Named Entity Recognition and Relation Extraction: State-of-the-Art

ZARA NASAR, SYED WAQAR JAFFRY, and MUHAMMAD KAMRAN MALIK,
PUCIT, University of the Punjab, Lahore, Pakistan

With the advent of Web 2.0, there exist many online platforms that result in massive textual-data production. With ever-increasing textual data at hand, it is of immense importance to extract information nuggets from this data. One approach towards effective harnessing of this unstructured textual data could be its transformation into structured text. Hence, this study aims to present an overview of approaches that can be applied to extract key insights from textual data in a structured way. For this, Named Entity Recognition and Relation Extraction are being majorly addressed in this review study. The former deals with identification of named entities, and the latter deals with problem of extracting relation between set of entities. This study covers early approaches as well as the developments made up till now using machine learning models. Survey findings conclude that deep-learning-based hybrid and joint models are currently governing the state-of-the-art. It is also observed that annotated benchmark datasets for various textual-data generators such as Twitter and other social forums are not available. This scarcity of dataset has resulted into relatively less progress in these domains. Additionally, the majority of the state-of-the-art techniques are offline and computationally expensive. Last, with increasing focus on deep-learning frameworks, there is need to understand and explain the under-going processes in deep architectures.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Information extraction**;

Additional Key Words and Phrases: Information extraction, named entity recognition, relation extraction, deep learning, joint modeling

ACM Reference format:
Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named Entity Recognition and Relation Extraction: State-of-the-Art. *ACM Comput. Surv.* 54, 1, Article 20 (February 2021), 39 pages.
https://doi.org/10.1145/3445965

1 INTRODUCTION

With the advent of Web 2.0, there exist many online platforms that are producing bulk of data on daily basis such as social networks, online blogs, magazines, and so on. This textual data is getting increased day by day and carries potential information insights. These insights could be used for

Authors' addresses: Z. Nasar and S. W. Jaffry, Graduate Block, National Centre of Artificial Intelligence, Punjab University College of Information Technology, University of the Punjab, Allama Iqbal Campus, Lahore, 54000, Punjab, Pakistan; emails: {zara.nasar, swjaffry}@pucit.edu.pk; M. K. Malik, Punjab University College of Information Technology, University of the Punjab, Quaid-e-Azam Campus, Samsani Road, Lahore, 54000, Punjab, Pakistan; email: kamran.malik@pucit.edu.pk.

0360-0300/2021/02-ART20 $15.00
https://doi.org/10.1145/3445965

variety of purposes to serve humanity more effectively. Therefore, it is of immense importance to extract potential information insights from this data. One approach towards effective harnessing of this unstructured textual data could be its transformation into structured text.

Next pertinent question is how this bulk of data can be processed. There are two major approaches to extract key insights from data. First approach is the manual analysis and the second approach is automated analysis that would employ computing technologies to perform text analysis. The latter approach is more appealing due to huge data sizes, as processing huge amount of text in a manual fashion will not be feasible. To automatically extract key information in a brief and concise form, Information Extraction (IE) is used.

IE deals with extraction of information from bulk of data. It primarily refers to extraction of structured information from unstructured or semi-structured text. The task of IE dates back to 1970, when a system named JASPER was proposed to extract certain chunks of information by means of template-driven approach and heuristics [2]. This system takes company press release as an input and tries to extract key insights out of the provided input data. This concept for IE is later extended in Message Understanding Conference in 1987. The problem of IE in general is composed of different sub-problems. This article majorly deals with Named Entity Recognition (NER) and Relation Extraction (RE) problems. NER deals with extraction of named entities, whereas RE deals with extraction of relation between named entities. Consider the following piece of text:

"**CHICAGO** (AP) — Citing high fuel prices, **United Airlines** said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. **American Airlines**, a unit of **AMR**, immediately matched the move, spokesman **Tim Wagner** said. United, a unit of **UAL**, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as **Chicago** to **Dallas** and **Atlanta** and **Denver** to **San Francisco, Los Angeles** and **New York**."

In the above passage, all highlighted tokens correspond to entities where *Tim Wagner* is a person name, *UAL* and *AMR* are organizations, whereas *Chicago, Dallas, Atlanta, Denver,* and so on, are locations. These extracted concepts are related, e.g., it can be clearly seen that phrase "**American Airlines**, a unit of **AMR"** and **"**United, a unit of **UAL"** establishes *UnitOf* relation between *American Airlines, AMR* and *United, UAL*. The process of extraction of these entities and their corresponding relationship is regarded as NER and RE, respectively. These information chunks form the back-bone of IE systems and carry immense importance in advanced Natural Language Processing (NLP) such as Machine Translation (MT), Question Answering (QA) systems, and Event Extraction. In this survey, the state-of-the-art against NER and RE is presented. Each of these sub-tasks of IE is extensively studied in the light of existing literature reviews and recent advancements. Every sub-task is further classified based on its applications. Furthermore, deep learning is being majorly focused in the context of this study; hence, approaches involving neural networks are being treated separately even though they can be treated as either supervised, un-supervised, or semi-supervised approaches depending on the application. This is done as, to the best of our knowledge, there exists no survey that is focused on state-of-the-art in IE in the light of recent advancements in deep learning.

Rest of this article is structured as follows: The background section covers major datasets and techniques that are being employed to solve and evaluate the respective tasks to provide brief overview of domain. Section 3 represents the methodology opted to perform this literature survey. Sections 4–5 briefly explain the state-of-the-art in domains of NER and RE, respectively. Section 6 provides conclusion of this study along with future directions.

2 BACKGROUND

This section includes existing surveys followed by brief introduction to widely used datasets that are being used to perform NER and RE. These existing surveys help in compiling early advancements in the field to present a holistic picture of undergoing advancements.

2.1 Existing IE Surveys

As IE comprises multiple sub-tasks, there exist surveys that are solely focused on IE as well. A survey study presented in Reference [1] is focused on various extraction patterns that are used to perform IE. These patterns exploit semantic, syntactic, and delimiters information. Furthermore, study also compiles various systems that perform IE from free-text and online text. Free-text here refers to plain grammatical English text, whereas online text carries hybrid text with grammatical, telegraphic, and ungrammatical text mixture. Available tools for both textual data types include LIEP, AutoSlog, PALKA, CRYSTAL, WebFoot, and HASTEN for free-text, whereas WHISK, RAPIER, and SRV are briefly explained for online text. Among these, WHISK is the only tool that can handle multiple lines.

Review study presented in Reference [2] decomposes overall task of IE into five major tasks. These tasks include segmentation, NER, RE, normalization, and coreference resolution. Major problems in segmentation include the variety of usage against hyphens, apostrophes, white-spaces, and full-stops. These problems are often resolved by means of employing a rule-base. Oriental languages such as Chinese pose additional challenges as well, which are often resolved by using N-gram-based models and Viterbi algorithm. For NER task, apart from simple rule-based techniques, major machine learning techniques employed in this regard include Support Vector Machines (SVM), Conditional random Fields (CRF), Maximum Entropy Markov Models (MEMM), Hidden Markov Models (HMM), and Decision Tree Classifier (DTC). RE task deals with extraction of relations between set of entities. The widely used learning approaches to perform RE include Markov models such as MEMM and CRF as well as context free grammars. HMMs are not used for this task, as this model is not suitable to capture long-term dependencies. Rule-based approaches can also be applied to perform RE, where incorporation of syntactic information tends to provide more generic rules. Remaining tasks that include normalization and coreference resolution tasks are less generic. This is because both tasks require information that is related to a respective domain of interest. Therefore, normalization task is mostly carried out using conversion rules and regular expressions, carefully designed by domain experts. Coreference resolution task, however, is being achieved by means of variety of approaches. Widely used approaches in this regard include rule-based approaches, DTC, and clustering techniques.

Majority existing surveys are focused towards statistical and rule-based approaches, and these studies are almost a decade old. Hence, many developments in this domain are not yet captured. Deep learning approaches are being widely used in textual problems after the advent of word vectors and their ability to encapsulate semantic as well as syntactic information. Therefore, this study is focused to present overall growth with respect to techniques, datasets, and results over time with major emphasis on deep learning approaches***. For the sake of conciseness, the brief introduction to underlying approaches is provided as supplementary material***.

2.2 Datasets

Datasets serve as the primary ingredient in IE tasks, as they are used for training and testing of various techniques. This section describes major benchmark datasets that are being widely used in literature for training and eventual evaluation of proposed techniques.

2.2.1 Message Understanding Conference (MUC) Corpus. Message Understanding Conference (MUC), which started in 1987, is focused on tasks related to IE. There were in total seven

Table 1. Details Regarding MUC Conference Series Partially Taken from Reference [4]

Year	Conf.	Language	Source Type	Data Sources	Task
1987	MUC1	English	Military reports	Fleet Operations	Open ended (no pre-defined template)
1989	MUC2	English	Military reports	Fleet Operations	IE in form of pre-provided template
1991	MUC3	English	Reports from News	Acts of terrorism in Latin America	IE in form of pre-provided template
1992	MUC4	English	Reports from News	Acts of terrorism in Latin America	IE in form of pre-provided template
1993	MUC5	English, Japanese	Reports from News	Corporate Joint Ventures, Microelectronic production	IE in form of pre-provided template
1995	MUC6	English	Reports from News	Negotiation of Labor Disputes and Corporate Management Succession	NER, Coreference Resolution, Description of NEs and scenarios
1997	MUC7	English	Reports from News	Reports on various aerial crashes, launch report of various missiles and rockets	NER, Coreference Resolution, Description of NEs and scenarios

conferences in this series from MUC-1 to MUC-7. Table 1 lists the name of conference and respective tasks along with text source and its potential domain.

Overall MUC focuses on the task of template extraction for identification of Named Entities (NEs), relations between entities, and event detection. In the latter, this task is divided into scenario-based template extraction, where entities are extracted along with their relationship information in context of some event; and template-based extraction, which primarily deals with task of RE between entities. In this series tasks from MUC-2 to onwards, template was already provided that needs to be filled by the algorithms. Whereas, in MUC-1, no template was pre-defined, so the task was rather open-ended. MUC-6 and MUC-7 extended previous versions by means of adding tasks of NER and Coreference resolution separately as well. MUC-3 and MUC-4 datasets are publicly available, whereas MUC-6 and MUC-7 are proprietary [3].

2.2.2 ACE Corpus. Automatic Content Extraction (ACE) corpus consists of data from broadcast transcripts, newswires, and newspapers data in English, Chinese, and Arabic languages. It is the most widely used dataset in context of RE. This dataset consists of training and testing data separately. ACE vocabulary consists of entities, mentions, and relations that represent objects, references to objects, and relation between the objects, respectively. In ACE, mention has three levels: names, nominal expressions, and pronouns, and ACE tasks are further classified into Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), Event Detection and Characterization (EDC), Entity Linking (LNK), and Time stamps. Table 2 presents brief overview of various ACE tasks that are being performed on ACE corpus over the course of time, along with the languages involved [5]. It is followed by Knowledge base population track in Text analytics conference [6].

2.2.3 Conference on Computational Natural Language Learning (CoNLL) Corpus. Conference on Computational Natural Language Learning (CoNLL) majorly focuses on natural language understanding. In context of NER tasks, it deals with four major types of NEs that include names of location, person, organization, and miscellaneous. All these entities are annotated from newswire articles. Details regarding CoNLL dataset are presented in Table 3.

2.2.4 OntoNotes Corpus. OntoNotes dataset was developed as a collaborative effort between various institutes across the United States. The primary purpose of this project was to prepare a

Table 2. Distribution of Various Tasks against Various ACE Corpus Releases

Corpus	Tasks	Language	Data Source
ACE 2002	EDT, RDC	English	Newswire
ACE 2003	EDT, RDC	English	Newswire, Broadcast
	EDT	Arabic	
ACE 2004	EDT, RDC, LNK	English, Arabic, Chinese	Newswire, Broadcast
ACE 2005	EDT, EDC, RDC, LNK, Time-Stamping	English, Chinese	Newswire, Newsgroups, Weblogs Broadcast
	EDT, EDC, RDC, LNK	Arabic	
ACE 2007	EDT, EDC, RDC, LNK	Arabic, Spanish	Newswire, Weblogs

Table 3. CoNLL Dataset Details

Dataset Name	Year	Language	Source Type	Data Source
CoNLL'02	2002	Dutch	Newswire Articles	Belgian newspaper "De Morgen"
		Spanish	Newswire Articles	Spanish EFE News Agency
CoNLL'03	2003	English	Newswire Articles	Reuters Corpus
		German	Newswire Articles	Frankfurter Rundschau

large human annotated corpus containing various textual-data genres including telephone speech, broadcast, news, talk shows, and so on, in different languages. It has been widely used to evaluate the problem of NER. Details regarding OntoNotes dataset are presented in Table 4. Each new release carries data from previous releases. Hence, source type column in following table only highlights new data sources that were not part of previous release. Due to wide variety of data sources, OntoNotes is one of the largest and the challenging benchmark dataset for NER comprising around 2,945,000 tokens in total.

2.2.5 Semantic Evaluation (Sem-Eval) Corpus. This is a yearly workshop that is focused towards solution of semantic-oriented problems. Its repository comprises various datasets that are being widely used to perform different IE tasks. Widely used ones include RE task of Sem-Eval 2010. Sem-Eval 2017 also included a task that is focused on entity and relation extraction from scientific articles [7].

2.2.6 Other Datasets. There exist various medical repositories. Widely used repositories include MEDLINE, PubMed, and PubMed Central (PMC). MEDLINE is the journal citation database for National Library of Medicine carrying about 24M references to biology and life sciences journals. PubMed consists of more than 27M citations from various biomedical literature repositories that include online books, life science journals, and MEDLINE. PMC carries full-text scientific articles against biomedical and life-sciences journal articles. Some of PMC journals are covered in MEDLINE as well. Many search studies tend to perform information extraction tasks on self-annotated subset of these collections. GENIA dataset [8] is one of the widely used medical source in IE-oriented tasks. It contains manually tagged named entities including various compounds as well as various biological oriented information related to proteins reaction. GENIA dataset is based on GENIA ontology and it currently carries 2,000 abstracts from MEDLINE.

Apart from medical datasets, there exist several workshops or tasks that are organized to address IE problems for low-resource languages. In addition, there exist several extensive datasets for such

Table 4. OntoNotes Details

Dataset Name	Year	Source Type	Language	Data Source
OntoNotes 1.0	2007	Newswire Articles	English	Wall Street Journal
	2007	Newswire Articles	Mandarin Chinese	Xinhua News Agency and Sinorama Magazine
OntoNotes 2.0	2008	Broadcast News	English	VoA, Public Radio International, NBC, CNN and ABC
	2008	Broadcast News	Mandarin Chinese	VoA, China Television System, China Broadcasting System, China Central TV, and China National Radio
OntoNotes 3.0	2009	Broadcast Conversation	English	Phoenix TV and China Central TV
	2009	Broadcast Conversation	Mandarin Chinese or Chinese	Phoenix TV and China Central TV
	2009	Newswire Articles	Standard Arabic or Arabic	An-Nahar
OntoNotes 4.0	2011	Weblogs, Newsgroups	English	English P2.5
	2011	Weblogs, Newsgroups	Mandarin Chinese or Chinese	Dev09, P2.5
	2011	Newswire Articles	Standard Arabic or Arabic	An-Nahar
OntoNotes 5.0	2013	Telephone, Pivot	English	English CallHome, New Testament, Old Testament
	2013	Telephone	Mandarin Chinese or Chinese	Chinese CallHome
	2013	Newswire Articles	Arabic	An-Nahar

languages as well. There exist comprehensive resources over some languages. Table 5 lists some of the non-English resources along with brief detail.[1]

3 METHODOLOGY

To collect the papers for IE sub-tasks being surveyed in this review study, research articles having relevant sub-domain name in their titles were filtered using IEEE and ACM Guide to Computing Literature collection. Exact phrase searching was performed within titles to acquire the relevant literature. The basic set of keywords include: "Relation Extraction," "Named Entity Recognition," "NER," and "Sequence Labeling." As the major focus was on deep learning systems for IE problems, that is why research articles carrying related words that include: "Long Short-Term Memory Models," "Recurrent Neural Network," "Neural Networks," "Deep Learning," "Neural Architectures," "LSTM," "GRU," "Gated Recurrent Unit," "Auto-encoder," "Encoder-Decoder," "Convolutional Network," "Convolution Network," "CNN," "RNN," and "Recursive Neural Network" were also acquired.

