Computer Science Review 29 (2018) 21–43

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Recent Named Entity Recognition and Classification techniques: A systematic review☆

Archana Goyal [a], Vishal Gupta [b,*], Manish Kumar [c]

[a] PG Department of Information Technology, GGDSD College, Chandigarh, India
[b] University Institute of Engineering & Technology, Panjab University, Chandigarh, India
[c] Panjab University Regional Centre, Muktsar, Punjab, India

ARTICLE INFO

Article history:
Received 2 December 2017
Received in revised form 1 June 2018
Accepted 2 June 2018
Available online 15 June 2018

ABSTRACT

Textual information is becoming available in abundance on the web, arising the requirement of techniques and tools to extract the meaningful information. One of such an important information extraction task is Named Entity Recognition and Classification. It is the problem of finding the members of various predetermined classes, such as person, organization, location, date/time, quantities, numbers etc. The concept of named entity extraction was first proposed in Sixth Message Understanding Conference in 1996. Since then, a number of techniques have been developed by many researchers for extracting diversity of entities from different languages and genres of text. Still, there is a growing interest among research community to develop more new approaches to extract diverse named entities which are helpful in various natural language applications. Here we present a survey of developments and progresses made in Named Entity Recognition and Classification research.

Contents

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cosrev.2018.06.001.

* Corresponding author.
E-mail addresses: archana.goyal@ggdsd.ac.in (A. Goyal), vishal@pu.ac.in (V. Gupta), mkjindal@pu.ac.in (M. Kumar).

https://doi.org/10.1016/j.cosrev.2018.06.001
1574-0137/

1. Introduction

Today in the era of the internet, an abundance of information is available in digital form in many languages. The information stored in the structured or unstructured form needs to be processed and extracted in various natural language processing tasks. Extraction of meaningful information out of voluminous data is a big challenge which demands to develop new technologies to handle such a big data. Many areas of information extraction and natural language processing require certain pre-processing tools to analyze the lexical, morphological, phonetic, syntactic and semantic structure of the text. Named Entity Recognition is one of the text pre-processing tools which plays a vital role in different natural language applications such as Automatic Text Summarization [1], Machine Translation [2], Information Retrieval [3], Question Answering [4], etc.

1.1. What is Named Entity?

The term "Named Entity" was first considered important for information extraction task by the MUC-6 [5]. A named entity is a word form that recognizes the elements having similar properties from a collection of elements. It is called as a rigid designator or an atomic element or member of the semantic class which may vary depending upon the domain of interest. For example — in Biomedicine domain, entities of interest are gene and gene products; in general domain, person, location, organization, number, date, time, etc. are important entities; in the homeopathic domain, drug names and disease names are recognized as entities.

1.2. Named Entity Recognition and Classification (NERC)

Named Entity Recognition and Classification, an important subtask of Information Extraction [6], points to identify and classify members of rigid designators from data suited to different types of named entities such as organizations, persons, locations, etc. [7]. The concept of named entity came into existence with the introduction of MUC-6 [5]. To achieve the main goal of the conference, named entities played an important role by extracting ENAMEX (person, location, organization) and NUMEX (time, currency and percentage expressions) entities out of the structured information related to company activities as well as the unstructured text of military messages. After that various scientific events such as Information Retrieval and Extraction (IREX) Program, 2000 [8], Conference on Natural Language Learning 2002 (CONLL 2002) [9], Conference on Natural Language Learning 2003 (CONLL 2003) [10], Automatic Content Extraction (ACE) Program [11], HAREM [12], etc. gave major contribution in emergence of NER. Since then, Named Entity Recognition has become a fascinating field to be studied.

