei et al. [80] have reached a higher performance by merging the word segmentation and the section information. Ghiasvand and Kate [42] have benefited from UMLS which provides a lot of entity terms that are declared as unambiguous. Thus, they do an exact matching for these terms to annotate a maximum number of unambiguous entities to partially solve the ambiguity. Xu et al. [155] have benefited from more context to solve the ambiguity problem by using categories representation by Word2vec, PoS, dependency relation, and semantic correlation knowledge. However, the method may miss some medical entities in the non-medical terms filtering step. The method of Zhou et al. [168] can solve the ambiguity problem by making two types of embeddings for more context, which are C-ELMo for word-level features and C-Flair for character-level. Likewise, the relationship between word-level contextual features can be captured by the C-ELMo model. However, this method fails to detect the boundary of complex phrase-level entities. Deng et al. [30] have used character embedding with BiLSTM-CRF to avoid feature engineering by learning the semantic information in the context. Thus, BiLSTM can provide more comprehensive contextual information and easily learn about contextual dependencies. Moreover, the CRF optimizes the result from the sentence level. However, this method is restricted by the entity labeling granularity where we can find some nested entities. The method of Zhao et al. [165] avoids the ambiguity as it automatically propagates some seed rules based on lexical or contextual clues which are strong indicators of entity recognition. In addition, the authors have fine-tuned a pre-trained contextual embedding model BERT in the biomedical domain. Also, they used a pre-trained contextual embedding model ELMO to give an average embedding for each rule to estimate the semantic relatedness between rules. Li et al. [82] have tested Word2vec with the help of BiLSTM to improve the results by capturing the contextual information. Indeed, BERT embedding alone is more effective than Word2vec and ELMo and even it does not need BiLSTM since it has already captured the contextual information. The method of Sui et al. [132] is based on the interactions among the words, entity triggers, and the whole sentence semantics to recognize the entity from its context.

7.2 Boundary detection

A method can recognize a part of a named entity and fail to determine its exact words. Thus, it can miss some words from the full named entity, or may add some surrounding words. For example, the entity “congenital heart disease” may be recognized partially by detecting only “heart disease”. Thus, the beginning and ending positions of the entity are named by the named entity boundary, where boundary detection is known as an important challenge. The most popular solution to this problem is the extraction of noun phrases while the most of named entities are noun phrases or overlapping with them [162]. Also, the sequence tagging with the IOB annotation format enables learning of the boundaries. Lei et al. [80] use a Chinese medical dictionary as a knowledge source for word segmentation. Indeed, most errors appear in long entities. Xu et al. [155] have extracted medical native noun phrases in a boundary detection step. In addition, they have exploited a knowledge-driven method to detect boundaries, by mapping text to concepts in offline and online lexical resources. Thus, the recognition performance is significantly improved. However, this method may still obtain inexact entities which show some decline in precision and recall. Ghiasvand and Kate [42] have trained a classifier by the medical terms found in UMLS to learn how to expand the boundary of words. Thus, the classifier is applied to all noun phrases in which the detected entity occurs to select the entity with the highest score. However, automatically obtaining noun phrases can make mistakes. Also, sometimes we may find named entities that are not noun phrases. To avoid incorrect identification for the entity boundary, Deng et al. [30] has ensured the integrity and accuracy of the named entity by the bidirectional storage of text information while the IOB labeling method is used. Also, they have used character-level embedding which can avoid poor segmentation. However, the phenomenon of nesting entities leads to an unclear definition of boundary and results in poor accuracy. Zhao et al. [165] have extracted all noun phrases from each sentence as candidate entity mentions based on a set of PoS patterns. Also, they are based on the BiLSTM-CRF model for an IOB labeling. The method of Li et al. [82] applies a relation classification on each pair of candidate entity fragments to determine if it is a discontinuous entity or not. However, BiLSTM-CRF detects entity boundaries more accurately than AAGCN by using label dependence.

7.3 Name Variation

A named entity can be written in different forms by adding and deleting some characters, changing the order of its component words, or changing some words by synonyms. Also, we may have typos in a narrative text written by humans. For example, the entity "left atrium dilation" can be written in another form such as "left atrium dilated", where it even changed from a noun phrase to a verbal phrase. Another example, the entity "Lung Diseases, Obstructive" can be written as "Obstructive Lung Disease", where there is a word order and syntactical variation in the entity. Thus, an exact search for named entities can not cover all the forms of named entities. As a common solution, the preprocessing step is often used to transform similar words into a unified form. Also, considering the surrounding words or characters may be useful to determine the named entity. Another solution is to use character embedding. The method of Quimbaya et al. [115] can solve the variation of named entities using exact, fuzzy, and lemmatized matching by a knowledge base. However, it can not take into consideration the context and the words surrounding the named entity, which could cause more ambiguity. In the work of Ghiasvand and Kate [42], the lemmatize and stemming form of the words surrounding the entity in addition to other features are fed to a decision tree-based classifier. Thus, that can tackle the variability problem. The method of Deng et al. [30] is based on the character embedding which can avoid the variability problem while it is not restricted by a vocabulary of words. Zhou et al. [168] can solve the variability by using character embedding to handle out-of-vocabulary words. Zhou et al. [165] can avoid the name variation problem where they define different types of rules which consider the lexical, contextual, and syntax information based on the clues to find entities. However, some rules may not be applicable due to the mismatch between the training set and a different dataset. Li et al. [82] have tested the ELMo character-level embedding which can represent out-of-vocabulary words. However, the characters can't capture the whole meaning of words and should be merged with word-level embedding. Sui et al. [132] have added entity triggers to help the model recognize the entity by the surrounding cue words. However, this method requires manual effort by annotators to annotate a large group of words to prepare entity triggers.