The resultant research query results were filtered based on publication type. Initially journal papers and conference papers were selected. The acquired paper set was manually pruned after reading abstracts and keywords. In addition to that, referenced articles in final set of papers were

[1]For detailed log of available datasets for NER in English and various other languages, visit https://github.com/juand-r/entity-recognition-datasets.

Table 5. Other Datasets

Dataset Name	Language	Data Source
MET [9]	Spanish, Japanese	MUC-6 dataset
IJCNLP [10]	Telugu, Bengali, Urdu, Hindi, Oriya	History of India including places and festivals
KPU-NE [11]	Urdu	Fifteen various sources including Education, Health, Science, Novels
Weibo [12]	Chinese	1,890 messages from social service provider "Weibo" with four entities GPE, person, location, and organization
Evalita	Italian	Tweets
		525 News stories taken from "L'Adige"
IREX	Japanese	Mainichi Newspaper
Mongolian [13]	Mongolian	33,209 sentences from news website

Table 6. Paper Distribution over the years Included in Survey for NER and RE

Study Type	Pre 2000	2001–2005		2006–2010		2011–2015		Post 2015	
	NER	NER	RE	NER	RE	NER	RE	NER	RE
Rule-based	0	0	0	1	2	0	0	1	0
Supervised	2	3	4	4	2	2	1	4	0
Semi Supervised	1	1	3	0	5	2	4	0	0
Distant Supervised	0	0	0	0	2	0	3	0	0
Unsupervised	0	1	2	1	2	1	1	1	0
Deep Learning	0	0	0	1	0	4	2	18	10
Joint Modeling	0	0	0	1	3	0	2	0	2
Transfer Learning	0	0	0	0	0	1	0	10	2
Survey	0	0	0	1	1	4	1	4	4
Total	3	5	9	9	17	14	14	38	18

also acquired using Google Scholar to enhance the overall resultant set of articles, so no potential study is missed. Initial queries resulted in approximately 544 studies. Resultant set was later manually pruned to filter studies that are related to current study. Finally, 112 studies are included in the state-of-the-art against two major sub-domains of IE including NER and RE.

Table 6 presents the yearly distribution of research studies included in the survey against NER and RE. Studies are classified based on the underlying approach used to solve a specific problem. It is evident that in the past couple of years, deep learning applications for NER have increased. Recent studies make use of hybrid models that employ deep learning along with statistical models. However, for RE, it can be observed that early studies were more focused on classification models. Whereas, recently widely used approaches include distant supervision, joint modeling, and deep-learning frameworks. In the current survey, studies are selected based on underlying technique and not by limiting specific set of entities or relations. Nonetheless, major classes of entities include general NER that deals with entities such as name of location, person, and organization, and so on, and their set of relations such as located, role, and so on. Other class includes domain-specific entities that vary across the data and problem at hand.

Rest of the article describes NER and RE in detail. Each section covers existing survey studies, current state-of-the-art using deep learning, and other technologies followed by conclusion. All

the results against evaluation measures that include precision, recall, f-score, and accuracy are reported in terms of percentages. P/R/F is used in rest of the article to denote precision, recall, f-score metrics reported in various studies.

4 NAMED ENTITY RECOGNITION

Named entity recognition (NER) refers to the task of identification followed by classification of various NEs from text. NER has many applications such as Question Answering system, IR system, RE system, and so on. The extraction of entities also enables us to perform various other research tasks such as to perform semantic as well as sentiment analysis between various entities or compilation of all various references to a single entity. NER can also be used to improve existing search engines as well as digital research repositories to provide users/researchers with various advanced parameterized search options based on author names, research question, domain, and so on. NER is highly dependent on the textual context, and thus word sequence is important in this problem. For example, the word "Washington" can either refer to a location or a person name. Its effective usage can only be understood if surrounding context is also analyzed. Thus, one of the widely used techniques to solve NER is Sequence Labeling.

Sequence labeling refers to classification of each token in a text stream into pre-defined classes, keeping word sequence and context in view. POS tagging, chunking, and NER are among the widely used sequence labeling tasks, though the concept can be extended to any other domain that requires algorithmic assignment of label per word given any textual stream. This section majorly covers the state-of-the-art in NER, whereas minor attention is also given to the problem of POS-tagging, as many studies tend to emphasis on sequence-labeling problems as whole.

There also exist several schemes to annotate NER data. Widely used tagging schemes include IO, IOB, and BILOU. If two tags appear consecutively, IO cannot distinguish between their boundaries. However, IOB and BILOU both can incorporate boundary information but differ with respect to their respective abilities to model finer context information. To better analyze their differences, consider the following excerpt from previous example:

"***American Airlines**, a unit of **AMR**, immediately matched the move, spokesman **Tim Wagner** said.*"

In IO format, it would be annotated as *"American/ORG Airlines/ORG, a/O unit/O of/O AMR/ORG, immediately/O matched/O the/O move/O, spokesman/O Tim/PER Wagner/PER said."*

Whereas in IOB and BILOU, it would be represented as *"American/B-ORG Airlines/I-ORG, a/O unit/O of/O AMR/ORG, immediately/O matched/O the/O move/O, spokesman/O Tim/B-PER Wagner/I-PER said."* and "*American/B-ORG Airlines/L-ORG, a/O unit/O of/O AMR/U-ORG, immediately/O matched/O the/O move/O, spokesman/O Tim/B-PER Wagner/L-PER said,"* respectively.

4.1 Existing Surveys

Various survey studies that are determined to present the state-of-the-art research in domain of NER are carried out in the literature [14]–[18]. These studies tend to highlight various aspects related to NER. Some are focused on classification of existing techniques; others are focused on respective feature sets that are used. This section explains the existing survey studies.

The review study performed in Reference [14] describes the work done on English as well as on other languages. It first describes various approaches for NER, including supervised and semi-supervised learning methods. It also classifies features for NER in three major categories that include list lookup features, word-related features, and features regarding complete document as well as corpus. Word-related features include features associated with a single word that include casing, POS tag, morphology, and so on. List lookup features refer to inclusion of additional information such as gazetteers or lexicons. Document and Corpus level features, however, provide

meta-information about documents and corpora statistics. Last, it describes various evaluation mechanisms for NER as employed in major conferences of NLP. Among these evaluation mechanisms, one is to perform exact matching while other gives credit to partial matching as well.

Research study carried out in Reference [15] broadly classified existing literature regarding NER into three major classes: supervised, semi-supervised, and unsupervised. In addition to this classification, it tends to list the potential features that are being widely used for NER. These features include entity-level features (entity type, entity frequency, entity distribution, entity neighbor), title-related features (whether word/entity exist in title or not), sentence-level features, and finally document level features. In addition to that, it first presents a baseline system that makes use of various domain-specific heuristic and lexicons. Later, it presents comparative results of various approaches including SVM, transformation-based learning, HMM, and hybrid approaches. Experiments are performed on manually tagged dataset and results show that SVM outperforms the rest. In addition to classification of existing approaches, major tagged datasets for NER are also described in the study that include ACE, CoNLL, MUC, and GENIA. Study also focuses on post-processing tasks of NER to improve the performance. This review study also covers the varieties of NER evaluation systems that include average P/R/F. Study also highlights various matching schemes that are being employed to define correctly classified instance. Among the open areas in NER, study highlights the importance of domain-independent and language-independent NER systems.

A survey study presented in Reference [16] divides task of NER into two broad categories: Generic NER, which deals with generic NE tags, including person name, organization, locations, and so on; and Domain-Specific NER, which deals with specialized NEs such as genes, and proteins information in medical sciences domain. In addition to that, this study tends to highlight the major challenges in NER, including open nature of vocabulary, dealing with abbreviations, polysemy of word capitalization, coreference resolution, and entity linkage issues. Later, it broadly classifies the literature into four classes, including rule-based, supervised, unsupervised, and NE-extraction (NEX). NEX class is closely associated with unsupervised approaches, and it refers to preparation of gazetteers containing NEs. Among the supervised approaches, multiple supervised algorithms, including SVM, HMM, and Maximum Entropy, are briefly explained. In addition to this classification, the progress of NER in non-English language is also briefly expressed, which include various Asian, European, and Indic languages.

Another literature study [17] classifies NER into supervised, semi-supervised, and unsupervised approaches. Supervised approaches are further classified into HMM, SVM, Maximum Entropy, and CRF. Semi-supervised approaches make use of bootstrapping approach to perform NER. Unsupervised approaches are explained using a widely used unsupervised NER system, namely, KnowItAll system, along with application of unsupervised approaches in context of non-English languages. In addition to existing literature classification, survey study is focused on various features for NER. Set of features included in this study include word-level features, digit pattern, common word ending, and list lookup features. Study later briefly explains the neural network framework along with the concept of word vectors and two widely used frameworks for word vector construction that include continuous bag-of-words and skip-gram models.

A survey study for NER in Indian and non-Indian languages is presented in Reference [18]. Primary issues in Indian languages in context of NER include unavailability of annotated corpus, lack of capitalization feature, and so on. This study highlights major corpora used for NER along with implementations for performing NER. In addition to that, it compiles research studies proposed for various Indian languages that include Bengali, Hindi, Nepali, Oriya, Punjabi, Tamil, Telugu, and Urdu,which are further classified on the basis of employed technique. Widely used techniques in context of non-English languages are HMM, MEMM, CRF, and SVM.

4.2 NER in English

As English is a resource-rich language, many researchers have contributed their research efforts to improve the performance of English NER over the past several decades. This section briefly covers the developments made to perform English NER using neural and non-neural approaches. this categorization is done to highlight the progress using deep learning approaches.

4.2.1 Non-neural Approaches. Widely used supervised approaches for NER include HMM, MEMM, SVM, and CRF. Traditional approaches relied solely on underlying algorithm and initial training data. However, now there is a rising phenomenon of semi-supervised or distantly supervised approaches. These approaches often involve external datasets or domain-specific heuristics to make the resultant models more robust. This section briefly explains the developments made with regard to non-neural approaches.

Supervised Approaches. A significant study [19] has pointed out major design challenges that are encountered during the process of NER. The major challenges include the selection of annotation scheme and inferencing algorithm, modeling of long and non-local contextual dependencies, and incorporation of external knowledge resources. It employs averaged perceptron as baseline system and has addressed each one of these challenges comprehensively. The study highlights the suitability of BILOU scheme over IOB, computational and performance effectiveness of greedy search in comparison to Viterbi and beam search for inferencing, exploitation of recent predicted tags, and context to incorporate non-local features and last the use of gazetteers and brown clusters to exploit external knowledge resources. The effect of each of these features is studied on CONLL, MUC, and Webpage datasets by employing averaged perceptron [20]. The proposed approach has outperformed public Stanford NER, and its codebase is also publicly available for research purpose.

A pioneer study to make use of character-level models is presented in Reference [21] to perform NER. It employs token character sequence and its length to train the models. State transitions are applied intuitively to allow only valid transitions. To start and end a character-level transition between words, special tags are introduced. This model is applied and experimented on HMM and MEMM. In addition, contextual features including preceding and following words and their respective POS tag and entity tag are also used. Results show that character models significantly improve the results in comparison to respective word models.

Study carried out in Reference [22] is the first one of its kind that employed joint model for NER and entity linking. The proposed system is based on the semi-CRF, which follows a relaxed Markov assumption between words. The system makes use of multiple features including unigram, bigram, brown clusters, WordNet, gazetteers, entity-level features, and correlation features along with external knowledge-bases such as Freebase and Wikipedia. This system is named JERL, which means "Joint Entity Recognition and Learning." The proposed system is evaluated using CoNLL dataset, which outperforms other state-of-the-art solution for NER. Another joint solution for NER and entity linking by means of supervised graphical modeling approach using variety of linguistic features is presented in Reference [23].

A semi-supervised learning algorithm that makes use of CRF to perform NER is presented in Reference [24]. The proposed algorithm relies on exploitation of evidence. This evidence is not associated with the employed features set, which are used during the process of classification. Hence, evidence is independent, and its exploitation results into high-precision annotations.

Unsupervised Approaches. NER is also carried out in unsupervised fashion. KnowItAll [25] system is renowned system to perform unsupervised NER. Total of three modules are added in this to improve the overall recall. These three modules include pattern learning, subclass extraction, and list extraction. Pattern learning requires set of rules that act as patterns for

further data extraction as well as validator for extracted patterns. Sub-class extraction refers to identification of further sub-concepts. For example, if teacher is to be found, search for professor, associate professor, assistant professor, and lecturer as well. List Extraction module first locates lists of class instances and after locating them, a wrapper function is learned that is further used in extraction of list elements. Results are evaluated on manually tagged corpus where three NEs are evaluated, namely, city, film, and scientist. Best precision achieved against entities city, film, and scientist are References 83, 72, and 77, respectively.

Another unsupervised system [26] mainly consists of two modules: gazette generation and ambiguity resolution. Gazette generation further involves multiple steps. First step is to generate seed query and retrieve web pages in response to the query. Second step is then to extract the required information from acquired web pages. This process is repeated as per the need of the system; at every step, newly identified entity is made part of seed query. After generation of gazette, second module is used to resolve ambiguity. There exist three major types of ambiguities, namely, entity-noun, entity-boundary identification, and entity-entity ambiguity. These ambiguities are resolved by means of several reported heuristics in literature. Experiments are conducted on MUC-7 dataset and results show that proposed system performs better in terms of recall due to gazetteers, on the cost of low precision.

4.2.2 Neural Network Approaches.

Multi-layer Perceptron. The study proposed in Reference [27] treats NER as one-word classification problem. It employs MLP to perform NER, where contextual information is managed by means of sliding window over the input documents. This contextual classifier is evaluated on two datasets that include commercial offers dataset and seminar announcements dataset. Among these datasets, commercial offers dataset consists of short informal sentences containing total of six NEs. Seminar announcement dataset, however, is a benchmark dataset for IE tasks containing four NEs. The proposed approach provides good solution to the domains having informal documents as well as domains where linguistic analysis does not produce good results.

The approach proposed in Reference [28] is a pioneer study in employing neural-based word embeddings to the task of NER. This study proposes a new way of learning word embeddings that exploits information from relevant knowledge sources, i.e., lexicons. The incorporation of external lexicons results in improved word embeddings. To learn the embeddings, basic Skip-gram Model is extended by predicting the lexicon membership of center word along with predicting the vectors of context words.