Till now, different entities have been recognized in different languages and in different domains using different approaches. Earlier systems were based on handcrafted rule-based algorithms which provided better results for restricted domains only but modern systems most often rely on machine learning based algorithms which overcome the drawbacks of rule-based systems. A number of factors are there which can make a difference to the performance of NERC like textual genres, types of entities, language, etc. NERC system built for one domain is quite challenging to port into another domain. To the extent of our knowledge, Named Entity Recognition techniques in different domains have not been found extensively which encouraged us to throw light on this field. So we present here a survey of developments and progresses made in NER research.

1.3. Motivation for conducting the survey

Today in the age of the internet, a vast amount of information is available online and is increasing every moment. The information stored can be in structured or unstructured form and needs to be extracted in a well-processed form suitable for the application using it. Robust information extraction techniques are needed to be explored to make the proper utilization of stored data on the web. Named Entity Recognition and Classification tool is the key component of Information Extraction. NERC recognizes and classifies the entity mentions of interest, useful in many natural language applications. The motivation of conducting this survey is to highlight the present status of NERC techniques developed

by research community yet and to identify numerous issues and challenges as well as factors affecting the NERC performance which are to be considered carefully while designing these systems.

The remaining article is structured as follows. Factors affecting the performance of NERC task is covered in Section 2. Section 3 highlights the issues and challenges that are to be handled while designing the NERC system. Applications of NERC task are discussed in Section 4. Section 5 outlines different types of NERC techniques. Section 6 throws light on various NERC systems proposed by different researchers and analysis of results is also highlighted in this section. An introduction to base classifiers used for NERC task is given in Section 7. Section 8 presents evaluation metrics used to measure results in NERC task. Section 9 indicates some future directions which will help the research community to enhance the research in this field. Section 10 concludes the article.

2. Factors affecting the performance of NERC task

There are certain factors affecting the performance of NERC task like the language in which NERC is to be performed; the textual genres or domains to be considered, the number of entities to be extracted, etc. Some languages or domains are resource-poor, thus making the NERC task challenging. The improvement in tag set increases the complexity of the system. More rules or features are to be identified for more number of entities to be recognized.

2.1. Language factor

Most of the earlier research in NERC focused on English as well as European languages. Capitalization clue is the major indication for identifying named entities in these languages. Later on, Asian and some other languages are also considered. English and Japanese are well explored in MUC-6 [5] and previous works. German, Dutch and Spanish are discussed in CONLL conferences. ACE Program has worked on audio, image and textual data in Arabic as well as English. Portuguese is studied at HAREM [12,13]. Chinese NERC task has enormous literature [11,14–20]. Some other languages have got attention in NERC task: Dutch [21], French [22,23], Turkish [24,25], Myanmar [26], Vietnamese [27–29], Arabic [30–32], Mongolian [33], Polish [34], Russian [35], etc. Indian Languages are well explored by different researchers. After the introduction of the IJCNLP-08 workshop at IIIT Hyderabad, research in South and South East Asian Languages (SSEAL) [36] is going at pace. Hindi is examined by [37], Gujarati by [38], Bengali by [39], Manipuri by [40], etc.

2.2. Textual genres or domain factor

Textual genres have a great impact on Named Entity Recognition task. In the last few years, various domains are explored for the NERC task. A substantial amount of work has been done in Biomedicine domain. Due to the complex structure of biomedical entities, Biomedicine NER showed less performance than general NERC systems [41]. JNLPBA-2004 [42] dataset has been well used for extracting various entities relating to gene and gene products. A good amount of research in NERC belongs to the general domain [43]. Ek et al. (2011) [44] proposed their work for extracting entities out of the unstructured text of short text messages (SMSes). Some other domains such as homeopathic domain [45], microtext [46], tweets [47], traditional Chinese medicine [48], scientific publications [49], electronic health records [50], Wikipedia articles [51], offline unstructured handwritten document images [52], clinical notes [53], web text [54], etc. A handful of work on other domains is also available. For example biochemistry [55], speech recognition [56], geographical information systems [57], etc.

Table 1
Different segment representation (SR) techniques and their applications [63].