7.4 Composed entity

The named entity can be composed of multiple words and even can be a long phrase. Consequently, we may find a nested named entity. For example, "excision of ulcer of stomach" is an entity of type procedure which contains "ulcer" as a nested entity of type disorder. Thus, the granularity level should be considered to recognize all the named entities that can be found in one longer named entity. To solve this problem, some work extract all possible noun phrases. Xu et al. [155] can handle nested entities by identifying entity candidates based on the dependency relationships between words. Thus, medical native noun phrases, such as single nouns and maximum noun phrases, are extracted. The method of Ghiasvand and Kate [42] can detect nested entities by obtaining all noun phrases, with nested ones, using a full parsing. The method of Li et al. [82] enumerates and represents all possible text spans [91] to recognize the overlapping entities. Thus, a relation classification is applied to judge whether a pair of entity fragments is overlapping or succession. Sui et al. [132] have proposed a cost-effective and efficient trigger-based graph neural network to cast the problem into a graph node classification task.

7.5 Discussion

Generally, to reduce the ambiguity, most methods [82, 132, 165] have included the context during the recognition and enriched the features by using domain-specific embedding and some other information. For the boundary detection, some methods [42, 155] have considered it as a separate step to correct the boundary of the recognized entities. Most methods [30, 82, 165] have used the combination of BiLSTM-CRF to perform a sequence tagging task by following the IOB annotation style to learn the entity boundary sequentially and take the context well into account. However, this solution is not appropriate for discontinuous and nested entities. Some recent methods [82] have proposed a new technique to transform the sentence into a graph of spans where a span can be a part of an entity. The variation is usually solved by using the context [132, 165], by preprocessing [42, 155] the text or by a character-level annotation [30, 168]. The character-level annotation can ignore the whole meaning of the word while preprocessing can cause ambiguity. Thus, the best solution is to use the context to recognize the entities. The detection of nested entities is mostly ignored, and sequence tagging-based methods with the IOB annotation are not appropriate for them. To detect these entities, Most methods [42, 155] are based on sentence parsing to extract all possible noun phrases. But automatic parsing remains not perfect. Also, few named entities are not overlapping with noun phrases. Some recent methods have proposed an interesting technique to deal with them by performing a spans graph tagging [82] rather than tokens sequence tagging. Thus, transforming a sentence into a graph instead of a sequence is an interesting direction to solve boundary detection and nested entity recognition. While using well the context and the entity triggers can solve the ambiguity and the name variation problems.

8 Limitations and challenges in the field

In this section, we summarize the general limitations and challenges in the medical field by focusing on some different axes such as the used techniques, deep learning models, training and supplemental resources, used features, languages, and social media.

8.1 Techniques

Recently, the studies have focused on deep learning and especially on the use of the sequence tagging models and the capture of the contextual dependency [43, 97, 104, 159]. Indeed, some recent methods turned to the use of graph techniques based on deep learning to better analyse the text in a more suitable manner [34, 132]. Although rule-based and dictionary-based approaches are more precise in some specific cases, they are limited and usually aimed to specific type of data and predefined cases [10, 21, 89]. Also, they need more manual effort to be prepared. Some solutions are proposed to make easily adaptable methods, but deep learning is still needed for better performance. Generally, different ways are proposed to exploit sufficient context during the analysis since the context is needed to handle with the difficulties of the medical terminology such as variability and ambiguity. Thus, deep learning based approaches represent the best direction and different techniques could be explored, especially the graph based models which can better exploit different type of useful relations. Thus, rule-based and dictionary-based techniques can be further enhanced to support the deep learning models.

8.2 Deep learning

Recently, we can see that the machine learning methods are well studied and especially which are based on deep learning [53, 97] where different models are tested with several types of features. Generally, CRF is the most popular model which is a sequence tagging model while the problems are mostly defined as token labeling, especially in the named recognition task. Recent work have shown that the fusion between BiLSTM and CRF gives better results [30, 126, 165], because BiLSTM can consider the order from double directions which makes it able to well understand a sequence of features. Hence, CRF can perform accurate labeling using the features provided by BiLSTM. However, the sequence tagging models are not able to handle the nested and the discontinuous entities. Li et al. [82] have adapted the GCN model by transformed the sentence into graph and performed a node (i.e. word) classification in order to fill this gap. Thus, the named entity recognition can be performed in a clearer and more appropriate way even if the sequential order is not well exploited. Also, by using GCN, the relation extraction task is able to exploit well the relations between words, especially the dependency relations [97].

8.3 Training data

The lack of data is one of the most important problems in this field. There is many languages and types of data, and the medical terminology is evolving day after day. Also, there is many type of information to be extracted. All these facts make the training data and supplemental resources not permanently useful. This problem is a big obstacle since the methods can’t be easily adaptable to other type of data. Usually the required data is prepared manually by medical expert. In fact, the most used supplemental resources are UMLS [15] and SNOMED-CT [32] that can cover a lot of concepts, languages, semantic types, and much information about terms, which make it possible to extract more information using some matching techniques or even partially annotate a raw text. However, they can’t cover everything especially with this constantly evolving field. However, some methods tried to reduce the manual effort to prepare high-quality data [42, 70, 165]. For example, Xu et al. [155] update supplemental resources by using a search engine online. Some methods expand or propagate few manually prepared data [26, 42, 108, 165] or fuzzily exploit them [5, 51, 133].

However, all these methods give lower results compared to others which use manually prepared data. Thus, providing high-quality data with the minimum of manual effort remains to be a big challenge. Furthermore, most datasets which contain medical articles are available only with annotated titles and abstracts. Recently, some researchers [56, 70] are trying to construct new annotated datasets with full-text articles, especially for NER, to provide more detailed information. These articles are usually collected from PMC [118]. Also, providing data in other format than raw text would be useful for many tasks to exploit other attributes such as formatting style [9].