Deep Learning. A semi-supervised approach based on a neural language model is presented in Reference [29]. This language model is used to train embeddings that can encode both the semantic and syntactic roles of words in context. Later, these context-aware embeddings are used to perform NER. Proposed approach employs RNN for neural language modeling as well as sequence modeling. Similar approach is used in Reference [30], which makes the underlying language model embedding more deep by means of bi-directional LSTMs. The proposed embeddings named Elmo are applied on variety of NLP tasks including NER. Both these approaches are evaluated on CONLL with f1-scores of 91.93 and 92.22, respectively. The distinctive feature of these approaches is to improve existing word-embeddings by incorporating context information. Furthermore, character convolutions are also used to incorporate character-level features.

Neuro-CRF (Deep Learning + CRF). An approach that employs neural networks (NN) with CRF is proposed in Reference [31] as NeuroCRF. In the proposed model, words are represented via word embeddings. The study focuses on two types of networks, namely, Low-rank NeuroCRFs and Full-rank NeuroCRF. Low-rank NeuroCRF refers to the learning of emission weights using

NN, which are being used to classify the words for label assignment. These weights are further complemented by constant transition weights of labels. Full-rank CRFs do not need transition matrix separately. Study shows that on ConLL 2003 task, Full-rank NeuroCRFs tend to outperform low-rank CRFs on chunking and NER tasks. This research study is extended in Reference [32], where shared parameters are incorporated into NeuroCRFs by exploiting similarities in labels. This extension resulted in f-score of 89.62 on ConLL 2003 dataset with ensemble learning.

A hybrid approach that uses CNN, Bi-LSTMs and CRF to solve the problems of POS tagging and NER is proposed in Reference [33]. In this research study, CNNs are used to get character-level embeddings. These character-level embeddings along with word vectors are used as input to Bi-LSTM. Output from LSTMs is later processed using final CRF layer. Furthermore, early stopping and drop-out mechanisms are also incorporated to improve the results.

Another approach that employs dilated CNN with CRF at output layer is presented in Reference [34]. It makes use of iterated dilations to capture the sentence context. The results are evaluated using OntoNotes and CONLL datasets. Experiments show state-of-the-art results in comparison to RNN-based hybrid models. The primary contribution of this approach is exploitation of parallel structures of CNN to improve overall efficiency of NER.

CRF is being used as training along with memory networks in Reference [35]. Memory networks focus on combination of reasoning, attention, and memory to better understand the language. Hence, a variation of CRF is presented that makes use of external memory, which is termed as "Memory-enhanced CRF (ME-CRF)." This integration allows CRF to include information that lies far from neighborhood steps, thus enabling the system to capture long-range dependencies. Whole system consists of two layers: memory layer and CRF layer. Memory layer is further broken down into three main components that deal with input memory, output memory, and current input step, respectively. Input memory and output memory are connected to each other by means of attention mechanisms. Memory weights are calculated using similarity between the current input and the memory contents. The output of this memory network is later processed by CRF to train the system. The proposed approach is applied on CONLL'03 dataset and shows improved CRF performance.

4.3 NER in other/multiple Languages

Other than English, there exist some high-resource languages and many low-resource languages. This section covers studies that are focused on NER from languages other than English; or studies that are addressing NER for multiple languages. Usually, initial developments in any low-resource language such as Malay, Urdu, Hindi, Bengali employ language-specific rules and heuristics. Some developments in this regard include References [36, 37], which employ language-specific rules to identify the NEs. However, nowadays these rules are combined with other machine learning approaches forming hybrid approaches to improve the overall accuracy of systems. Rest of the section describes machine learning–based developments to perform NER in various languages.

4.3.1 Non-neural Approaches.

Supervised. Study carried out in Reference [38] makes use of HMM along with various word-related and number-related features to perform NER. MUC-6 and MET-1 datasets, which are in English and Spanish languages, respectively, were used to perform experiments. In the light of results, f-score has significantly improved in comparison to other state-of-the-art approaches. Results show best f-score of 94.47 when mixed case features are used. Experiments also show that increase in training data tends to improve overall results. As the future directions, authors recommended to improve the bi-gram model, as it does not capture long-range dependencies. In addition to that, they emphasize on hierarchical modeling to capture nested names. Another

future direction is to design training heuristics to assist with corpus annotation. Other studies that employed HMM for NER include Reference [39].

A research carried out in Reference [40] is among the pioneer study to employ MEMM to perform NER. It made use of various features that include lexical, binary, external-lexicons, external-system, section-related, dictionary features, reference resolution, consistency, and compound features. Results are reported on MUC-7 dataset for English, as well as on MET-2 and IREX for Japanese language. Results are compared using various features as well as variations of training and testing datasets.

A pioneer study to employ CRF for NER is presented in Reference [41]. This study made use of multiple external lexicons, where each external lexicon has set of widely used entities. Results are reported on CoNLL 2003 dataset of German and English articles with f-scores of 68.11 and 84.04, respectively.

HMM-based NER for Chinese language is presented in Reference [42], which makes use of two-stage process. First, words are segmented using bigrams followed by NER using lexicalized HMM. This approach is evaluated using character- and word-level lexicalized HMMs. Furthermore, a lexicon is also used to identify known words. Major three entities, namely, person, location, and organization, are being identified. The proposed approach is evaluated by means of three different Chinese datasets.

A recent study [43] is focused on employment of Character Level language Model (CLM) and various features to improve the performance of multi-lingual NER systems. It employs CogCompNLP system as baseline [44] for NER. Total of eight languages are used for demonstration of proposed approach, including Tagalog, Somali, Hindi, Farsi, Bengali, Arabic, Amharic, and English. For English, CoNLL'03 dataset is used, whereas LORELEI language pack [45] is used for remaining seven languages. The proposed CLM model treats a complete word as sentence, whereas each character is treated as a word in language model. Moreover, variety of language models including n-gram, skip-gram, continuous bag-of-words, and bi-linear are employed. Two separate CLMs are trained for entity and non-entity tokens using existing annotated datasets. Other features describing the languages and entity status using CLM's perplexities are also used. Results show that the proposed approach is simple and straightforward in nature and performs at par with state-of-the-art deep learning approaches. Keeping recent concerns regarding environmental impact of training of deep neural architectures in view, the proposed approach carries immense significance.

Semi-supervised. There exist various semi-supervised approaches for NER. First one in this regard is a Transformation-based learning (TBL) approach. TBL is a machine learning–oriented, rule-based approach that learns and improves rules over time [46]. It is an iterative approach to refine and modify rules. Its first phase is to annotate the data, which include random class assignment or sophisticated algorithms. This assignment is later followed by the iterative process. An objective function is used to calculate distance/similarity between gold-standard and annotated corpora. By making use of this objective function, in every iteration, transformation rules are updated until a defined accuracy criterion is met or transformation rules no longer get updated. This approach is employed to perform NER in Reference [47] for Filipino language.

A bootstrapping-based approach proposed in Reference [48] makes use of entity-level information and contextual information to perform NER. First, candidate entities are identified via chunking algorithm followed by their soft classification in respective NE. Boundaries on entities are determined using very basic rules such as position features, capitalization features, word length, and so on. Experiments are conducted on Dutch and Spanish languages.

Another semi-supervised approach for Indonesian language is presented in Reference [49]. This study makes use of DBPedia for quality improvement of training data. First, CRF is used to perform

initial NER. Later, information from DBPedia is exploited to assign tags to the unlabeled data. This labeled data is later added to overall training set by means of a scoring mechanism for next iteration, which consequently improves the results. Results are evaluated on manually annotated dataset of 75 Wikipedia articles, which show improvements in f-score after every iteration.

4.3.2 *Neural Network Approaches.*

Multi-layer Perceptron. Study presented in Reference [11] performs comparative analysis of HMM and MLP on Urdu dataset KPU-NER to recognize three distinct entities that include person, organization, and location. Different input features for MLP, word-embeddings, and context window size are used. Among these two, MLP outperformed HMM.

Deep Learning. Research study presented in Reference [50] makes use of bi-directional GRUs to perform NER. Input layer comprises surface forms of words along with context words, with optional information of lemmas and tags. Each word can be modeled by means of word vectors, character-level embeddings that are being computed by means of bi-directional GRU, prefix/suffix features, and manually designed classification features. In proposed architecture, hidden layer consists of parametric rectified linear units, whereas output layer is a Softmax layer. To train the network, AdaGrad is used along with dropout at hidden layer. Ablation study is performed to analyze the effect of various word models on performance of resultant system. Experiments conducted on Czech news data using proposed architecture resulted in f-score of 89.92.

Hybrid (Neuro-CRF). An alternate study that employs CLM using LSTMs and CRF for sequence labeling problems including chunking, NER, and POS tagging is carried out in Reference [51]. The underlying neural architecture employs character-level embeddings and highway neural layers and co-training to achieve improved performance in specific tasks. The proposed system named LM-LSTM-CRF is employed in Reference [52] to perform NER on Vietnam text.

Research study carried out in Reference [53] combines RNN with CRF to perform NER. It employs output-label dependencies with transition features, which is the core idea of RNN. Additionally, it also employs sequence information to calculate objective function in the similar way as that of CRF. In this combined model, cross-entropy function is used for training via back propagation through time with Stochasitc Gradient Descent (SGD). The study provides comparative analysis between variants of RNN such as GRU and Structurally Constraint Recurrent Networks (SCRN), Bi-LSTMs, Bi-GRUs, and their combinations with CRFs. Results show that Bi-LSTM-CRF tends to outperform the rest with f1-score of 90.24 for CoNLL 2003 dataset. Experiments were conducted on ETRI Korean dataset as well and results show that GRU-CRF and LSTM-CRF both tend to outperform other techniques with f1-score of 90.89. Thus, the major contribution of this study is its comparative analysis of different RNN architectures with and without the incorporation of CRF along with the state-of-the-art results on various benchmarked datasets.

Bi-LSTM-based approach along with CRFs is proposed in Reference [54] for multi-lingual NER. The model employs character-level embeddings using Bi-LSTMs along with pretrained word embeddings and dropout. The proposed model outperformed state-of-the-art NER systems in Spanish, German, and Dutch languages, whereas for English language, it performed just at par with state-of-the-art system, which employed many features [22] and external knowledge-bases such as Wikipedia and Freebase. The major contribution of this study is the generalization of proposed approach across multiple languages without any need of language-specific rules or gazetteers. Similar approach that employs both Bi-LSTM and CRF is proposed in Reference [55] for Mongolian NER.

4.3.3 Transfer Learning. Direct transfer learning for cross-lingual NER is employed to deal with data scarcity in various languages in Reference [56]. To carry out this task, cross-lingual word clusters [57] are used that can help in transfer of linguistic structure across languages. Further learning is carried out using self-training approach and native language word clusters. Experiments conducted on CONLL 2002 and 2003 datasets show proposed approach gives promising results and is particularly useful to shorten the annotation time.

A study with intuitive idea to apply transfer learning with varied target labels is presented in Reference [58]. Here, target language carries more labels than the source ones. Hence, it makes use of CRF that is trained on bulk of source language data. This training is used later to learn the correlation between source and target datasets. Afterwards, another linear chain CRF is used to learn domain-specific patterns using target dataset. This way, even in the presence of limited target dataset, proposed approach achieves better results. Similar objective of varied NE labels between source and target datasets is also studied in Reference [59] using deep neural networks.

An eminent study [60] incorporates the novel concept of translation of lexical resources from high-resource to any low-resource language. Hence, the work is focused on performing cross-lingual NER. Furthermore, if target language has its own Wikipedia, then proposed approach improves performance by incorporating linguistic knowledge learned through Wikipedia resources. The proposed approach is employed on eight different languages. These languages include Dutch, German, Spanish, Turkish, Bengali, Tamil, Yoruba, and comprehensive case study on Uyghur, which is a resource-poor language. For initial translation, the most resource-rich language, i.e., English, is used. It is also studied that English works better for European languages, i.e., Dutch, German, and Spanish. Hence, for non-European language, lexicons are prepared using common set of entities across languages.

Study emphasizing on low-resource languages [61], including Marathi, Tamil, Bengali, Hindi and Malayalam, performs NER using multilingual learning by means of Bi-LSTMS and CRFs. Multilingual training makes use of annotated concepts that are present in other languages for training. Hence, in this particular research study, multi-lingual training is employed using Hindi as an assisting language due to the availability of annotated Hindi datasets. This study also employs CNN to extract word-level features. Similar work for Dutch, Spanish, and Chinese is performed in Reference [62], which further employs cross-lingual knowledge transfer approach and lexicon extension strategy to deal with out-of-lexicon words using English as high-resource language.

The study presented in Reference [63] makes use of transfer learning. The authors have pointed out the similarities in the nature of Chinese word segmentation (CWS) and NER. Hence, they have proposed a joint model to solve the problems of CWS and NER. It is performed, as in views of authors, both problems are very similar. Also, the resources available for Chinese NER are very small in comparison to CWS resources. Hence, a transfer learning framework based on adversarial CWS is presented to perform NER. This approach is pioneer in application of adversarial framework for transfer learning in Chinese NER. It makes use of shared and task-specific feature extractors using Bi-LSTMs.

Another transfer learning approach employing zero-shot and few-shot learning for multilingual NER is presented in Reference [64]. It proposes various models, including supervised models with hundred instances in target language, unsupervised models, single source transfer, and multi-resource transfer using truth inferencing. Comprehensive experimentation is conducted using total of 41 different languages to demonstrate the effectiveness of proposed approach. A cross-lingual transfer learning approach presented in Reference [65] performs Japanese's NER using English as a source employing deep neural framework. The distinctive feature of proposed approach is employment of orthographic models to deal with writing difference in source and target languages.

Other approaches to transfer learning include exploitation of phonology rather than characters in neural frameworks to improve performance across domains [66] and employment of deep learning networks [67].

4.4 NER in Specific Domain

In addition to language-specific progress, many researchers have studied NER problem in a specific domain. This means domain-specific entities are extracted tailored to information need of respective community. There exist some comprehensive surveys that cover state-of-the-art in a particular domain, e.g., scientific entity extraction [68], chemical compound extraction [69], biomedical entities extraction [70, 71]. Following section briefly covers application of NER in various areas, mainly social media sites such as Twitter, judgments, and disease extraction.

4.4.1 Non-neural Approaches. A pioneer study employs rule-based approach to extract dietary recommendations [72]. Total of three tags including food, nutrient, and quantity are dealt in this study. To extract these entities, study makes use of domain-specific lexicons along with rule-bases to extract the required information. No prior annotated corpus is available for this task. Thus, results are evaluated on 100 manually selected documents where half of them comprise research papers' abstracts related to the domain of "dietary intake" or "food-composition" and rest include dietary recommendations passages from 12 different websites. Results are reported in terms of TP, FP, and FN against every entity. Self-calculated P/R/F using provided information against food, nutrient, and quantity are (99.08, 95.56, 97.28), (99.64, 97.03, 98.31), and (100.0, 88.65, 93.98), respectively. This study is pioneer in applying NER to extract the information from dietary recommendation.