SR techniques (tags)	Application
IO (Inside, Outside)	Shallow parsing
IOB1 [64] (Inside, Outside, Begin)	Noun phrase chunking
IOB2 [65] (Inside, Outside, Begin)	Part of speech tagging
IOE1 [66] (Inside, Outside, End)	Noun phrase chunking
IOE2 [66] (Inside, Outside, End)	Noun phrase chunking
IOBE [66] (Inside, Outside, Begin, End)	Noun phrase chunking
IOBES [67] (Inside, Outside, Begin, End, Single)	NERC

2.3. Entity type factor

Selection of the tag set is a big challenge in NERC task. Majority of the earlier research has contributed to the extraction of limited entity types. MUC-6 [5] involved recognition of entities like people, organization, place names, temporal expressions and numerical expressions. MUC-7 [58] considered only three classes: person, location and organization. CoNLL-03 [10] included one more entity type i.e. miscellaneous. ACE [11] worked out for four entity types namely geo-political entities, weapons, vehicles and facilities. The introduction of GENIA corpus [42] made many studies easier dedicated to entity types namely, protein, RNA, DNA, cell_type, cell_line [43,59]. Studies are also conducted to recognize drug names and disease names [46], diagnosis and treatment entities [51], chemical names [60], symptom names [49], etc. Many studies are concerned with the extraction of fine-grained entities [61,62]. Working with fine-grained entities are quite cumbersome than coarse-grained entities due to data sparsity.

3. Issues and challenges in NERC task

Named Entity Recognition and Classification task has many issues which make it quite challenging such as nested entities, ambiguity in the text, availability of resources, etc. These challenges should be carefully handled to make the Named Entity Recognition Systems robust. Some of these challenges are discussed as follows:

Nested entities: Entities inside other named entities are called nested entities. Recognition of nested entities is a major challenge, affecting the performance of NERC task. Segment labeling is the solution to this issue as found by many researchers. Different types of segment representation techniques [63] have been proposed by different researchers as shown in Table 1.

Ambiguity in text: Text is ambiguous if it appears as a named entity at one place and common noun at another place or if it is used to refer to different entity types. For example, Jordan refers to the location as well as person name also. To resolve this issue, named entity disambiguation task provides inference ability to a system so that it can determine that a chunk is actually a named entity or not. Besides it, contextual information can give a major clue to determine the type of entity.

Annotation of training data: The labeled data or annotated data is the essential input to supervised learning methods. These methods build the model by learning the training examples and then this model is used to detect similar examples of the same kind in testing data. Annotating the training data is a tedious and time taking task and requires the engagement of domain experts to efficiently annotate it. This issue can be resolved using semi-supervised techniques requiring a small amount of annotated data as seed examples for further classification or using unsupervised learning ones.

Lack of resources: A large annotated dataset (corpus), as well as gazetteers, are the excellent resources which can be relied upon while implementing and testing the performance of NERC systems. Besides it, some other resources such as POS Tagger, Morphological

Analyzer, Chunker, etc. play an important role in recognizing entities. Some languages such as Arabic, Mongolian, Indonesian, Indian languages like Hindi, Punjabi, Bengali, Urdu, etc. are resource-poor languages, thus making the NERC task more challenging.

Above all, there are some other issues which make the NERC task hard like capitalization issue, agglutinative nature of the text, spelling variations, non-local dependencies, etc. These issues require proper consideration to make the NERC system efficient.

4. Applications of Named Entity Recognition

Named Entity Recognition acts as the base for many crucial areas to manage abundant of digital information stored in the structured or unstructured form. It acts as a pre-processing tool to solve many complex NLP applications. Some of them are highlighted as below:

4.1. Information Extraction

The accuracy of Information Extraction (IE) System depends upon proper names i.e. named entities as they carry important information about the text itself. So NERC is considered important step to Information Extraction. Named Entity Detection plays a crucial role in Protein–Protein Interaction (PPI) Information Extraction task [68] by exploiting dictionary lookup strategy and Conditional Random Field (CRF) based machine learning. NERC is a pre-requisite for an event extraction as well as a relation extraction task as relation extraction [69,70] pipelines start with recognizing named entities to identify the relations expressed between entities and concepts and then ends with the classification of the relation type.