8.4 Features

Most recent studies focus on distributed representation for words or characters such as BERT, Flair, and Word2vec, which can well extract semantic information for the machine learning methods. Furthermore, the models which generate these embeddings can be used in different manners, where they can be pre-trained from other sources, trained from scratch during the whole model training, or fine-tuned. Especially the pre-trained and fine-tuned contextual language models such as BERT have achieved state-of-the-art performance on many natural language processing tasks. For example, the work of Yang et al. [159] has shown that clinical pre-trained transformers achieve better performance for RE. In addition, other types of features can be beneficial such as knowledge and syntactic features. Indeed, there is some information that is not well exploited although they have shown a good potential to improve the IE tasks. Generally, these tasks can be improved by giving more context. Indeed, the study of Lei et al. [80] has proved that section information can improve the entity recognition task, but it is not well exploited in this field. In addition, Tran and Kavuluru [140] have used sub-headings to improve the RE. Furthermore, the formatting style of the document can be very important, while it has proved its ability to detect section titles in the study of Beel et al. [9] using only font-size information. A medical document is generally created in PDF format which provides more useful information than raw text. However, all available benchmark datasets are provided only as annotated raw text.

8.5 Languages

A lot of studies in the medical IE are destined for the Chinese language. Although the English language is easier for many tasks, the Chinese researchers are trying to improve the medical IE for them while they are very interested in the evolution of the medical field generally. However, the Chinese language in the medical field is more difficult compared to English especially in the segmentation while it has complicated syntax rules and a lack of Chinese data. Hence, many studies are destined to deal with this language problem [30, 51, 155]. The segmentation is generally used to provide samples and extract features from them. It can be performed on word, phrase, sentence, or section level. Indeed, Deng et al. [30] found that making features for a sequence of characters is more suitable, especially for the Chinese language. However, character embedding can't capture the whole meaning of words [30, 168].

8.6 Social media

The social media can be a very critical area for the medical information extraction, where the data is noisier and may contain many grammatical mistakes and especially false information by which people can be influenced. Some recent work are aimed at social media [66, 126], for example, Shi et al. [126] have used machine learning models with the help of Concern Graph to apply entity recognition on pandemic concern entities in Twitter. Recently, social media have played an important role during the COVID-19 pandemic period and it is crucial to automatically understand and supervise a lot of people's interactions. Indeed, social media is a rich source of information that mostly contains unstructured and confusing textual and other multimedia data. Thus, some studies are applied to extract information from that data to perform some tasks such as associating tags to posts [75], identifying relevant information [95], etc. It is worth noting that the graph of users' relations such as the following relations can be exploited too to enhance the medical IE. Thus, some important tasks such as community detection and influence identification [46, 59, 68, 103] can be combined with medical IE tasks in social media.

9 Conclusion and future research directions

In this paper, we have introduced an up-to-date survey about the medical information extraction field. We have made a comprehensive review of methods that are aimed at different tasks which are: section detection, entity recognition, and relation extraction. In addition, we have shown the impact of the data on these methods and we have discussed the nature of medical data and information about useful resources and datasets for this field. Also, we have compared numerically some interesting existing approaches based on their published results to get a high level of comparison. Thus, we have shown the directions in which they succeed and eventually fail.

The IE in the medical field is very interesting especially to find information about diseases. Generally, the tasks of this field are aimed to explore knowledge, support persons to find relevant information, and help doctors to release the best decision, for example, choosing the right treatment, making the appropriate drug prescription, or discovering the causes and effects of some diseases. Also, extracting information from EHR have a positive impact on medical clinical practice and investigation in many fields such as the cardiology field [62]. A large amount of unstructured medical textual information is terrible to be manually analyzed by doctors while is considered a heavy treasure of information. In our survey, we conclude that the rule-based and hybrid methods are generally the promising techniques for IE where they have shown the best results. However, the rules depend highly on a specific domain. Thus, it is difficult to adapt these methods to a new type of data since a manual effort by domain experts is needed. Thus, generating rules that dynamically adapt to a new data type is a promising direction. For the deep learning approach, most methods focus on the combination of CRF and BiLSTM models which is very beneficial for sequence-tagging tasks. However, the CRF model is usually used for flat NER, which is not appropriate for nested and discontinuous named entities. Recently, to solve this problem, another technique is adopted which is based on transforming the sentence into a graph to annotate its nodes by a GCN model. Domain-specific embeddings, especially the ones provided by the BERT model, are used by many methods and give better results. There is a lack of medical data where these data contain more privacy compared to other fields. It is a challenging issue to find an easier way to provide high-quality data for training and other additional data which can cover the new medical terms, several types of data, all possible cases, multiple languages, different types of labels, etc. Also, all datasets are available in the textual format while the formatting style is very important and should be more exploited. Indeed, it adds very useful information to the text which is especially used to understand the structure of the document and even the meaning of words.

Indeed, section detection is a challenging issue and showed a positive impact on the performance of several IE tasks. Mainly, the position of a concept in a document can provide more contextual information. However, this task is not well covered by the research work and methods. Thus, benchmark datasets should be constructed for it. Another issue is combining rule-based, dictionary-based, and deep-learning approaches to well benefit from them in one hybrid method. Many ideas have been proposed and can be exploited further. Indeed, we can use rules to prepare data for machine learning or we can use machine learning to generate rules. Besides, we can ameliorate the dictionary matching by machine learning or we can annotate a lot of data with a dictionary to train a machine learning model. Also, we can use the rules by constructing regular expressions to perform a dictionary matching or we can use a dictionary as a supplemental resource to support the rules. In addition, we can make features by using a dictionary and rule-based techniques.