Another approach proposed in Reference [73] employs a filter-stream-based NER technique to extract NEs from Twitter data. Twitter streams are usually short and informal, making conventional NER approaches not very suitable for them. In this approach, filters are used to extract NEs. The proposed approach uses five different filters based on nouns, affixes, terms, dictionaries, and context information. These filters employ probabilistic analysis to determine the suitable label for tokens in given message and do not have any dependency on grammar rules. Micropost data was used to evaluate the results on four NEs: person, location, organization, and miscellaneous. In this study, NER task is distributed in several recognition processes due to usage of multiple filters. The major achievement of this study is its computational efficiency and simplicity in nature.

4.4.2 Neural Network–based Approaches.

Deep Learning. The model proposed in Reference [74] employs NN in unsupervised fashion to extract NEs from Italian text. The study majorly focuses to cover the weaknesses of existing models. Existing models primarily suffer with two limitations: first, only local information is considered, and second, due to limited labeled data, models tend to over fit. To resolve these issues, study proposes a context-window-based RNN along with a new recurrent feedback mechanism to ensure proper modeling of dependencies. In addition to that, a network pre-training technique using weakly labeled data is also used, followed by employment of early stopping, weight decay, and dropout to avoid over fitting and consequently improving generalization. The results are reported on Evalita 2009 benchmark with P/R/F of (85.69, 80.10, 82.81), when gazetteers are employed as well. By using pre-trained network and context-window-based recurrent architecture, respective P/R/F is (82.74, 80.14, 81.42). The major contribution of research study is a recurrent architecture that has the capability to model the dependency of output tags.

A Bi-LSTM-based semi-supervised approach is presented in Reference [75] to perform NER on Chinese social networks. Proposed approach incorporates self-training that is being performed using similarity and sentence ranking functions. Furthermore, out-of-domain datasets are also

used. Results are evaluated on social network "Sena Weibo" using "SIGHAN" annotated corpus as out-of-domain dataset.

A CNN-based model is presented in Reference [76] to perform disease NER. It takes character-based embeddings, word embeddings, and lexicon feature embeddings as input. Later, it employs a CNN network with a methodology, called multi-label strategy (MLS). The MLS enables models to capture correlation between labels in neighborhood by means of predicting contextual words. The proposed approach gives state-of-the-art results on CDR [77] and NCBI [78] datasets.

The study presented in Reference [79] makes use of CNNs to perform NER on Chinese electronic medical records. To train CNN classifiers, one-versus-rest approach is used, i.e., separate classifier is trained against every NE in the dataset. The dataset consists of 992 clinical notes with total of five NEs, including treatment, disease group, disease, test, and system. The proposed model is evaluated using Chinese discharge summaries and progress notes.

Hybrid (Neuro-CRF). Another approach presented in Reference [80] is focused on domain-specific Chinese NER by means of Bi-LSTMS and CRF. Its fundamental focus is to employ document-level features rather than sentence-level features. Furthermore, integer linear programming is carried out using various features. System training is carried out using OntoNotes 5.0 Chinese collection. Whereas, to evaluate the robustness and generalization of system, legal judgments are used.

4.6 Nested NER

In addition to above studies, there are research studies dealing with nested NER as well. Nested NER refers to phenomenon when an entity itself can be further decomposed into several sub-entities; e.g., the entity "University of Punjab" itself is an organization, but the token "Punjab" is a location. One widely used method for nested NER is to extract top-level entities followed by further classification of their tokens and/or clauses in respective categories.

4.5.1 Non-neural Approaches. Pioneer study in this regard [81] employs sentence parse trees, jointly modeled POS tags, and named entities using constituency parsing. The proposed approach gives promising results on top-level NER as well as nested NER. However, it is computationally very expensive.

Another prominent study [82] for nested mention detection makes use of hypergraph and CRFs. These hypergraphs can express wide variety of nested mentions. Furthermore, for improvement, various manually crafted features are used. The proposed approach is experimented on wide range of corpora, including ACE English dataset, Genia, and CoNLL. This idea of hypergraphs for nested NER is further employed in Reference [83], which relies solely on greedy approach for hypergraph population and also does not use any manual features.

4.5.2 Neural Approaches. A deep learning approach for nested NER [84] makes use of dynamic framework. The proposed approach stacks flat LSTM-CRFs layers until no nested entity remains. Each flat LSTM-CRF layer employs bidirectional LSTM to capture sequential context, whereas CRF deals with global context. Experiments are conducted on a variety of corpora, which show promising results.

An empirical comparative study for English and Japanese fine-grained NER is performed in Reference [85]. The study conducts a comparison between deep LSTM frameworks and traditional CRF and SVM-based models using external resources and different features. In the latter model, CRF is used for top-level NER, whereas SVM is used for classification of top NEs into fine-grained ones. Furthermore, due to huge count of Japanese characters, CNN-based word embedding models

that present state-of-the-art results in English do not perform well. Hence, the authors have proposed a dictionary-based solution to replace CNN layer for Japanese language. The results show that regardless of data size, for English language, neural frameworks outperform CRF+SVM, whereas for Japanese, in small-sized datasets CRF+SVM outperform LSTM-CRF with wide margins.

4.7 Conclusion

NER is of utmost importance in major applications that require natural language processing such as question answering as well as querying systems. With the advent of word vectors, recent studies represent words as word vectors to effectively harness the semantic as well as syntactic nature of words in improved fashion. Combination of deep learning frameworks with statistical methods is also being used in recent studies. If the modeling is based on token-level sequence, there is a chance of dealing with unknown tokens while dealing with unseen data. Whereas, if character-based sequence models are used, it is highly unlikely to encounter an unseen character. This has led to the wider incorporation of character-level models. Some of the schemes only rely on character-level models, whereas others apply character-level embeddings in the first layer and word level on the subsequent layer. These decisions related to the primary tagging scheme, word embeddings, character embeddings, and incorporation of dropout and early stopping are among the factors that hugely impact the overall result. A recent study in Reference [86] comprehensively explains and elaborates the effects of various available options when recurrent neural networks are applied to solve the sequence labeling problems. Hyper-parameters that are evaluated in this study include pre-trained word embeddings, character representations, optimizer, gradient clipping and normalization, various tagging schemes, dropout rate, number of recurrent units, number of stacked LSTM layers, mini-batch size, and classifier (that is used in final layer of network). In the light of this study, some of these features have a huge impact on overall system accuracy, whereas others do not impact much. Thus, it is of immense importance to understand the impact of various hyper-parameters on overall system. In other words, comprehensive study of various models' hyper-parameters and their effects on overall result is one of the major research questions in this domain. In addition to that, features selection also has a great impact on overall results. Thus, such analytical studies are of utmost importance in context of NER. Table 7 presents summary of NER literature covered in this study. In following table, *Features/ Properties* column, sub-entry *E* denotes external features such as lexicon, *W* denotes any word features such as its orthography, embeddings; *C* denotes incorporation of character-level features, and *O* denotes any other distinction. *Typ* column corresponds to respective NER group, *Results* column shows the respective precision, recall, and F-measure values against a study, *Lang* denotes primary language, whereas *Dataset* represents either the name of the employed dataset or its genre. Furthermore, *Y* as value denotes presence of a particular feature, whereas *D* denotes domain-specific in Features/Properties column.

5 RELATION EXTRACTION

Relation extraction (RE) is one of the key tasks involved in information extraction. It refers to classification of semantic relationship that can exist between entities. This type of information is essential to construct semantic knowledge bases (KBs), which can be further employed to infer the relationships that exist between various entities. In addition to that, relation extraction is useful to develop question answering systems to perform text summarization as well to construct taxonomy of concepts. Many approaches to RE make use of already extracted named entities and later establish link between them by means of heuristics or machine learning–based algorithm. Recent algorithms tend to solve the problem on NER and RE in a joint fashion. Remaining section

Table 7. NER Literature Summary

	Technique[1]	Features/ Properties				Typ[2]	Results			Lang.	Dataset
		E	W	C	O		P	R	F		
[21]	HMM, MEMM	-	Y	Y		HR				English	CONLL
										German	CONLL
[22]	Semi-CRF, JM	Y	Y		Brown Clusters, Wiki	HR	91.5	91.4	91.2	English	CONLL
[26]	US	Y			Heuristics		Low	High			MUC-7
[27]	MLP		Y		Sliding Window	HR	87.41	86.15	86.76	English	Commercial offers
							85.57	86.22	85.95		Seminar Announ.
[28]	MLP	Y			Skip-gram	HR			90.9	English	CONLL
									82.3		OntoNotes
[29]	RNN		Y	Y	Language Model	HR			91.93	English	CONLL
[30]	Bi-LSTM		Y	Y	Language Model	HR			92.22	English	CONLL
[32]	Neuro-CRF		Y			HR			89.62	English	CONLL
[33]	Neuro-CRF		Y	Y	Bi-LSTM	HR				English	CONLL
[54]	Neuro-CRF		Y	Y	Bi-LSTM	HR	Multiple languages are used.				
[34]	Neuro-CRF		Y		CNN, Iterated Dilation	HR			90.65	English	CONLL
									84.53		OntoNotes5
[35]	Neuro-CRF		Y		Memory Network				89.5	English	CONLL
[42]	HMM	Y	Y	Y	Lexicalized HMM	OTH				Chinese	Multiple Chinese Datasets
[11]	MLP	-	Y	-	Context Window	OTH	81.05	87.54	84.17	Urdu	KPU-NE
[47]	SS				TBL	OTH	76.45	99.20	86.36	Filipino	Asian Hist. Ref.
[48]	SS		Y	Y	Bootstrapping, linguistic rules	OTH	73.03	71.62	72.31	Dutch	CONLL
							78.19	76.14	77.15	Spanish	CONLL
[49]	SS	Y	Y		Iterative	OTH				Indonesian	75 Wikipedia Articles
[74]	RNN	Y	Y		Early Stopping, Weight Decay		85.69	80.10	82.81	Italian	Evalita (Tweets and News)
[50]	DNN		Y	Y	Bi-GRU, AdaGrad	OTH			89.92	Czech	News
[52]	DNN		Y	Y	Co-training	OTH			94.56	Vietnam	VLSP
[53]	Neuro-CRF		Y	Y	LSTM, GRU, SCRN	OTH			90.89	Korean	ETRI
[72]	Heuristic	D	-	-		DOM	99.57	93.75	96.52	English	Dietary Recom.
[73]	CRF	D	Y			DOM	67.81	52.52	58.46	English	Micropost Twitter
[87]	US				Phrase Chunking	DOM			15.2	English	GENIA
									26.5		Pittsburgh
[74]	RNN	Y	Y		Early Stopping, Weight Decay	DOM	85.69	80.10	82.81	Italian	Evalita
[88]	LSTM		Y	Y		DOM	82.70	86.70	84.60	English	Pubmed Abstracts
[75]	LSTM	Y	Y	Y	Cross domain learning	DOM			59.78	Chinese	Social Media
[79]	CNN		Y		One vs rest approach	DOM			88.64	Chinese	Discharge Summ.
									91.13		Progress Note
[80]	Neuro-CRF				Document level features		87.38	87.38	87.38	Chinese	Marriage Judge.
							94.49	88.60	91.45		Contract Judge.

(Continued)

Table 7. Continued

	Technique[1]	Features/ Properties				Typ[2]	Results			Lang.	Dataset
		E	W	C	O		P	R	F		
[38]	HMM	-	Y	-	-	MUL	96.00	93.00	94.47	English	MUC-6
									90.00	Spanish	MET-1
[40]	MEMM	Y	Y		Reference Resolution	MUL			90.25	English	MUC-7
									83.80	Japanese	MET-2
									77.37	Japanese	IREX
[41].	CRF	Y	Y						84.04	English	CONLL
									68.11	German	CONLL
study [43]	CLM	Y	Y	Y	Language Models, CogCompNLP	MUL	Performance at par with recent DL frameworks			Tagalog, Somali, Hindi, Farsi, Bengali, Arabic, Amharic and English	
[61]	Neuro-CRF	Y	Y		Bi-LSTM and CRF for NER, CNN for word features	MUL			70.90	Marathi	
									55.57	Bengali	
									64.27	Malayalam	
									60.25	Tamil	
[60]	TL	Y	Y		Wikipedia, Translation of lexical resources, Cross-lingual NER	MUL	Training of each model using English and one relevant language			Dutch, German, Spanish, Turkish, Bengali, Tamil, Yoruba, Uyghur	

[1]SS, US, TL denote semi-supervised, unsupervised, respectively, and transfer learning.
[2]HR, OTH, MUL denotes high-resource, others, and multiple languages, respectively.

briefly explores the existing research surveys conducted in the context of RE followed by brief classification of existing approaches.

5.1 Existing Surveys

Multiple survey studies that cover the research in domain of RE are presented in References [89–91]. These studies tend to highlight various aspects related to RE. Some are focused on classification of existing techniques, whereas others are focused on respective feature sets that are used. In addition to that, some surveys elaborate existing RE systems in brief fashion. This section highlights the findings of existing survey studies briefly.

The review study presented in Reference [89] reports classification of existing literature for RE on the basis of various approaches. Later, it presents brief description and widely used systems against each approach along with comparative analysis. In addition to that, it lists the widely used evaluation mechanisms that exist to evaluate various systems of RE. Major datasets that are being used for RE are also compiled and briefly mentioned. Last, it highlights the application of RE in domain of question-answering systems and biological natural language processing. Subsequent paragraphs briefly explain the findings of this survey study.

This review study classifies existing RE literature in two major classes: supervised and semi-supervised approaches. Supervised approaches are further distributed into kernel- and feature-based approaches. Widely used kernels, which include bag of features kernels and tree kernels, are briefly explained. This study concludes that among feature-based and kernel-based supervised approaches for RE, the latter gives better results. For semi-supervised approaches, this study briefly explains multiple RE systems for Open IE and Web, including DIPRE, SnowBall, KnowitAll, and TextRunner systems [92–95] in detail. Here, Open IE refers to the task of RE when applied on domains, having very little or no training data. Respective research study further provides pros and

cons of each described system while performing Open IE. Primary difference between TextRunner and the rest of approaches is that TextRunner learns relations, entities, and classes from data on its own in self-supervised fashion, whereas rest require *a priori* knowledge about the relation types, which are of interest.

Regarding evaluation of RE, supervised approaches are mostly evaluated by means of precision, recall, and f-score. Semi-supervised approaches, however, are rather difficult to evaluate due to multiple factors including Web data. One widely used process, which is followed in the evaluation of References [92, 95], involves forming small set of data that is randomly drawn from system-generated output. This output set is later manually validated from human. System precision can later be calculated with system output and human input, but recall cannot be calculated using this process, as actual records that should be part of result are not known.

Major datasets that are being used for RE include MUC, ACE, and MEDELINE, where MUC is intended for NER in its nature. In authors' view, MUC dataset can be extended to perform RE. Authors also discuss and present several ideas to extract non-binary relations in light of research study carried out in Reference [96] that makes use of heuristics and graph-based approach. Among the systems, TextRunner claims to be able to extract n-ary relations, but the complexity involved is not explicitly mentioned. Another major limitation identified in the existing systems is that they tend to perform RE on sentence level, thus relations spanned across sentences cannot be captured by any system.