4.2. Question-Answering

Question Answering (QA) is concerned with building systems that generate answers to questions asked by human beings in natural language. These systems are classified according to the type of questions asked by users. One important type of questions is factoid type questions which generally start with wh-word (What, When, Which, Where, Who) and require answers in a phrase or small sentence [71]. So Named Entity Recognition System works as an important element in a Question Answering System [72–74]. The reason behind the employment of NERC as a component in a QA System is to find the answers of many fact-based questions and these answers are entities that can be detected by the NERC system only. Therefore, incorporating the NERC in a QA system makes the task of finding answers to some of the questions considerably easy.

4.3. Machine Translation

Automatic Machine Translation is the procedure of converting text or speech from source language to target language by the computer automatically without human intervention. Correct named entities' identification is a challenging task in Machine Translation [2,75] because proper names need to be tackled in a different way than other type of words. Named entities require different approaches for translation due to specific translation rules that apply to them. Failure in correct identification of named entities not only effects lexical and global syntactic structure of translation but also immediate and local context in the text. The quality of Machine Translation System can be improved with the use of automatic Named Entity Recognition System. It increases the BLEU [76] score of Machine Translation System.

4.4. Automatic Text Summarization

Automatic Text Summarization is the problem of building a system that extracts short, and an accurate summary with all the important points of the original document. Text summarization techniques include extraction of text segments based on statistical or heuristic methods. For extracting the text, topic identification is considered as a prime task. Named entities are considered as an important indication of the topic of the text. They are the useful key expressions for text summarization [1]. Therefore, integrating Named Entity Recognition significantly improves the performance of resulting summaries [1,77]. It is also found by [78] that appropriate weight should be given to named entities while summarizing text segments in order to avoid repetition found in resulting summary.

4.5. Text Clustering

Text Clustering involves grouping a set of texts in such a way that the texts in one group (cluster) contain same properties than the texts in other groups or clusters. It is aimed at classifying and grouping up the data of common attributes together. The process of Text clustering is mainly used for knowledge discovery and data mining. It includes Keyword Extraction and Named Entity Recognition because keywords present in the text assist in making text clusters, as well as word forms, recognized as persons, locations, and organizations, are also included in grouping the data points. Combination of named entities and keywords improves the clustering quality [79]. Named Entity Recognition has shown a remarkable improvement in Suffix Tree Clustering (STC) [80] which clusters the news searching results returned by the search engine.

4.6. Information Retrieval

Information Retrieval is the process of fetching relevant information out of a collection of information resources in response to the queries made by users [81]. The queries consist of a collection of strings which includes keywords or named entities and these keywords or entities are matched with the information content stored in large databases to make the information access quick [82]. Information Retrieval Systems do not analyze the information to be searched. It mainly focuses on applying the robust technique to retrieve the accurate information. Robust Named Entity Recognition simplifies the task of information retrieval problem. Content-based Information Retrieval System has been developed by [83] that combines the methods category tagging done by Named Entity Recognition and content tagging done by Semantic Role Labeling.

4.7. Knowledgebase or ontology population

A knowledge base is mentioned as a knowledge unit which stores information used by the computer system for knowledge management and sharing such as taxonomies, ontologies, Thesaurus, etc. An ontology is a semantic and formal encoding of concepts for a domain of interest [84]. It consists of the concepts and relations that make up a conceptualization of a domain, while a knowledge base contains the instances of the classes and the relations in the ontology [85]. Building knowledge bases or ontologies involves extraction of entities and concepts from data and learning the semantic and conceptual relationship between them. Therefore, it needs support from Relation Extraction and Named Entity Recognition. KnowItAll [86] tool is one of the examples of such kind of system.