Document summarization is another very challenging task. This task consists in making a text summarization that contains only the most important information. It can depend on the NER where this idea is already used by Sandhya and Kantesaria [123] in a non-medical field. For that, as a first step, they have identified ordinary entities such as persons, organizations, places, time and measurement, etc. Then, they have used them to select the most important words which would be useful to make a document summary. In the medical field, medical entities can be used too to support the medical document summarization task. Hence, we can easily recognize a relevant and a very reduced readable part of the text in a document instead of reading a whole text. Thus, the useless part of the text can be eliminated even for other tasks of IE. Due to the diversity and the growing quantity of medical information, people need to quickly assimilate and determine the content of a medical document. Thus, document summarization helps persons to quickly determine the main points of a document. However, this research field has not yet reached maturity, since a variety of challenges still need to be overcome such as handling large-scale data, providing sufficient annotated data, etc.

Another important issue and possible research direction, which has been discovered especially during the COVID-19 pandemic, is about analyzing the propagated medical information on social media. The content on social media is very different from documents especially when the information is written by normal users and not medical experts. Thus, natural language processing will be much more difficult while we can find unstructured texts and many typos. As well, we can easily find a big number of users influenced by false medical information which represents a critical problem. Likewise, the social media environment is rich in useful information more than just a document. Therefore, we can easily benefit from the reactions to the post, the owner profile, contacts, etc. Thus, IE in social media is very challenging than in medical documents and is needed to detect the spread of false information and understand the people's interactions with medical information.

Availability of data and materials Not applicable.

Declarations

Conflict of Interests The authors declare that they have no competing interests.

References

1. Akbik A, Bergmann T, Blythe D et al (2019) FLAIR: an easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 Conference of the north american chapter of the association for computational linguistics (Demonstrations), pp 54–59
2. Alex B, Grover C, Tobin R et al (2019) Text mining brain imaging reports. J Biomed Semant 10(1):1–11
3. Angeli G, Premkumar MJJ, Manning CD (2015) Leveraging linguistic structure for open domain information extraction. In: Proceedings of the 53rd Annual meeting of the association for computational linguistics and the 7th International joint conference on natural language processing (vol 1: Long Papers), pp 344–354
4. Apostolova E, Channin DS, Demner-Fushman D et al (2009) Automatic segmentation of clinical texts. In: 2009 Annual international conference of the IEEE engineering in medicine and biology society, IEEE, pp 5905–5908
5. Arbabi A, Adams DR, Fidler S et al (2019) Identifying clinical terms in medical text using Ontology-Guided machine learning. JMIR Med Inform 7(2):e12,596
6. Aronson AR, Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 17(3):229–236
7. Aydar M, Bozal O, Ozbay F (2020) Neural relation extraction: a survey. arXiv e-prints pp arXiv–2007
8. Batista DS (2018) Named-Entity evaluation metrics based on entity-level. http://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation
9. Beel J, Gipp B, Shaker A et al (2010) SciPlore xtract: extracting titles from scientific PDF documents by analyzing style information (font size). In: International conference on theory and practice of digital libraries, Springer, pp 413–416
10. Ben Abdessalem Karaa W, Alkhammash EH, Bchir A (2021) Drug disease relation extraction from biomedical literature using NLP and machine learning. Mob Inf Syst, p 2021
11. Berrazega I (2012) Temporal information processing: a survey. Int J Naturel Lang Comput 1(2):1–14
12. Bethard S, Savova G, Chen WT et al (2016) Semeval-2016 task 12: Clinical tempeval. In: Proceedings of the 10th International workshop on semantic evaluation (SemEval-2016), pp 1052–1062
13. Bethard S, Savova G, Palmer M et al (2017) SemEval-2017 task 12: Clinical TempEval. In: Proceedings of the 11th International workshop on semantic evaluation (SemEval-2017). Association for computational linguistics, Vancouver, Canada, pp 565-572. https://doi.org/10.18653/v1/S17-2093
14. Bhatia P, Celikkaya B, Khalilia M (2019) Joint entity extraction and assertion detection for clinical text. In: Proceedings of the 57th Conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, vol 1: Long Papers. Association for computational linguistics, pp 954–959. https://doi.org/10.18653/v1/p19-1091
15. Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(suppl_1):D267–D270
16. Bottou L (1999) On-line learning and stochastic approximations. Cambridge University Press, USA, pp 9–42
17. Bramsen P, Deshpande P, Lee YK et al (2006) Finding temporal order in discharge summaries. In: AMIA annual symposium proceedings, american medical informatics association, p 81
18. Carrell DS, Halgrim S, Tran DT et al (2014) Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. Am J Epidemiol 179(6):749–758
19. Chapman W, Dowling J, Chu D (2007) ConText: an algorithm for identifying contextual features from clinical text. In: Biological, translational, and clinical language processing, pp 81–88
20. Chapman WW, Savova GK, Zheng J et al (2012) Anaphoric reference in clinical reports: characteristics of an annotated corpus. J Biomed Inform 45(3):507–521
21. Chirila OS, Chirila CB, Stoicu-Tivadar L (2019) Named entity recognition and classification for medical prospectuses. Stud Health Technol Inform 262:284–287
22. Chirila OS, Chirila CB, Stoicu-Tivadar L (2019) Improving the prescription process information support with structured medical prospectuses using neural networks. Stud Health Technol Inform 264:353–357
23. Cohen KB, Lanfranchi A, MJy Choi et al (2017) Coreference annotation and resolution in the colorado richly annotated full text (CRAFT) corpus of biomedical journal articles. BMC Bioinforma 18(1):1–14
24. Cohen KB, Verspoor K, Fort K et al (2017) The colorado richly annotated full text (craft) corpus: Multi-model annotation in the biomedical domain. In: Handbook of linguistic annotation. Springer, pp 1379–1394
25. Dai X, Karimi S, Hachey B et al (2020) An effective transition-based model for discontinuous NER. arXiv:200413454