Another survey presented, in context of RE [90], tends to classify the existing systems into four major categories. These categories include knowledge-based, supervised, self-supervised, and joint models. Knowledge-based systems are quite efficient if domain is well-defined, as they mostly make use of domain specific rules but primary weakness of knowledge-based system is that it involves lots of human effort and it does not perform well when applied on other domains. Supervised systems, on the contrary, show better performance across domains as well, as long as training data is available. But these systems are costly, as they require annotated dataset for training, where data annotation is human intensive and laborious task. Supervised systems that make use of set of seed patterns and later perform bootstrapping are classified in the category of weak-supervised systems in the context of this survey study. Last, self-supervised systems are those that tend to label datasets by means of independent extraction rules.

In addition to providing primary classification of existing RE systems, this study tends to briefly express the notion of Open-IE along with major systems that perform Open-IE, such as TextRunner. This survey study last presents state-of-the-art RE systems that include distant-supervision and joint systems. Distant-supervision systems employ external RE resources, such as FreeBase and Wikipedia, which contain entities and their relations. Basic idea regarding distant supervision involves filtering of sentences carrying existing entities. These filtered sentences are later processed, and features are learned to extract relations between the entities. Last, joint approaches to RE are discussed, which tend to collectively model various NLP tasks. The primary benefit of joint modeling approach is that it results in improvement of cascading errors, which usually occur in pipeline tasks.

A literature survey that is primarily focused on the task of cause-effect RE is presented in Reference [91]. It majorly classifies the literature in two paradigms of non-ML and ML. Non-ML approaches usually make use of pattern matching and count-based approaches to identify relations. On the contrary, ML-approaches are employed for extraction of implicit relations. A brief overview of these paradigms is presented along with relevant literature and comparative analysis. In addition to that, difficulties and challenges associated with the task of causal RE are highlighted, which include contextual dependency and data sparsity. Authors argue that by employing feature construction techniques and lexico-semantic resources like WordNet, results are being improved using

sophisticated ML algorithms. Future directions in this area, as per the authors' view, include comparative analysis of various techniques on empirical datasets, employment of deep learning, and combination of various relation classifiers for performance improvement.

A survey study presented in Reference [97] is focused on determining the recent developments for the task relation extraction when performed using different variants of CNN. This study first provides brief explanation of the major constructs that are part of deep learning frameworks, including word embeddings and position embeddings. Later, it classifies CNN approaches in two major classes, supervised and distant-supervised, where distant supervision refers to automatically training of bulk data by exploiting the information residing in existing KB. The supervised CNN approaches include CNNs with max-pooling and various kernels. Here, kernels are employed to capture n-gram features. However, CNNs that are being widely used for distant supervision include employment of piece-wise kernels, cross-document max-pooling, and various attention mechanisms. Study concludes that majority of models now employ distant supervision, as deep learning models require extensive data to be trained. Recent models focus on minimizing noise that becomes part of data, due to the very nature of distant supervision.

A comprehensive study covering the state-of-the-art with respect to distant supervision is presented in Reference [98]. It mainly identifies primary challenges involved in distant supervision that includes wrong label assignment and coverage of KBs. As distant supervision is used to label large set of unlabeled data using existing annotated datasets, it exploits data from existing KBs as well. Due to its primary nature, two issues identified tend to affect the overall quality of annotations. First issue regarding wrong word labels is being addressed using classifiers for entity mentions. Second issue, however, has been handled by means of prior refinement of underlying KBs before applying distant supervision. As for future directions, study points that incorporation of crowd sourcing and human-in-the-loop technology can help in dealing with noisy labels. The major research question in this aspect is the identification of efficient ways that result in minimal human efforts and improved system performance. In addition, feedback loop that can learn from human feedback and refine the results accordingly is another open direction. Other research questions include application of various deep learning approaches, incorporation of more contextual information, and effective ways to extract infrequent relations.

5.2 Non Neural-network-based Approaches

As in this study, primary aim is to highlight the progress made using neural approaches. Therefore, classification is made on the basis of neural framework employment. This particular section is focused on covering state-of-the-art in the light of approaches that do not use neural frameworks. Following sub-sections classify existing literature of non-neural approaches into further categories.

5.2.1 Rule/Heuristic-based Approaches. A rule-based approach is proposed in Reference [99] that first generates dependency parse trees and later employs knowledge-base/rule-base to extract relations. These extracted relations are later processed to check for negation, effect detection, and conjunction handling. If relations are preceded with any negation word, that particular relation is discarded. Effect detection primarily deals with detection of effector and effectee and it is carried out using heuristics regarding sentence formation, i.e., active voice or passive voice. Conjunctions are also handled using pre-defined rules. Results are reported on dataset that was prepared for Learning Language in Logic (LLL) 2005 workshop [100] that resulted in P/R/F of References [68, 83, 75].

An approach proposed in Reference [101] makes use of syntactic parser to perform RE. By using DBPedia dataset, rules are written to extract patterns using 188 articles of training set. Quero

news, news source in French language, [102] is used for evaluation. Other rule-based studies for RE include References [103, 104].

5.2.2 Supervised Approaches. The research study reported in Reference [105] makes use of feature-based approach using MEMM as a training medium using lexical, semantic, and syntactic features. Proposed model uses words, NE tags, mention labels (name, nominal, and pronoun), overlap information (number of words between two mentions) and parse tree information as features streams. All syntactic information is acquired using syntactic parse tree as well as dependency tree that are computed using Ratnaparkhi's tagger [106] trained on PennTree Bank dataset. The proposed approach shows competitive results with other techniques proposed on ACE corpus with improved scalability in terms of relation types.

An approach that employs SVM using various features along with base-phrase chunking and semantic information from WordNet is reported in Reference [107]. An external lexicon of people names is also used to further improve the results. Results show that incorporation of base-phrase chunking has resulted in the increase of (4.1, 5.6, 5.2) in P/R/F, respectively. This study is the pioneer in employment of WordNet within a supervised framework for performance gain.

A knowledge-driven RE approach is presented in Reference [108]. Based on the relations that exist between entities, it argues that relations can be expressed using semantic and syntactic expressions. It further states that these expressions are relatively easy to identify. Based on this insight, the study demonstrates the effectiveness of employing these semantic and syntactic expressions to perform RE. To carry out this task, major four relation expressions from ACE datasets are extracted by means of simple rule, which cover around 80% of the entire corpus. Separate classifiers are developed using SVM and variety of features for binary, coarse- and fine-grained entities. Analysis conducted using various experiments shows that proposed approach presented very good results and effectively reduced the mention pairs having no valid relation between them.

Aforementioned supervised studies relied on features. Supervised approaches also use kernels to improve the results. There exist studies that perform comparison of various kernels suited to solve a problem. One such study proposed in Reference [109] employs SVM and voted Perceptron learning with kernel functions to perform relation extraction. This study takes shallow parse along with recognized NEs as input. Two types of kernels are described in study that include contiguous sub-tree kernels and sparse sub-tree kernels, whereas kernel functions are computed by means of dynamic programming. Evaluation conducted on 200 newswire articles and publications shows that proposed kernel-approach outperforms existing approaches. It also reports that kernel-based methods are superior to feature-based supervised approaches for RE. Similar approach is employed in Reference [110], which makes use of tree kernels to measure similarity between sentences.

Research carried out in Reference [111] makes use of convolutional tree kernels along with SVMs to extract relations. Study majorly focuses on use of parse trees along with convolution tree kernels to improve the RE performance, as tree kernels have capability to explore syntactic information that is exhibited in parse trees. In addition to employment of various kernels, a composite kernel is also proposed that is based on a linear and a tree kernel. Results are evaluated on ACE'04 dataset, which show that the composite kernel outperforms existing approaches. Another kernel-based approach is proposed in Reference [112].

5.2.3 Semi-supervised Approaches. A pioneer system that tends to perform RE in autonomous fashion is KnowItAll system [94]. This system consists of four major modules. These modules include data extractor, an interface for search engine, an evaluation module termed "Assessor," and database. It uses ontologies and bootstrapping to extract relations. Assessor is the module that deals with assigning probabilities to extracted relations. The major contribution of this study

is an automatic system that tends to perform relation extraction in scalable, domain-independent, and autonomous fashion.

Another bootstrapping approach is proposed in Reference [113]. It employs radial bias kernel function and SVM to measure similarity. Apart from kernel functions, lexical, syntactical, deep-parsing, and entity features are being used during classification. In addition to that, three variants of bootstrapping are used. First one is traditional baseline bootstrapping approach. Second one incorporates concept of bagging„ i.e., multiple classifiers are trained in an iterative fashion. Third approach performs bootstrapping with random projection of features in sub-spaces, whereas, in each projected space, committee of classifiers tends to mark unlabeled points. Among these three approaches for bootstrapping, last approach employing random projection performs better than the rest, and this algorithm is called "BootProject" algorithm. To evaluate, first SVM was trained using whole training data (4,328 records) and subset of training data (100 records) marking ceiling and floor accuracies. Later, aim was to achieve accuracy closer to the ceiling one using bootstrapping. BootProject algorithm's key advantage is that it reduces the need for labeled training data significantly by making use of random feature projection. Other approaches based on bootstrapping are proposed in References [93] and [114].

CRF-based semi-supervised learning approach for RE is presented in Reference [115]. It tends to extract relations by learning relational and contextual patterns. In addition to that, a database is also used to hold the existing relationships. Multiple experiments are reported using different parameters and thresholds. This approach tends to employ top-down pattern discovery by means of data mining to discover relational patterns.

An alternate methodology is proposed in Reference [116], which is focused on dealing with primary limitation of semi-supervised learning that include incorrect or unreliable set of relation extractions. To overcome this limitation, it performs simultaneously learning of classifiers in the presence of an ontology. This enables training for various entities and relations. Ontology here defines constraints that control the simultaneous training of classifiers. In other words, it employs bootstrapping approach along with set of constraints that tend to couple different relations. To evaluate the results, first dataset is prepared that contains self-crawled 200 pages belonging to categories of sports and companies. Further, noise is added in the dataset by adding documents from other domains. Mechanical Turk was used in study to tag the dataset. Results show that by incorporating various constraints and by coupling the training of multiple data extractors, results are improved. In addition to that, it is also validated empirically that coupling can solve the problem of semantic drift that usually occurs in bootstrapping-based approaches.

Another approach proposed in Reference [117] learns semantic patterns and partial patterns to perform relation extraction using initial seed patterns. Using identified patterns and noun labels, SVM classifier is trained to acquire relations. Results are reported on TSUBAKI Japanese corpus, which shows that proposed approach produces good results.

A tensor decomposition-based approach is proposed in Reference [118], which creates embeddings for knowledge-base, where tensor decomposition is a mathematical tool that is used for data analysis. This study primarily uses recent advancements in RE approaches. These include integration of entity-relation triples, from variety of textual data as well as KBs. Here, each triple represents a fact. Using this data, a tensor-based decomposition approach is presented. The proposed approach is computationally very efficient as well as relatively scalable than existing approaches. Furthermore, it also uses domain knowledge about various types of entities. Hence, it can consequently determine new relations that are absent from the existing relation databases. In addition to that, the proposed system performs at par with existing systems that are comparatively computationally expensive. The proposed KB embeddings are used in task of RE on dataset used in Reference [119] and it results into weighted Mean Average Precision value of 57.

5.2.4 Distantly Supervised Approaches. A distant supervision (DS) approach is presented in Reference [120] that does not require any labeled data; instead it relies on an existing KB that is used as primary source of information. Here, it is assumed that if a sentence holds two entities having a relation, then all sentences carrying those two entities will have the similar relation. As this assumption will lead to many noisy features, a logistic regression classifier for multi-class is trained that learns the features of noise. Moreover, experiments are carried out with lexical features, syntactical features, and both. Mechanical Turk was used to prepare the test set for human evaluation.

Another approach to deal with primary limitation of distant learning, i.e., noisy data [121], uses graphical models. An undirected graphical model is proposed to predict relationship between entities along with sentences that can carry such relations. Training of this model is performed using constrained-driven semi-supervised approach via SampleRank algorithm [122].

Another distant learning approach proposed in Reference [123] improves distant supervision approach by means of improving relation label assignment. Due to the nature of DS, it extracts relations but labels them incorrectly, which in turn affect precision. To deal with the issue of incorrect labeling, a heuristic labeling process by means of a novel generative model is proposed. This model incorporates latent variables and employs expectation maximization algorithm to generate model. In addition to that, multi-class logistic regression is used to learn the relations. The model is evaluated on Wikipedia articles using Freebase as KB. Another approach to improve DS algorithm by means of improving features is proposed in Reference [124].

A rather different approach for RE is followed in Reference [119], which revolves around concept of Universal Schema, i.e., collection of multiple schemas. These schemas carry relations from Open IE relations as well as from structured sources. Main hypothesis of this study is to predict source data rather than semantic equivalence; thus, KB data is modeled with text by means of collaborative filtering. In addition to that, a probabilistic model is represented via matrix with rows depicting entity pairs and columns representing their respective relationship. The proposed matrix factorized model learns the latent feature vectors for entities and their respective relation as well. Further, to increase the relation confidence for any tuple, neighborhood model is being employed along with matrix factorization. To train the model, SGD is used. Results are evaluated on self-sampled collection of New York Times Corpus using MAP metric.

5.2.5 Unsupervised Approaches. There exist some unsupervised approaches for RE as well. K-means clustering-based approach is presented in Reference [125] to identify important contextual words that will eventually help in determining various relation types. Several feature ranking mechanisms such as frequency and entropy are employed to select these words. Results are evaluated using ACE corpus, where corpus is divided into three main classes: PER-ORG, ORG-GPE, and ORG-ORG, respectively. Relatedness measure is used to evaluate the relation extraction task. Results show that the proposed approach is useful in determining subset of features and to estimate number of context clusters.

Another approach using hierarchical clustering for paraphrase extraction makes use of keyword identification in Reference [126]. The process is divided into multiple phases. First, one identifies keywords in phrases and later joins the phrases having similar keywords into one collection. Second stage deals with linking those collections having similar individual NEs. This study uses chunking information rather than full parse, which is considered as a potential future direction as per the authors. Four different newswire corpora published in 1995: the *Los Angeles-Times, The New York Times,* Reuters, and the *Wall Street Journal* are used as datasets.

An unsupervised approach using language models is proposed in Reference [127] to extract sparse relations, where language model consists of HMM and N-gram model. HMM model is used for type-checking, i.e., to check for the correctness of relation attributes, whereas relational n-gram

model is primarily used as assessment module, which allows estimating context distributions for any pair of arguments. Proposed approach outperforms existing systems when sparse relations were extracted. In addition to that, average precision of 85.1 is achieved on manually prepared dataset from different web pages.

Another unsupervised approach is presented in Reference [95] that consists of three major modules: Self-Supervised Learner, which classifies relations as trustworthy or not; Single Pass Extractor, which extracts all possible tuples depicting some relation in a single-pass of entire corpus; and last, Redundancy-based Assessor, which assigns probability to every tuple. This system is called TextRunner, and evaluations are performed on a 9M page dataset. Comparison is made with KnowItall [94] system, which is a closed Information Extraction system and results show that TextRunner outperforms KnowItAll.

Unsupervised relation extraction from web is studied in Reference [128], which makes use of entity similarity graphs and hypernymy graphs, whereas DIRT algorithm is used to calculate the pairwise similarity graph for phrases depicting some relation information. Furthermore, it proposed a Web Relation Extraction (WEBRE) system that carries two major modules. In the first module, discovery of semantic classes is performed resulting in large set of relations. Then, second module deals with grouping similar relations together. To handle polysemy of objects, HAC is employed. Evaluation is performed on subset of Cluewebset 09 dataset containing 503M pages.