26. Dai HJ, Syed-Abdul S, Chen CW et al (2015) Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. BioMed Research International, p 2015
27. De Bruijn B, Cherry C, Kiritchenko S et al (2011) Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 18(5):557–562
28. Del Corro L, Gemulla R (2013) Clausie: clause-based open information extraction. In: Proceedings of the 22nd international conference on World Wide Web, pp 355–366
29. Deléger L, Névéol A (2014) Automatic identification of document sections for designing a french clinical corpus (identification automatique de zones dans des documents pour la constitution d'un corpus médical en français) [in french]. In: TALN
30. Deng N, Fu H, Chen X (2021) Named entity recognition of traditional chinese medicine patents based on BiLSTM-CRF. Wirel Commun Mob Comput, p 2021
31. Devlin J, Chang M, Lee K et al (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, vol 1 (Long and Short Papers). Association for computational linguistics, pp 4171–4186. https://doi.org/10.18653/v1/n19-1423
32. Donnelly K (2006) SNOMED-CT: the advanced terminology and coding system for ehealth. Stud Health Technol Inform 121:279
33. Doğan RI, Leaman R, Lu Z (2014) NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 47:1–10
34. drissiya El-allaly E, Sarrouti M, En-Nahnahi N et al (2022) An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation. J Biomed Inform 125(103):968
35. Edinger T, Demner-Fushman D, Cohen AM et al (2017) Evaluation of clinical text segmentation to facilitate cohort retrieval. In: AMIA Annual symposium proceedings, american medical informatics association, p 660
36. Elhadad N, Pradhan S, Gorman S et al (2015) SemEval-2015 task 14: Analysis of clinical text. In: Proceedings of the 9th International workshop on semantic evaluation (SemEval, vol 2015, pp 303–310
37. Eriksson R, Jensen PB, Frankild S et al (2013) Dictionary construction and identification of possible adverse drug events in danish clinical narrative text. J Am Med Inform Assoc 20(5):947–953
38. Fader A, Soderland S, Etzioni O (2011) Identifying relations for open information extraction. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp 1535–1545
39. Ford E, Carroll JA, Smith HE et al (2016) Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 23(5):1007–1015
40. Fundel K, Küffner R, Zimmer R (2007) RelEx—Relation extraction using dependency parse trees. Bioinformatics 23(3):365–371
41. Garvin JH, DuVall SL, South BR et al (2012) Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc 19(5):859–866
42. Ghiasvand O, Kate RJ (2018) Learning for clinical named entity recognition without manual annotations. Inform Med Unlocked 13:122–127
43. Goenaga I, Lahuerta X, Atutxa A et al (2021) A section identification tool: Towards HL7 CDA/CCR standardization in spanish discharge summaries. J Biomed Inf 121(103):875
44. Grishman R, Sundheim BM (1996) Message understanding conference-6: A brief history. In: COLING 1996 vol 1: The 16th International conference on computational linguistics
45. Guo F, He R, Dang J (2019) Implicit discourse relation recognition via a BiLSTM-CNN architecture with dynamic chunk-based max pooling. IEEE Access 7(169):281–169,292
46. Hafiene N, Karoui W, Romdhane LB (2020) Influential nodes detection in dynamic social networks: A survey. Exp Syst Appl 159(113):642
47. Hahn U, Oleynik M (2020) Medical information extraction in the age of deep learning. Yearb Med Inform 29(01):208–220
48. Hallersten A, Fürst W, Mezzasalma R (2016) Physicians prefer greater detail in the biosimilar label (SmPC)–results of a survey across seven european countries. Regul Toxicol Pharmacol 77:275–281
49. Hasan F, Roy A, Pan S (2020) Integrating text embedding with traditional NLP features for clinical relation extraction. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, pp 418-425
50. Haug PJ, Wu X, Ferraro JP et al (2014) Developing a section labeler for clinical documents. In: AMIA Annual symposium proceedings, american medical informatics association, p 636
51. He S, Sun D, Wang Z (2022) Named entity recognition for chinese marine text with knowledge-based self-attention. Multimed Tool Appl 81(14):19,135–19,149

52. Henry S, Buchan K, Filannino M et al (2020) 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc 27(1):3–12
53. Hong WS, Haimovich AD, Taylor RA (2018) Predicting hospital admission at emergency department triage using machine learning. PloS one 13(7):e0201,016
54. Honnibal M, Montani I (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, to appear
55. Hsu W, Han SX, Arnold CW et al (2015) A data-driven approach for quality assessment of radiologic interpretations. J Am Med Inform Assoc 23(e1):e152–e156
56. Islamaj R, Leaman R, Kim S et al (2021) NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. Sci Data 8(1):1–12
57. Jagannatha A, Liu F, Liu W et al (2019) Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 42(1):99–111
58. Jancsary J, Matiasek J, Trost H (2008) Revealing the structure of medical dictations with conditional random fields. In: Proceedings of the 2008 Conference on empirical methods in natural language processing, pp 1–10
59. Jaouadi M, Romdhane LB (2019) Influence maximization problem in social networks: An overview. In: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), IEEE, pp 1–8
60. Jelier R, Jenster G, Dorssers LC et al (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics 21(9):2049–2058
61. Johnson AE, Pollard TJ, Shen L et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3(1):1–9
62. Jonnalagadda SR, Adupa AK, Garg RP et al (2017) Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials. J Cardiovasc Transl Res 10(3):313–321
63. Karlsson I, Boström H (2016) Predicting adverse drug events using heterogeneous event sequences. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, pp 356–362
64. Kim Y, Heider PM, Lally IR et al (2021) A hybrid model for family history information identification and relation extraction: Development and evaluation of an End-to-End information extraction system. JMIR Med Inform 9(4):e22,797
65. Koleck TA, Dreisbach C, Bourne PE et al (2019) Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 26(4):364–379
66. Komariah KS, Shin BK (2021) Medical entity recognition in twitter using conditional random fields. In: 2021 International Conference on Electronics, Information, and Communication (ICEIC), IEEE, pp 1–4
67. Komninos A, Manandhar S (2016) Dependency based embeddings for sentence classification tasks. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1490–1500
68. Kouni IBE, Karoui W, Romdhane LB (2021) WLNI-LPA: detecting overlapping communities in attributed networks based on label propagation process. In: Proceedings of the 16th International conference on software technologies, ICSOFT 2021, Online Streaming, July 6-8, 2021. SCITEPRESS, pp 408–416. https://doi.org/10.5220/0010605904080416
69. Kreuzthaler M, Schulz S (2015) Detection of sentence boundaries and abbreviations in clinical narratives. BMC Medical Inform Decis Mak 15:S4–S4
70. Kroll H, Pirklbauer J, Ruthmann J et al (2020) A semantically enriched dataset based on biomedical NER for the COVID19 open research dataset challenge. arXiv:2005.08823
71. Kropf S, Krücken P, Mueller W et al (2017) Structuring legacy pathology reports by openEHR archetypes to enable semantic querying. Method Inform Med 56(03):230–237
72. Kumar S (2017) A survey of deep learning methods for relation extraction. arXiv:170503645
73. Lai KH, Topaz M, Goss FR et al (2015) Automated misspelling detection and correction in clinical free-text records. J Biomed Inform 55:188–195
74. Lan M, Wang J, Wu Y et al (2017) Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In: Proceedings of the 2017 Conference on empirical methods in natural language processing, pp 1299–1308
75. Landolsi MY, Mohamed HH, Romdhane LB (2021) Image annotation in social networks using graph and multimodal deep learning features. Multimed Tools Appl 034(8):12,009–12

76. Laparra E, Su X, Zhao Y et al (2021) SemEval-2021 task 10: Source-free domain adaptation for semantic processing. In: Proceedings of the 15th International workshop on semantic evaluation (SemEval-2021). 348–356
77. Laparra E, Xu D, Elsayed A et al (2018) SemEval 2018 task 6: Parsing time normalizations. In: SemEval@ NAACL-HLT, pp 88–96
78. Lee W, Choi J (2018) Temporal segmentation for capturing snapshots of patient histories in korean clinical narrative. Healthc Inform Res 24(3):179–186
79. Lee J, Yoon W, Kim S et al (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
80. Lei J, Tang B, Lu X et al (2014) A comprehensive study of named entity recognition in chinese clinical text. J Am Med Inform Assoc 21(5):808–814
81. Leroy G, Chen H (2001) Filling preposition-based templates to capture information from medical abstracts. In: Biocomputing 2002. World Scientific. 350–361
82. Li F, Lin Z, Zhang M et al (2021) A Span-Based model for joint overlapped and discontinuous named entity recognition. arXiv:2106.14373
83. Li Y, Lipsky Gorman S, Elhadad N (2010) Section classification in clinical notes using supervised hidden markov model. In: Proceedings of the 1st ACM International health informatics symposium, pp 744–750
84. Li W, Shi S, Gao Z et al (2018) Improved deep belief network model and its application in named entity recognition of chinese electronic medical records. In: 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA), IEEE, pp 356-360
85. Li J, Sun Y, Johnson RJ et al (2016) BioCreative v CDR task corpus: a resource for chemical disease relation extraction. Database, p 2016
86. Liu F, Chen J, Jagannatha A et al (2016) Learning for biomedical information extraction: Methodological review of recent advances. arXiv:1606.07993
87. Liu Y, Ott M, Goyal N et al (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692
88. Liu Y, Wei L, Yao Z et al (2016) The practice and experience of emergency information system construction. Chin Digit Med 11(5):53–55
89. Lohr C, Luther S, Matthies F et al (2018) CDA-compliant section annotation of german-language discharge summaries: Guideline development, annotation campaign, section classification. In: AMIA 2018, American medical informatics association annual symposium, San Francisco, CA, November 3-7, 2018. AMIA
90. Lohr C, Luther S, Matthies F et al (2018) CDA-compliant section annotation of german-language discharge summaries: guideline development, annotation campaign, section classification. In: AMIA Annual symposium proceedings, american medical informatics association, p 770
91. Luan Y, Wadden D, He L et al (2019) A general framework for information extraction using dynamic span graphs. In: Proceedings of the 2019 Conference of the north american chapter of the association for computational linguistics: Human language technologies, vol 1 (Long and Short Papers). Association for computational linguistics, Minneapolis, Minnesota, pp 3036–3046. https://doi.org/10.18653/v1/N19-1308
92. Ludwick DA, Doucette J (2009) Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 78(1):22–31
93. Lupșe O, Stoicu-Tivadar L (2018) Supporting prescriptions with synonym matching of section names in prospectuses. Stud Health Technol Inform 251:153–156
94. Lupșe O, Stoicu-Tivadar L (2018) Extracting and structuring drug information to improve e-prescription and streamline medical treatment. Appl Med Inf 40(1-2):7–14
95. Mabrouk O, Hlaoua L, Omri MN (2021) Exploiting ontology information in fuzzy SVM social media profile classification. Appl Intell 51(6):3757–3774
96. Mahendran D, McInnes BT (2021) Extracting adverse drug events from clinical notes. In: AMIA Annual symposium proceedings, american medical informatics association, p 420
97. Mahendran D, Tang C, McInnes B (2022) Graph convolutional networks for chemical relation extraction. In: Proceedings of the semantics-enabled biomedical literature Analytics (SeBiLAn)
98. Manning CD, Surdeanu M, Bauer J et al (2014) The stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60