5.2.6 Joint Modeling Approaches. Joint Modeling refers to the modeling approach where multiple related approaches are modeled together, and each sub-problem being modeled results in resultant performance improvement. Many state-of-the-art approaches for RE employ joint modeling approaches.

Research presented in Reference [129] uses linear programming to perform RE. It primarily uses a variant of Viterbi algorithm having some capabilities of CRF, and makes use of multiple features and constraints to perform classification. Furthermore, to identify entities and relations, inference is also being applied. Experiments are conducted on subset of sentences acquired from TREC documents having at least one relation. Results are reported using individual models for tasks and joint model, whereas upper bound is calculated using an omniscient classifier.

Study presented in Reference [130] makes use of graphical model to perform RE. Graphs are used to reduce the joint extraction tasks to mere joint node labeling task. A binary directed graph is used that encodes every possible entity and relation that could exist in any sentence. To mark no relation and no entity, *NR* and *Other entity* marks are used. Proposed parsing algorithm uses beam searching to maintain a queue of elements at every graph node, with leaf nodes storing entity label with its respective probability as beam elements. SVM classifiers and kernels are also employed in parsing algorithms where classification is required. Features such as POS tags and words in neighboring context are also used. Results are reported on dataset prepared in Reference [129] containing three NEs: location, person, and organization.

Another graph-based model primarily making use of CRFs and semi-Markov chains is presented in Reference [131]. This approach also proposes a new algorithm, namely, collective iterative classification, to find optimal relations assignments in an iterative fashion. This algorithm consists of two steps, namely, bootstrapping and iterative classification. Results are evaluated on Wikipedia collection of 441 pages that were manually tagged. Results were compared with other variants of CRF, such as linear combination of CRFs (pipe-line model) and joint model based on CRF. Proposed model outperforms existing results of relation extraction.

A rather different approach based on history information is proposed in Reference [132] via structured learning. The proposed approach employs tables to manage history information. Here, mapping of the table is performed between each table cell and the respective entity label.

Furthermore, to map the input word sequence into tables, first of all, table transformation is required. This is achieved through static ordering that transforms the tabular structure into a one-dimensional form. Furthermore, while adding further labels in the table cells, existing cell assignments are also being considered. This consequently avoids any illegal assignment. After transformation of input into tabular form, structured learning approach is employed for training, using margin. In addition, multiple training algorithms are used including AdaGrad [133], Perceptron [134], AROW [135, 136], and DCD-SSVM [137]. After training, learned weights eventually aid in performing mapping of relations and respective entities in a table. Results are reported on corpus described in Reference [129]. This study transforms joint model of relation extraction into a table-filling problem where selection of learning methodologies and searching order heavily impacts the performance of resultant model.

5.3 Neural-network-based Approaches

This section covers the developments made by means of neural frameworks. These are further classified based on underlying technique that is carried out to perform RE such as distant supervision, joint modeling, and transfer learning.

5.3.1 Supervised Approaches. A research study reported in Reference [138] employs a Matrix Vector Recursive Neural Network (MV-RNN). Proposed solution is focused to learn vector embeddings for phrases as well as sentences, instead of learning word vector representations. Each sentence word is represented by means of word vector along with a matrix of parameters. In this study, word vectors used are 50-dimensional and trained using Wikipedia data, whereas each matrix is initialized with identity and some Gaussian noise. To train the classifier, cross-entropy error function is used along with Softmax classifier to learn the corresponding weights. Eventually every vector associated with each node in a sentence captures its semantic and syntactic features. Matrix, however, captures the resultant transitions in meanings of contextual words caused by a word itself. This approach is used to predict the movie review ratings, which resulted in 79 accuracy against movie-review data. On relation classification task, MV-RNN resulted in f-score of 79.1 on Sem-Eval task. By using three additional features of POS tags, NER tags of words and WordNet hypernyms, f-score of 82.4 was achieved on Sem-Eval 2010, which outperformed the state-of-the-art RE solutions.

CNN is being widely used for IE problems. Study employing CNN for RE is presented in Reference [139], which majorly relies on multiple features. The proposed system makes use of external knowledge resources, e.g., WordNet, position-related features, and word pairs information along with set of lexical features and word embeddings. To extract sentence level features, CNN classifier is employed that makes use of word embeddings as well as word position embeddings. After extraction of all these features, they are being concatenated to form a vector. Later, Softmax classifier is employed that takes this concatenated feature vector as input and in return performs relation classification. Sem-Eval 2010 dataset is used for evaluation and experiments result in f-score of 82.7, outperforming existing state-of-the-art systems for RE.

Research study in Reference [140] presents another approach for RE using CNN. This study incorporates multi-level attention model in CNN to capture attentions that are specific to entity as well as relation. To improve the overall network, study has proposed a new objective function that is based on margin-based distance. Model takes word vector embeddings to incorporate lexical-semantic features, and word position embeddings to incorporate relative distances from already marked entities within a sentence. Results are evaluated on Sem-Eval 2010 Task 8 and f-score of 88 is achieved, outperforming other systems for RE. Another study that employs RNN to perform RE on SemEval 2010 dataset is reported in Reference [141], which results in f-score of 79. This

study employs skip-gram model to learn word embeddings. A recent survey study carried out in Reference [97] extensively covers various variation of CNN that are briefly covered in Section 5.1.

In addition to these research studies, recent Sem-Eval task 2017 focuses on information extraction from scientific publications. One of the key problems in this task was relation extraction among identified NEs along with their classes. Total of three major NEs are used to annotate the open-access articles from ScienceDirect. These NEs include task, process, and material. The results of this task conclude that NN-based approaches perform better than non-neural network approaches for relation extraction. CNN gives the best results with incorporation of rule-based processing and argument ordering strategy. Apart from CNN, RNN and rule-based classification approaches are also used to solve this task. The brief overview of data generation, corpus details and results are presented in Reference [7]. Another study for multiple relation extraction using LSTMs is presented in Reference [142].

Capsule networks are also recently used for feature clustering in multi-labeled RE system [143]. Initial layer of the proposed system deals with learning semantic information using Bi-LSTM units, followed by feature clustering layer based on attention and capsule networks. Proposed framework is applied on NYT and SemEval-2010 for evaluation. Proposed approach is tailored to deal with highly overlapping relations and is pioneer in employment of capsule networks for the task of RE.

5.3.2 Distantly Supervised Approaches. A deep learning–based distant supervised approach using LSTMs with attention mechanism is proposed in Reference [144] that helps in mitigate shortcomings of previous distant supervised learning approaches. The problems it addresses include avoidance of existing false positives in labeled data as well as efficiency achievement by means of not adopting human-designed rules during extraction process. This study thus makes use of word-level attention features to extract relations along with instance-level attention mechanism to deal with the issue of false positives in data. Results are compared with various feature-based as well as neural-based methods on dataset used in Reference [121], and results show that proposed approach outperforms all other approaches.

A CNN-based approach that employs Jaccard-based similarity coefficient to minimize wrong label assignment during DS is presented in Reference [145]. In total, four relations are being explored in this study that include created, isAffiliatedTo, diedIn, and wasBornIn. Experiments on randomly selected Wikipedia articles show that the proposed approach employing word embeddings and semantic Jaccard similarity resulted in average accuracy of 86.2. However, on *New York Times* dataset, the proposed approach resulted in accuracy of 77.3.

A study employing piece-wise CNN is presented in Reference [146], which uses word-level model for attention to identify crucial words. This approach favors the important words and consequently results in improved precision and recall. Other CNN-based approaches for DS evaluated on this dataset are presented in Reference [147], which makes use of sentence and word embeddings along with multi-path CNNs. Experiments show that the proposed approach resulted in precision of 77.

5.3.3 Joint Modeling Approaches. There exist many joint modeling approaches based on neural frameworks. One such approach is proposed in Reference [148]. It uses structured learning along with history for entity extraction. Structured perceptron with beam search is employed to fulfill the desired task. It further uses local and global features to extract entity mentions and relation mentions.

An approach discussed in Reference [149] employs Bi-Directional LSTM with entity recognition conceived as Sequence Labeling problems. Model consists of three layers, including embeddings, sequence, and dependency. Embedding layer primarily deals with embeddings of word, dependency types, POS tags, and entity labels. Sequence layer consists of Bi-Directional

LSLTM to effectively maintain word sequence within a sentence. This layer is followed by an entity detection phase by means of a Neural Network, where entity labels are assigned in greedy fashion from left to right. Label of last word is used to predict the current word label. This is followed by extraction of relations between extracted entities.

Research study presented in Reference [88] also uses Bi-Directional LSTM to perform relation classification. It uses character-level word embeddings followed by performing NER using Bi-LSTM neural network model. After identifying NEs, relation classification is performed using another Bi-LSTM neural network model that exploits information of Shortest Dependency Paths as well. This system takes word embeddings, word POS tag, and character-level embeddings (processed via CNN) against respective word as input. Later, using stacked LSTMs, first LSTM layer deals with NER task and second layer deals with relation classification, thus a joint model is proposed in this study. Results are reported on dataset of 1,644 abstracts of PubMed data. Experiments show P/R/F of (67.5, 75.8, 71.4).

5.3.4 Transfer Learning. Collection of KBs are used in Reference [150] to perform transfer learning for RE. This work is focused to develop new KBs, given scarce resources and other relevant KBs in any particular domain of interest. To start with, using available text, initial weights are assigned. As suitability of a KB to any domain is dependent on its semantics, two techniques are used to identify the KBs. First technique determines correlation between KB and target corpus. Whereas, second focuses on discriminative ability of KB in a latent space. After selecting relevant KBs, risk minimization and approximate estimation techniques are used to perform domain-aware transfer learning. By means of a domain correlation and discriminative ability concepts, most relevant KBs are selected. Extensive experimentation is performed with increasing size of KBs on DBPedia, Wiki-KBP [151], and NYT datasets. Here, DBPedia KBs are used to train the model, whereas the remaining two are used for evaluation. As the size of KBs grow, performance increases. But, with the inclusion of too many irrelevant KBs, performance starts to decrease. Best reported results are achieved when total of 25 KBs were employed.

A transfer learning approach to deal with low quality and noisy data using distant supervision is presented in Reference [152]. The main idea is to initialize the network weights using entity weights to deal with noise. Words and position embeddings are major inputs that are being modeled using sub parse trees. This is followed by employment of hierarchical attention frameworks at both sentence- and work-level to improve the results. Finally, the relation prediction layer employs GRU units. This approach basically performs parameter-level transfer learning.

5.4 Conclusion

Relation extraction serves as baseline for many advanced text-based intelligent tasks, including question answering systems and text generation. A recent study makes use of keyphrase and relation extraction along with semantic information to generate survey [153]. State-of-the-art researches to perform relation extraction include end-to-end modeling, which does not require any additional information. In addition to that, recent studies have employed joint models for relation extraction, which improve its performance by solving the task of NER and RE in a joint fashion. Deep Learning frameworks are being widely used for RE and perform significantly better than other approaches when multi-label instances are also incorporated. Table 8 presents the summary of studies included in review. It can be observed that joint modeling approaches tend to improve the performance on entity recognition as well. In following table, three figures in evaluation metrics denote P/R/F, respectively. Hence, if study has reported any other metric or subset of these metrics, it is explicitly mentioned in Evaluation Metrics column. Furthermore, if

Table 8. RE Literature Summary

Study	Technique	Evaluation Metrics			Features/ Model Properties	Dataset/ Genre
		P	R	F		
[105]	MEMM			52.8	Lexical, Semantic and Syntactic	ACE'02
				55.2		ACE'03
[88]	Bi-LSTM	67.5	75.8	71.4	Stacked LSTM Model	PubMed abstracts
[99]	Heuristic	68	83	75	Conjunction, Negation	LLL'05 workshop
[101]	Heuristic	75.5	62.1	68.1	Syntactic Parser, DBPedia	Quaero News
[107]	SVM	77.2	60.7	68.0	Lexical, Semantic, Syntactic, External Lexicon	ACE'03
[109]	SVM	82.7	91.3	86.0	Kernels and voted perceptron	200 newswire and publications
[110]	SVM	70.3	26.3	38.0	Tree Kernel	ACE
[111]	SVM	76.1	68.4	72.1	Tree Kernel	ACE'03
[113]	Bootstrapping with SVM	63.2	61.5	60.3	Radial Bias Kernel	Self-annotated
[115]	CRF	73.4	56.1	63.6	Relational pattern features. Word, external	Wikipedia articles
[94]	SS				BootStrapping, Ontology	Sports and Companies web pages
[117]	SVM				Semantic Classes, Partial Pattern	TSUBAKI
[118]	SS	57.0			KBs, Tensor Decomposition	New York Times dataset [119]
[119]	Collaborative Filtering	69.0			KB, Universal Schemas	New York Times dataset
[116]	Multi-class Logistic Regression	68.0			KBs, Lexical and Syntactical Features	Self-annotated using Mechanical Turk
[121]	DS	87.0			SampleRank, CRF, FreeBase	New York Times dataset
[123]	Logistic Regression	78.2	68.2	66.7	Freebase, EM	Wikipedia articles
[144]	LSTM				Attention Mechanism	NYT
[145]	CNN	Accuracy: 86.2 77.3			Semantic Jaccard	Wikipedia Articles New York Times
[146]	Piece-wise CNN	46.9	44.5	45.7	Word-level attention model	NYT [121]
[147]	Multi-path CNN	77.0			Word and sentence level attention model	NYT [121]
[154]	Clustering	77.5	78.5	77.5	Hierarchy NER, Complete Linkage	NYT
[125]	US				Unsupervised Feature Subset Selection, K-means	
[126]	US	Accuracy: 79.5			Chunking Information, Hierarchical Clustering	News
[127]	HMM	85.1				Web-pages
[128]	US	89.7	68.4	77.6	Hierarchical Clustering	Cluewebset'09

(Continued)

Table 8. Continued

Study	Technique	Evaluation Metrics			Features/ Model Properties	Dataset/ Genre
		P	R	F		
[138]	RNN	82.4		82.4	POS Tags, NER Tags, Wordnet Hypernyms	SemEval 2010
[139]	CNN	82.7		82.7	Wordnet	SemEval 2010
[140]	CNN	88.0		88.0	Multi-level Attention Model	SemEval 2010
[141]	RNN	79.0		79.0	Skip-gam-based Word Vectors	SemEval 2010
[142]	LSTMs	72.9	70.8	67.9	Dynamic models	CONLL'04
[129]	Viterbi	54.0	68.4	58.14	Inferencing	TREC documents
[130]	Joint Model	90.1 73.0	91.8 62.7	91.3 66.0	POS Tags, Context Words, Hybrid Model including SVM, CYK-Parsing	TREC documents [129]
[155]	Joint Model	94.0 76.0			Graph	New York Times data
[131]	Joint Model	93.4 72.6	93.4 64.3	93.4 68.2	BootStrapping with Markov Models and CRF, Joint Model	Wikipedia
[148]	Joint Model	83.5 64.7	76.2 38.5	79.7 48.3	Casing, Gazetteer, Relation Features, Perceptron	ACE'04
		85.2 68.9	76.9 41.9	80.8 52.1		ACE'05
[132]	Joint Model	92.4 83.7	92.4 59.9	92.4 69.8	History Info., Structured Learning	TREC documents [129]
[149]	Joint Model	80.8 48.7	82.9 48.1	81.8 48.4	Bi-directional LSTM	ACE'04
		82.9 57.2	83.9 54.0	83.4 55.6		ACE'05
[143]	LSTM, Capsule Networks	30.8	63.7	41.6	Attention re-routing, position embedding	NYT
				84.5		SemEval-2010
[150]	Transfer Learning				Knowledge bases	Wiki-KBP NYT

study has presented a joint model, in that case, results on NER and RE are being reported together, with first row highlighting NER and second row highlighting RE results.