99. Mausam SM, Bart R et al (2012) Open language learning for information extraction. In: Proceedings of the 2012 Joint conference on empirical methods in natural language processing and computational natural language learning. Association for computational linguistics, USA, EMNLP-CoNLL '12, pp 523–534
100. Mehrabi S, Krishnan A, Roch AM et al (2015) Identification of patients with family history of pancreatic cancer-investigation of an nlp system portability. Stud Health Technol Inform 216:604
101. Mercorelli L, Nguyen H, Gartell N et al (2022) A framework for de-identification of free-text data in electronic medical records enabling secondary use. Australian Health Review
102. Meystre SM, Savova GK, Kipper-Schuler KC et al (2008) Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inf 17(01):128–144
103. Mnasri W, Azaouzi M, Romdhane LB (2021) Parallel social behavior-based algorithm for identification of influential users in social network. Appl Intell, pp 1–19
104. Nair N, Narayanan S, Achan P et al (2022) Clinical note section identification using transfer learning. In: Proceedings of 6th International congress on information and communication technology, Springer, pp 533–542
105. Nasar Z, Jaffry SW, Malik MK (2021) Named entity recognition and relation extraction: State-of-the-art. ACM Comput Surv (CSUR) 54(1):1–39
106. Nayel HA, ShashrekhaH L (2019) Integrating dictionary feature into a deep learning model for disease named entity recognition. arXiv:1911.01600
107. Neumann M, King D, Beltagy I et al (2019) ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP workshop and shared task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019. Association for computational linguistics, pp 319–327. https://doi.org/10.18653/v1/w19-5034
108. Ni J, Delaney B, Florian R (2015) Fast model adaptation for automated section classification in electronic medical records. Stud Health Technol Inform 216:35–39
109. Peters ME, Neumann M, Iyyer M et al (2018) Deep contextualized word representations. In: Proceedings of the 2018 Conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, vol 1 (Long Papers). Association for computational linguistics, pp 2227–2237. https://doi.org/10.18653/v1/n18-1202
110. Pomares-Quimbaya A, Kreuzthaler M, Schulz S (2019) Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. BMC Med Res Methodol 19(1):155
111. Popejoy LL, Khalilia MA, Popescu M et al (2014) Quantifying care coordination using natural language processing and domain-specific ontology. J Am Med Inform Assoc 22(e1):e93–e103
112. Popovski G, Seljak BK, Eftimov T (2020) A survey of named-entity recognition methods for food information extraction. IEEE Access 8(31):586–31,594
113. Pradhan S, Elhadad N, Chapman WW et al (2014) SemEval-2014 task 7: Analysis of clinical text. In: SemEval@ COLING, pp 54–62
114. Qi P, Zhang Y, Zhang Y et al (2020) Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual meeting of the association for computational linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020. Association for computational linguistics, pp 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14
115. Quimbaya AP, Múnera AS, Rivera RAG et al (2016) Named entity recognition over electronic health records through a combined dictionary-based approach. Procedia Computer Science 100:55–61
116. Ramshaw LA, Marcus MP (1999) Text chunking using transformation-based learning. In: Natural language processing using very large corpora. Springer, pp 157–176
117. Rebholz-Schuhman D, Jimeno-Yepes A, Li C et al (2011) Assessment of NER solutions against the first and second CALBC silver standard corpus. J Biomed Semantics 2(5):1–12
118. Roberts RJ (2001) PubMed central: The GenBank of the published literature
119. Rochefort CM, Buckeridge DL, Forster AJ (2015) Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. Implement Sci 10(1):1–9
120. Rosario B, Hearst MA (2004) Classifying semantic relations in bioscience texts. In: Proceedings of the 42nd Annual meeting of the association for computational linguistics (ACL-04), pp 430–437
121. Rundo L, Pirrone R, Vitabile S et al (2020) Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. J Biomed Inf 108:103,479
122. Sadoughi N, Finley GP, Edwards E et al (2018) Detecting section boundaries in medical dictations: toward real-time conversion of medical dictations to clinical reports. In: International conference on speech and computer, Springer, pp 563–573