6 CONCLUSION

Although brief summaries and conclusions are expressed in each section, this particular section presents the overall big picture of recent trends in domain of IE along with open research areas. This survey study presents state-of-the-art in domain of IE with major focus on advances via deep learning approaches. This review study is majorly focused on two sub-domains of IE that include NER and RE. Papers were collected using major research repositories including ACM, IEEE, DBLP, and Google Scholar.

In the light of surveys conducted, recent trend in context of IE domain is the employment of joint modeling, which tends to model multiple problems simultaneously in a joint fashion. This type of modeling results in improved precision and recall. One primary reason for this accuracy gain is that IE tasks are dependent on one another. If task earlier in pipeline is not executed well, it will eventually affect the tasks to come. Thus, by jointly modeling the tasks, the error of overall system decreases. Majority systems, in context of joint modeling, model NER and RE in a joint

fashion. These joint models make use of various machine learning techniques including graphical models, linear programming, as well as deep learning approaches.

Historically, statistical classification approaches were common for NER as well as RE. But recently, after the advent of word vectors, deep learning is being widely applied to solve IE-oriented tasks. Recent approaches for NER employ recurrent neural networks as well as convolution neural networks in combination with CRF. These hybrid models currently give state-of-the-art results against majority benchmark datasets for NER. End-to-end modeling is also being applied nowadays that deals with solution of a problem given the input and output requirements without making use of any other information. These end-to-end models are majorly powered via deep learning approaches. RE, however, is currently being widely addressed using distant supervision approaches as well as deep learning–based joint models. Currently, rule-based approaches are still in use for languages lacking linguistic resources. Such approaches require time to write rule-bases or to learn heuristics using data at hand. Also, with increase in rules, system complexity is increased.

Major challenges in IE-oriented tasks is the feature selection and hyper-parameter tuning. The study presented in Reference [86] provides a thorough experimentation and describes the effect of various network architectures decision on overall system accuracy for sequence labeling tasks. These types of studies against other machine learning models and various problems are among open research questions in the domain of IE.

In addition, various approaches used to perform IE above have their own pros and cons. Some of these techniques are computationally very expensive, such as CRF and RNN. Complexity of majority techniques depends on how the technique is modeled, which type of smoothing, post-processing, and regularization parameters are used. Additionally, most of these approaches are offline learning approaches. It means that whenever a new type of data feature or new data genre has to be incorporated in model training, complete model needs to be re-trained, which in itself is a memory- and computation-hungry process. Thus, online training methods, which have ability to be trained incrementally, are one of the open areas under IE techniques.

Another area that is relatively under-addressed in RE is how to effectively deal with the problem of nested relations. Just like in NER, there are some real-life problems that require nested RE. One such study dedicated to perform nested RE for Chinese bond prospectus data is presented in Reference [156], which employs deep learning frameworks in an iterative fashion.

Keeping the application of deep learning approaches and their increasing employment across multiple domains in view, it is also of interest to design techniques that efficiently utilize energy and provide low-cost solutions while maintaining the accuracy of current systems. These types of solutions require understanding of the datasets at hand along with the idea about current and future computational requirements, as discussed in Reference [157]. Such type of trend analysis based on state-of-the-art employment of DNN in IE can serve as an open future direction.

As the whole domain of IE is highly dependent upon the datasets at hand, nature of datasets is extremely important. Currently, major literature is dedicated towards well-structured natural languages text as major benchmark datasets. Most of the data sources include collection of news articles as well as scientific articles. Due to advent of World Wide Web and social media, informal and ill-structured text is also increasing. With studies being performed on Twitter data to extract keywords and NEs in recent years, research on informal text and respective benchmark datasets are one of the major open areas in this domain.

Another major open research area is related to evaluation measures. Mostly evaluation measures perform exact matching or partial matching between systems-generated results and gold-standard results. There is dire need to incorporate semantic information to identify and rate semantically equivalent results. Prevailing and widely used metrics in current domain of IE does not exploit semantic similarity. Thus, design of efficient evaluation metrics for IE is of immense importance.

REFERENCES

[1] I. Muslea et al. 1999. Extraction patterns for information extraction tasks: A survey. In *Proceedings of the AAAI Workshop on Machine Learning for Information Extraction*.

[2] G. Simoes, H. Galhardas, and L. Coheur. 2009. Information extraction tasks: A survey. In *Proceedings of the INForum*.

[3] Linguistic Data Consortium. 2017. MUC Data Sets. Retrieved from http://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html.

[4] A. Rodriguez. 2017. MUC - Cohen Courses. Retrieved from http://curtis.ml.cmu.edu/w/courses/index.php/MUC.

[5] Linguistic Data Consortium. 2002. Annotation Tasks and Specifications. Retrieved from https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications.

[6] National Institute of Standards and Technology (NIST). 2017. TAC Knowledge Base Population (KBP). In *Proceedings of the Text Analytic Conference*.

[7] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum. 2017. SemEval 2017 Task 10: ScienceIE - extracting keyphrases and relations from scientific publications. *ArXiv170402853 Cs Stat*, Apr. 2017.

[8] L. Neve. 2019. GENIA Corpus. *The ORBIT Project*. Retrieved from https://orbit.nlm.nih.gov/browse-repository/dataset/human-annotated/83-genia-corpus.

[9] R. Merchant, M. E. Okurowski, and N. Chinchor. 1996. The multilingual entity task (MET) overview. In *Proceedings of a Workshop on held at Vienna, Virginia: May 6–8, 1996*. 445–447. DOI: 10.3115/1119018.1119075

[10] Asian Federation of Natural Language Processing. 2008. IJCNLP-08 Workshop on NER for South and South East Asian Languages. Retrieved from http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5.

[11] M. K. Malik. 2017. Urdu named entity recognition and classification system using artificial neural network. *ACM Trans Asian Low-Resour Lang. Inf. Proc.* 17, 1 (2017), 2:1–2:13. DOI: 10.1145/3129290

[12] N. Peng and M. Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 548–554.

[13] W. Wang, F. Bao, and G. Gao. 2015. Mongolian named entity recognition using suffixes segmentation. In *Proceedings of the International Conference on Asian Language Processing (IALP'15)*. 169–172. DOI: 10.1109/IALP.2015.7451558

[14] D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investig.* 30, 1 (2007), 3–26. DOI: 10.1075/li.30.1.03nad

[15] N. Kanya and T. Ravi. 2012. Modelings and techniques in named entity recognition-an information extraction task. In *Proceedings of the IET Chennai 3rd International on Sustainable Energy and Intelligent Systems (SEISCON'12)*. DOI: 10.1049/cp.2012.2199

[16] G. K. Palshikar. 2013. Techniques for named entity recognition. *Bioinforma. Concepts Methodol. Tools Appl.* 400 (2013).

[17] R. Sharnagat. 2014. Named entity recognition: A literature survey. Report 11305R013. Cent. Indian Lang. Technol.

[18] N. Patil, A. S. Patil, and B. Pawar. 2016. Survey of named entity recognition systems with respect to Indian and foreign languages. *Int. J. Comput. Appl.* 134, 16 (2016).

[19] L. Ratinov and D. Roth. 2019. *Design challenges and misconceptions in named entity recognition*. 147–155. Retrieved from http://dl.acm.org/citation.cfm?id=1596374.1596399.

[20] N. Rizzolo and D. Roth. 2007. Modeling discriminative global inference. In *Proceedings of the International Conference on Semantic Computing (ICSC'07)*. 597–604.

[21] D. Klein, J. Smarr, H. Nguyen, and C. D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL, Volume 4*. 180–183. DOI: 10.3115/1119176.1119204

[22] G. Luo, X. Huang, C.-Y. Lin, and Z. Nie. 2015. Joint named entity recognition and disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. 879–880.

[23] D. B. Nguyen, M. Theobald, and G. Weikum. 2016. J-NERD: Joint named entity recognition and disambiguation with rich linguistic features. *Trans. Assoc. Comput. Linguist.* 4 (2016), 215–229. DOI: 10.1162/tacl_a_00094

[24] W. Liao and S. Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing*. 58–65. Retrieved from http://dl.acm.org/citation.cfm?id=1621829.1621837.

[25] O. Etzioni et al. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.* 165, 1 (2005), 91–134. DOI: 10.1016/j.artint.2005.03.001

[26] D. Nadeau, P. Turney, and S. Matwin. 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Adv. Artif. Intell. Lecture Notes in Computer Sciences, vol. 4013*. Springer, 266–277. DOI: 10.1007/11766247_23

[27] I. Gallo, E. Binaghi, M. Carullo, and N. Lamberti. 2008. Named entity recognition by neural sliding window. In *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*. 567–573. DOI: 10.1109/DAS.2008.13

[28] A. Passos, V. Kumar, and A. McCallum. 2017. Lexicon infused phrase embeddings for named entity resolution. *ArXiv14045367 Cs*, Apr. 2014.

[29] M. Peters, W. Ammar, C. Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1756–1765. DOI: 10.18653/v1/P17-1161

[30] M. Peters et al. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2227–2237. DOI: 10.18653/v1/N18-1202

[31] M. Rondeau and Y. Su. 2015. Full-rank linear-chain NeuroCRF for sequence labeling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 5281–5285. DOI: 10.1109/ICASSP.2015.7178979

[32] M. A. Rondeau and Y. Su. 2015. Recent improvements to NeuroCRFs for named entity recognition. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'15)*. 390–396. DOI: 10.1109/ASRU.2015.7404821.

[33] X. Ma and E. Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF. *ArXiv Prepr. ArXiv160301354*, 2016.

[34] E. Strubell, P. Verga, D. Belanger, and A. McCallum. 2017. Fast and accurate sequence labeling with iterated dilated convolutions. *ArXiv170202098 Cs*, Feb. 2017.

[35] F. Liu, T. Baldwin, and T. Cohn. 2017. Capturing long-range contextual dependencies with memory-enhanced conditional random fields. *ArXiv Prepr. ArXiv170903637*, 2017.

[36] K. Riaz. 2010. Rule-based named entity recognition in Urdu. In *Proceedings of the Named Entities Workshop*. 126–135. Retrieved from http://dl.acm.org/citation.cfm?id=1870457.1870476.

[37] R. Alfred, L. C. Leong, C. K. On, and P. Anthony. 2014. Malay named entity recognition based on rule-based approach. *Int. J. Mach. Learn. Comput.* 4, 3 (2014), 300.

[38] D. M. Bikel, R. Schwartz, and R. M. Weischedel. 1999. An algorithm that learns what's in a name. *Mach. Learn.* 34, 1–3 (1999), 211–231. DOI: 10.1023/A:1007558221122

[39] R. Ageishi and T. Miura. 2008. Named entity recognition based on a hidden Markov model in part-of-speech tagging. In *Proceedings of the 1st International Conference on the Applications of Digital Information and Web Technologies (ICADIWT'08)*. 397–402. DOI: 10.1109/ICADIWT.2008.4664380

[40] A. E. Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. Dissertation. New York University, New York, NY.

[41] A. McCallum and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, Volume 4. 188–191. DOI: 10.3115/1119176.1119206

[42] G. Fu and K.-K. Luke. 2005. Chinese named entity recognition using lexicalized HMMs. *SIGKDD Explor. Newsl.* 7, 1 (2005), 19–25. DOI: 10.1145/1089815.1089819

[43] X. Yu, S. Mayhew, M. Sammons, and D. Roth. 2019. On the strength of character language models for multilingual named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics*. 3073–3077. Retrieved from https://aclweb.org/anthology/papers/D/D18/D18-1345/.

[44] D. Khashabi et al. 2018. Cogcompnlp: Your Swiss army knife for NLP. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*.

[45] S. Strassel and J. Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. 3273–3280.

[46] E. Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Comput. Linguist.* 21, 4 (1995), 543–565.

[47] Q. L. L. Buco, J. L. L. Capcap, J. C. A. Hermocilla, C. S. Yumul, R. A. Sagum, and A. G. Pastrana. 2013. The application of transformation-based learning in the development of a named entity recognition system for Filipino text. *J. Ind. Intell. Inf.* 1, 1 (2013).

[48] S. Cucerzan and D. Yarowsky. 2002. Language independent NER using a unified model of internal and contextual evidence. In *Proceedings of the 6th Conference on Natural Language Learning*, Volume 20. 1–4. DOI: 10.3115/1118853.1118860

[49] R. A. Leonandya, B. Distiawan, and N. H. Praptono. 2015. A semi-supervised algorithm for Indonesian named entity recognition. In *Proceedings of the 3rd International Symposium on Computational and Business Intelligence (ISCBI'15)*. 45–50. DOI: 10.1109/ISCBI.2015.15

[50] J. Straková, M. Straka, and J. Hajič. 2016. Neural networks for featureless named entity recognition in Czech. In *Proceedings of the International Conference on Text, Speech, and Dialogue*. 173–181.

[51] L. Liu et al. 2018. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[52] X.-D. Doan, T.-T. Dang, and M. L. Nguyen. 2019. Effectiveness of character language model for Vietnamese named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Retrieved from https://aclweb.org/anthology/papers/Y/Y18/Y18-1018/.

[53] C. Lee. 2017. LSTM-CRF models for named entity recognition. *IEICE Trans. Inf. Syst.* 100, 4 (2017), 882–887.

[54] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT, Association for Computational Linguistics*. 260–270. DOI: 10.18653/v1/N16-1030

[55] W. Wang, F. Bao, and G. Gao. 2016. Mongolian named entity recognition with bidirectional recurrent neural networks. In *Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI'16)*. 495–500. DOI: 10.1109/ICTAI.2016.0082

[56] O. Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*. Retrieved from http://dl.acm.org/citation.cfm?id=2390426.2390435.

[57] O. Täckström, R. McDonald, and J. Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Retrieved from http://dl.acm.org/citation.cfm?id=2382029.2382096.

[58] L. Qu, G. Ferraro, L. Zhou, W. Hou, and T. Baldwin. 2016. Named entity recognition for novel types by transfer learning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*. 899–905. DOI: 10.18653/v1/D16-1087

[59] L. Chen, A. Moschitti, G. Castellucci, A. Favalli, and R. Romagnoli. 2018. Transfer learning for industrial applications of named entity recognition. In *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI'18) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018)*. 129–140. Retrieved from http://ceur-ws.org/Vol-2244/paper_12.pdf.

[60] S. Mayhew, C.-T. Tsai, and D. Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*. 2536–2545. DOI: 10.18653/v1/D17-1269

[61] R. Murthy, M. M. Khapra, and P. Bhattacharyya. 2018. Improving NER tagging performance in low-resource languages via multilingual learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Proc.* 18, 2 (2018), 9:1–9:20. DOI: 10.1145/3238797

[62] X. Feng, X. Feng, B. Qin, Z. Feng, and T. Liu. 2018. Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 4071–4077. Retrieved from http://dl.acm.org/citation.cfm?id=3304222.3304336.

[63] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. 182–192. Retrieved from https://aclweb.org/anthology/papers/D/D18/D18-1017/.

[64] A. Rahimi, Y. Li, and T. Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. 151–164.

[65] A. Johnson, P. Karanasou, J. Gaspers, and D. Klakow. 2019. Cross-lingual transfer learning for Japanese named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. 182–189.

[66] A. Bharadwaj, D. Mortensen, C. Dyer, and J. Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1462–1472.

[67] J. Xie, Z. Yang, G. Neubig, N. A. Smith, and J. G. Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 369–379. DOI: 10.18653/v1/D18-1034

[68] Z. Nasar, S. W. Jaffry, and M. K. Malik. 2018. Information extraction from scientific articles: A survey. *Scientometrics*. DOI: 10.1007/s11192-018-2921-5

[69] M. Abdelmagid, M. Himmat, and A. Ahmed. 2014. Survey on information extraction from chemical compound literatures: Techniques and challenges. *J. Theor. Appl. Inf. Technol.* 67, 2 (2014), 284–289.

[70] G. Duck, G. Nenadic, M. Filannino, A. Brass, D. L. Robertson, and R. Stevens. 2016. A survey of bioinformatics database and software usage through mining the literature. *PLoS One* 11, 6 (2016), e0157989. DOI: 10.1371/journal.pone.0157989

[71] B. Shickel, P. Tighe, A. Bihorac, and P. Rashidi. 2017. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *ArXiv Prepr. ArXiv170603446*, 2017.

[72] T. Eftimov, B. Koroušić Seljak, and P. Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS One* 12, 6 (2017), e0179488. DOI: 10.1371/journal.pone.0179488

[73] D. M. de Oliveira, A. H. F. Laender, A. Veloso, and A. S. da Silva. 2013. FS-NER: A lightweight filter-stream approach to named entity recognition on Twitter data. In *Proceedings of the 22nd International Conference on World Wide Web*. 597–604. DOI: 10.1145/2487788.2488003

[74] D. Bonadiman, A. Severyn, and A. Moschitti. 2015. Deep neural networks for named entity recognition in Italian. In *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC It'15)*.

[75] J. Xu, H. He, X. Sun, X. Ren, and S. Li. 2018. Cross-domain and semisupervised named entity recognition in Chinese social media: A unified model. *IEEE/ACM Trans. Audio Speech Lang. Proc.* 26, 11 (2018), 2142–2152. DOI: 10.1109/TASLP.2018.2856625

[76] Z. Zhao et al. 2016. ML-CNN: A novel deep learning based disease named entity recognition architecture. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'16)*. 794–794. DOI: 10.1109/BIBM.2016.7822625

[77] J. Li et al. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Datab.- J. Biol. Datab. Curat.* May (2016). DOI: 10.1093/database/baw068

[78] R. I. Doğan, R. Leaman, and Z. Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* 47 (2014), 1–10. DOI: 10.1016/j.jbi.2013.12.006

[79] X. Dong, L. Qian, Y. Guan, L. Huang, Q. Yu, and J. Yang. 2016. A multiclass classification method based on deep learning for named entity recognition in electronic medical records. In *Proceedings of the New York Scientific Data Summit (NYSDS'16)*. 1–10. DOI: 10.1109/NYSDS.2016.7747810

[80] L. Wang, S. Li, Q. Yan, and G. Zhou. 2018. Domain-specific named entity recognition with document-level optimization. *ACM Trans. Asian Low-Resour. Lang. Inf. Proc.* 17, 4 (2018), 33:1–33:15. DOI: 10.1145/3213544

[81] J. R. Finkel and C. D. Manning. 2009. Nested named entity recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Volume 1. 141–150. Retrieved from http://dl.acm.org/citation.cfm?id=1699510.1699529.

[82] W. Lu and D. Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 857–867.

[83] A. Katiyar and C. Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 861–871.

[84] M. Ju, M. Miwa, and S. Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1446–1459.

[85] K. Mai et al. 2018. An empirical study on fine-grained named entity recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*. 711–722.

[86] N. Reimers and I. Gurevych. 2017. Optimal hyperparameters for deep LSTM-Networks for sequence labeling tasks. Retrieved from https://arxiv.org/abs/1707.06799.

[87] S. Zhang and N. Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *J. Biomed. Inform.* 46, 6 (2013). DOI: 10.1016/j.jbi.2013.08.004

[88] F. Li, M. Zhang, G. Fu, and D. Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinf.* 18 (2017). DOI: 10.1186/s12859-017-1609-9

[89] N. Bach and S. Badaskar. 2007. A review of relation extraction. Unpublished Report. Retrieved from www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf.

[90] N. Konstantinova. 2014. Review of relation extraction methods: What is new out there? In *Analysis of Images, Social Networks and Texts*. 15–28. DOI: 10.1007/978-3-319-12580-0_2

[91] N. Asghar. 2016. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *ArXiv Prepr. ArXiv160507895*, 2016.

[92] S. Brin. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of the International Workshop on the World Wide Web and Databases*. 172–183.

[93] E. Agichtein and L. Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM Conference on Digital Libraries*. 85–94.

[94] O. Etzioni et al. 2004. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th International Conference on World Wide Web*. 100–110.

[95] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)*. 2670–2676.

[96] R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White. 2005. Simple algorithms for complex relation extraction with applications to biomedical IE. In *Proceedings of the 43rd Meeting on Association for Computational Linguistics*. 491–498. DOI: 10.3115/1219840.1219901

[97] S. Kumar. 2017. A survey of deep learning methods for relation extraction. *ArXiv170503645 Cs*, May 2017.

[98] A. Smirnova and P. Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Comput. Surv.* 51, 5 (2018), 106:1–106:35. DOI: 10.1145/3241741

[99] K. Fundel, R. Küffner, and R. Zimmer. 2006. RelEx—Relation extraction using dependency parse trees. *Bioinformatics* 23, 3 (2006), 365–371.

[100] C. Nédellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL'05)*. 31–37.

[101] Kamel Nebhi. 2013. A rule-based relation extraction system using DBpedia and syntactic parsing. In *Proceedings of the 2013th International Conference on NLP & DBpedia (NLP-DBPEDIA'13)*, Vol. 1064. 74–79. Retrieved from http://dl.acm.org/citation.cfm?id=2874479.2874487.

[102] S. Rosset, C. Grouin, K. Fort, O. Galibert, J. Kahn, and P. Zweigenbaum. 2012. Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proceedings of the 6th Linguistic Annotation Workshop*. 40–48.

[103] G. Leroy and H. Chen. 2001. Filling preposition-based templates to capture information from medical. In *Proceedings of the Pacific Symposium on Biocomputing*.

[104] C. Blaschke and A. Valencia. 2001. The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform.* 12 (2001), 123–134.

[105] N. Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. 22–25. DOI: 10.3115/1219044.1219066

[106] A. Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Mach. Learn.* 34, 1–3 151–175. DOI: 10.1023/A:1007502103375

[107] Z. GuoDong, S. Jian, Z. Jie, and Z. Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Meeting on Association for Computational Linguistics*. 427–434. DOI: 10.3115/1219840.1219893

[108] Y. S. Chan and D. Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Meeting of the Association for Computational Linguistics: Human Language Technologies*, Volume 1. 551–560. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002542.

[109] D. Zelenko, C. Aone, and A. Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, Volume 10. 71–78. DOI: 10.3115/1118693.1118703

[110] A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 423–429. DOI: 10.3115/1218955.1219009

[111] Z. Min, Z. GuoDong, and A. Aiti. 2008. Exploring syntactic structured features over parse trees for relation extraction using kernel methods. *Inf. Proc. Manag.* 44, 2 (2008), 687–701. DOI: 10.1016/j.ipm.2007.07.013

[112] K. Tymoshenko and C. Giuliano. 2017. FBK-IRST: Semantic relation extraction using cyc. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. 214–217. Retrieved from http://dl.acm.org/citation.cfm?id=1859664.1859711.

[113] Z. Zhang. 2004. Weakly-supervised relation classification for information extraction. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*. 581–588.

[114] P. Pantel and M. Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Meeting of the Association for Computational Linguistics*. 113–120.

[115] A. Culotta, A. McCallum, and J. Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. 296–303. DOI: 10.3115/1220835.1220873

[116] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr, and T. M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. 101–110. DOI: 10.1145/1718487.1718501

[117] S. De Saeger et al. 2011. Relation acquisition using word classes and partial patterns. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145524.

[118] K.-W. Chang, S. W. Yih, B. Yang, and C. Meek. 2014. Typed tensor decomposition of knowledge bases for relation extraction. Retrieved from https://www.microsoft.com/en-us/research/publication/typed-tensor-decomposition-of-knowledge-bases-for-relation-extraction/.

[119] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL'13)*. 74–84.

[120] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Volume 2. Retrieved from http://dl.acm.org/citation.cfm?id=1690219.1690287.

[121] S. Riedel, L. Yao, and A. McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 148–163. DOI: 10.1007/978-3-642-15939-8_10

[122] M. Wick, K. Rohanimanesh, K. Bellare, A. Culotta, and A. McCallum. 2011. SampleRank: Training factor graphs with atomic gradients. In *Proceedings of the 28th International Conference on Machine Learning*. 777–784. Retrieved from http://dl.acm.org/citation.cfm?id=3104482.3104580.

[123] S. Takamatsu, I. Sato, and H. Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Meeting of the Association for Computational Linguistics: Long Papers*, Volume 1. 721–729. Retrieved from http://dl.acm.org/citation.cfm?id=2390524.2390626.

[124] H. Zhang and Y. Zhao. 2013. Improving few occurrence feature performance in distant supervision for relation extraction. In *Advanced Data Mining and Applications*, H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, and W. Wang (Eds). Springer Berlin, 414–422.

[125] J. Chen, D. Ji, C. L. Tan, and Z. Niu. 2005. Unsupervised feature selection for relation extraction. In *Companion Volume to the Proceedings of Second International Joint Conference on Natural Language Processing, [Online]*. Retrieved from https://www.aclweb.org/anthology/I05-2045.

[126] S. Sekine. 2005. Automatic paraphrase discovery based on context and keywords between NE pairs. In *Proceedings of the International Workshop on Paraphrasing (IWP'05)*. 4–6.

[127] D. Downey, S. Schoenmackers, and O. Etzioni. 2017. Sparse information extraction: Unsupervised language models to the rescue. In *Proceedings of the Meeting of the Association for Computational Linguistics*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.130.8780.

[128] B. Min, S. Shi, R. Grishman, and C.-Y. Lin. 2012. Towards large-scale unsupervised relation extraction from the web. *Int. J. Seman. Web Inf. Syst.* 8, 3 (2012), 1–23. DOI: 10.4018/jswis.2012070101

[129] D. Roth and W. Yih. 2007. Global inference for entity and relation identification via a linear programming formulation. In *Introduction to Statistical Relational Learning*. The MIT Press, Cambridge, MA, 553–580.

[130] R. J. Kate and R. J. Mooney. 2010. Joint entity and relation extraction using card-pyramid parsing. In *Proceedings of the 14th Conference on Computational Natural Language Learning*. Retrieved from http://dl.acm.org/citation.cfm?id=1870568.1870592.

[131] X. Yu and W. Lam. 2010. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. 1399–1407. Retrieved from http://dl.acm.org/citation.cfm?id=1944566.1944726.

[132] M. Miwa and Y. Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1858–1869.

[133] J. Duchi, E. Hazan, and Y. Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* 12 (2011), 2121–2159.

[134] M. Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Volume 10. 1–8.

[135] A. Mejer and K. Crammer. 2010. Confidence in structured-prediction using confidence-weighted models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 971–981.

[136] K. Crammer, A. Kulesza, and M. Dredze. 2009. Adaptive regularization of weight vectors. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 414–422.

[137] M.-W. Chang and W. Yih. 2013. Dual coordinate descent algorithms for efficient large margin structured prediction. *Trans. Assoc. Comput. Linguist.* 1 (2013), 207–218.

[138] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 1201–1211.

[139] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the International Conference on Computational Linguistics (COLING'14)*. 2335–2344.

[140] L. Wang, Z. Cao, G. de Melo, and Z. Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1298–1307. DOI: 10.18653/v1/P16-1123

[141] S. Takase, N. Okazaki, and K. Inui. 2016. Modeling semantic compositionality of relational patterns. *Eng. Appl. Artif. Intell.* 50, C (2016), 256–264. DOI: 10.1016/j.engappai.2016.01.027

[142] J. Liu, H. Ren, M. Wu, J. Wang, and H.-J. Kim. 2018. Multiple relations extraction among multiple entities in unstructured text. *Soft Comput.* 22, 13 (2018), 4295–4305. DOI: 10.1007/s00500-017-2852-8

[143] X. Zhang, P. Li, W. Jia, and H. Zhao. 2019. Multi-labeled relation extraction with attentive capsule network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 33 (2019), 7484–7491.

[144] D. He, H. Zhang, W. Hao, R. Zhang, and K. Cheng. 2017. A customized attention-based long short-term memory network for distant supervised relation extraction. *Neural Comput.* 29, 7 (2017), 1964–1985. DOI: 10.1162/NECO_a_00970

[145] C. Ru, J. Tang, S. Li, S. Xie, and T. Wang. 2018. Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Inf. Proc. Manag.* 54, 4 (2018), 593–608. DOI: 10.1016/j.ipm.2018.04.002

[146] J. Qu, D. Ouyang, W. Hua, Y. Ye, and X. Li. 2018. Distant supervision for neural relation extraction integrated with word attention and property features. *Neural Netw.* 100, C (2018), 59–69. DOI: 10.1016/j.neunet.2018.01.006

[147] Y. Li, Z. Zhong, and N. Jing. 2018. Multi-path convolutional neural network for distant supervised relation extraction. In *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*. 119:1–119:7. DOI: 10.1145/3207677.3278063

[148] Q. Li and H. Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the Meeting of the Association for Computational Linguistics*. 402–412.

[149] M. Miwa and M. Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *ArXiv Prepr. ArXiv160100770*, 2016.

[150] S. Di, Y. Shen, and L. Chen. 2019. Relation extraction via domain-aware transfer learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1348–1357. DOI: 10.1145/3292500.3330890

[151] X. Ling and D. S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 94–100.

[152] T. Liu, X. Zhang, W. Zhou, and W. Jia. 2018. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2195–2204. DOI: 10.18653/v1/D18-1243

[153] S. Yang, W. Lu, D. Yang, X. Li, C. Wu, and B. Wei. 2017. KeyphraseDS: Automatic generation of survey by exploiting keyphrase information. *Neurocomputing* 224 (2017), 58–70. DOI: 10.1016/j.neucom.2016.10.052

[154] T. Hasegawa, S. Sekine, and R. Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. 415–422. DOI: 10.3115/1218955.1219008

[155] L. Yao, S. Riedel, and A. McCallum. 2010. Collective cross-document relation extraction without labelled data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1013–1023. Retrieved from http://dl.acm.org/citation.cfm?id=1870658.1870757.

[156] Y. Cao, D. Chen, H. Li, and P. Luo. 2019. Nested relation extraction with iterative neural network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1001–1010. DOI: 10.1145/3357384.3358003

[157] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *ArXiv Prepr. ArXiv170309039*, 2017.

Received February 2019; revised August 2020; accepted October 2020