123. Sandhya P, Kantesaria ML (2020) Named entity recognition in document summarization. In: Trends and applications of text summarization techniques. IGI Global. 125–149
124. Shen J, Robertson N (2021) Bbas: Towards large scale effective ensemble adversarial attacks against deep neural network learning. Inf Sci 569:469–478
125. Shi J, Li W, Yang Y et al (2021) Automated concern exploration in pandemic Situations-COVID-19 as a use case. In: Pacific rim knowledge acquisition workshop, springer, pp 178–185
126. Shi J, Li W, Yongchareon S et al (2022) Graph-based joint pandemic concern and relation extraction on twitter. Exp Syst Appl 195(116):538. https://doi.org/10.1016/j.eswa.2022.116538
127. Sohrab MG, Duong K, Miwa M et al (2020) BENNERD: a neural named entity linking system for COVID-19. In: Proceedings of the 2020 Conference on empirical methods in natural language processing: System demonstrations, pp 182–188
128. Song HJ, Jo BC, Park CY et al (2018) Comparison of named entity recognition methodologies in biomedical documents. Biomed Eng Online 17(2):1–14
129. Sorgente A, Vettigli G, Mele F (2013) Automatic extraction of cause-effect relations in natural language text. DART@ AI* IA 2013:37–48
130. Stubbs A, Kotfila C, Uzuner Ö (2015) Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. J Biomed Inf 58:S11–S19
131. Stubbs A, Kotfila C, Xu H et al (2015) Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2. J Biomed Inform 58:S67–S77
132. Sui Y, Bu F, Hu Y et al (2022) Trigger-GNN: a Trigger-Based graph neural network for nested named entity recognition. 2204.05518
133. Sun Q, Bhatia P (2021) Neural entity recognition with gazetteer based fusion. In: Findings of the association for computational linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, Findings of ACL, vol ACL/IJCNLP 2021. Association for computational linguistics, pp 3291–3295. https://doi.org/10.18653/v1/2021.findings-acl.291
134. Sun W, Cai Z, Li Y et al (2018) Data processing and text mining technologies on electronic medical records: a review. J Healthcare Eng
135. Sun W, Cai Z, Liu F et al (2017) A survey of data mining technology on electronic medical records. In: 2017 IEEE 19th International conference on e-health networking, applications and services (Healthcom), IEEE, pp 1-6
136. Suominen HJ, Salakoski TI (2010) Supporting communication and decision making in finnish intensive care with language technology. J Healthcare Eng 1(4):595–614
137. Tang B, Cao H, Wu Y et al (2013) Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. In: BMC Medical informatics and decision making, BioMed Central. 1–10
138. Tchraktchiev D, Angelova G, Boytcheva S et al (2011) Completion of structured patient descriptions by semantic mining. In: Patient safety informatics. IOS Press, pp 260–269
139. Tepper M, Capurro D, Xia F et al (2012) Statistical section segmentation in free-text clinical records. In: Lrec, pp 2001–2008
140. Tran T, Kavuluru R (2019) Distant supervision for treatment relation extraction by leveraging MeSH subheadings. Artif Intell Med 98:18–26
141. Tran V, Tran VH, Nguyen P et al (2021) CovRelex: a COVID-19 retrieval system with relation extraction. In: Proceedings of the 16th Conference of the european chapter of the association for computational linguistics: System demonstrations, pp 24–31
142. Uzuner Ö, Solti I, Cadag E (2010) Extracting medication information from clinical text. J Am Med Inform Assoc 17(5):514–518
143. Uzuner Ö, South BR, Shen S et al (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 18(5):552–556
144. Vunikili R, Supriya H, Marica VG et al (2020) Clinical NER using spanish BERT embeddings. In: IberLEF@ SEPLN, pp 505–511
145. Wang L, Foer D, MacPhaul E et al (2021) PASCLex: a comprehensive Post-Acute sequelae of COVID-19 (PASC) symptom lexicon derived from electronic health record clinical notes. J Biomed Inf, p 103951
146. Wang Y, Fu S, Shen F et al (2020) The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. JMIR Med Inform 8(11):e23,375
147. Wang P, Hao T, Yan J et al (2017) Large-scale extraction of drug–disease pairs from the medical literature. J Assoc Inform Sci Technol 68(11):2649–2661
148. Wang X, Hripcsak G, Markatou M et al (2009) Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. J Am Med Inform Assoc 16(3):328–337

149. Wang S, Ren F, Lu H (2018) A review of the application of natural language processing in clinical medicine. In: 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp 2725–2730
150. Wang Y, Wang L, Rastegar-Mojarad M et al (2018) Clinical information extraction applications: a literature review. J Biomed Inform 77:34–49
151. Wei WQ, Feng Q, Jiang L et al (2014) Characterization of statin dose response in electronic medical records. Clin Pharmacol Ther 95(3):331–338
152. Wei Q, Ji Z, Si Y et al (2019) Relation extraction from clinical narratives using pre-trained language models. In: AMIA annual symposium proceedings, American medical informatics association, p 1236
153. Weiskopf NG, Hripcsak G, Swaminathan S et al (2013) Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 46(5):830–836
154. Wu Y, Jiang M, Xu J et al (2017) Clinical named entity recognition using deep learning models. In: AMIA Annual symposium proceedings, american medical informatics association, p 1812
155. Xu J, Gan L, Cheng M et al (2018) Unsupervised medical entity recognition and linking in chinese online medical text. J Healthcare Eng, p 2018
156. Yang Z, Dai Z, Yang Y et al (2019) Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst, p 32
157. Yang J, Han SC, Poon J (2021) A survey on extraction of causal relations from natural language text. arXiv:2101.06426
158. Yang Z, Lin H, Li Y (2008) Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature. Comput Biol Chem 32(4):287–291
159. Yang X, Yu Z, Guo Y et al (2021) Clinical relation extraction using transformer-based models. arXiv:2107.08957
160. Yang X, Zhang H, He X et al (2020) Extracting family history of patients from clinical narratives: exploring an end-to-end solution with deep learning models. JMIR Med Inform 8(12):e22,982
161. Zhang R, Chu F, Chen D et al (2018) A text structuring method for chinese medical text based on temporal information. Int J Environ Res Public Health 15(3):402
162. Zhang S, Elhadad N (2013) Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. J Biomed Inform 46(6):1088–1098
163. Zhang T, Huang Z, Wang Y et al (2022) Information extraction from the text data on traditional chinese medicine: A review on tasks, challenges, and methods from 2010 to 2021. Evidence-Based Complementary and Alternative Medicine
164. Zhang Y, Yan X, Gao X et al (2016) Demand analysis of decision support system of grass-roots health. Chinese Gen Pract 19:2636–2639. https://doi.org/10.3969/j.issn.1007-9572.2016.22.005
165. Zhao X, Ding H, Feng Z (2021) GLaRA: graph-based labeling rule augmentation for weakly supervised named entity recognition. In: Proceedings of the 16th Conference of the european chapter of the association for computational linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021. Association for computational linguistics, pp 3636–3649. https://doi.org/10.18653/v1/2021.eacl-main.318
166. Zheng C, Rashid N, Koblick R et al (2015) Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation. Clin Ther 37(9):2048–2058
167. Zhou J, Fu Bq (2018) The research on gene-disease association based on text-mining of pubmed. BMC bioinformatics 19(1):1–8
168. Zhou Y, Ju C, Caufield JH et al (2021) Clinical named entity recognition using contextualized token representations. arXiv:2106.12608
169. Zweigenbaum P, Deléger L, Lavergne T et al (2013) A supervised abbreviation resolution system for medical text. In: CLEF (Working Notes)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.