4.8. Opinion mining

Opinion mining, also known as sentiment analysis, involves building a system to collect and categorize opinions about an entity. This system studies people's attitudes, opinions, emotions and

appraisals towards entities. These entities can refer to products, individuals, organizations, services, events, etc. [87]. People freely express their opinions on social web on a variety of topics or products and many people make their final decisions by looking into these opinions. The role of user's comments is of particular importance when there is a little differentiation between the product offers. Therefore, Named Entity Recognition and Classification plays a considerable role in opinion mining. It is important in determining roles. The system OPINE has been developed for the extraction of attributes of products and the analysis of the related opinions [88].

4.9. Semantic search

Semantic search aims to search for the information and the knowledge on the web by better understanding users' intentions and directly answers the query than traditional search. It brings the ability to extract relevant answers and delivers more personalized results. Extraction of named entities and concepts from documents makes semantic search more powerful and robust [89]. Named Entity Recognition task has been widely used in web search queries as it assists in better understanding their semantics by exploiting the representation of contextual cues around the named entities. Detecting and analyzing the named entities consisting of a search query makes search engines possible for meeting users' search intent.

4.10. Other applications

Named Entity Recognition has many other applications in the medical field. For example interaction between drug–drug [90,91], detection of adverse drug effect [92,93], classification of diagnosis [94], interactions between gene–gene and protein–protein [95], identification of heart disease risk factors [96], extraction of bio medical entities, etc. Biomedical Named Entity Recognition aims at searching and classifying biomedical entities into classes like genes, proteins, diseases, etc. [97]. Named Entity Recognition is very challenging in biomedical domain due to the presence of semantically related entities in the data, variations in names of same concepts, common acronyms and abbreviation, etc.

5. Techniques used in NERC

Techniques for NERC is broadly classified into 3 main streams: rule-based approaches, learning-based approaches and hybrid approaches.

5.1. Rule-based approaches

Earlier systems are most often based on hand-crafted rules as noted by [7]. These systems include usage of information lists such as gazetteers as well as rules based on syntactic-lexical patterns to identify and classify named entities. Rule-based NERC systems are considered as highly efficient because they exploit the properties of language-related knowledge [98]. They employ domain specific features to obtain the sufficient accuracy. However, some limitations of these systems are that they are quite expensive, domain-specific and non-portable. Furthermore, these systems require human expertise with regard to knowledge of the domain and language along with programming skills [99] for its development. Besides it, rule-based systems cannot be transferred across domains. Therefore, such kind of systems made for one domain cannot be ported into other domains which shifts the interests of researchers towards machine learning based approaches.

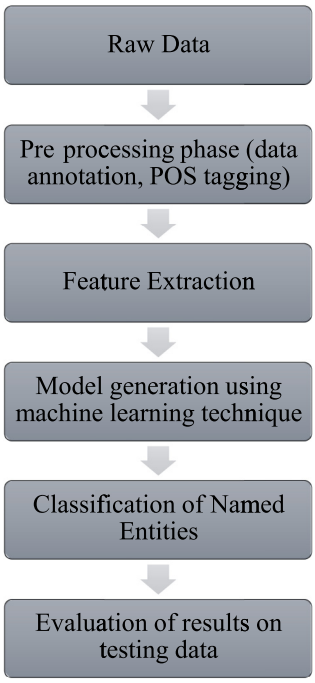

Fig. 1. Block diagram of supervised NERC system.

5.2. Learning-based approaches

Machine learning refers to the science of automatically learning complex patterns or sequence tagging algorithms which further makes efficient decisions about the data. Learning-based approaches can be divided into following three categories:

5.2.1. Supervised learning

Supervised learning based approaches are based on the idea of providing labeled training data involving positive and negative examples; constructing the adaptive features associative with examples; selecting appropriate learning algorithm that distinguishes positive from negative examples by consuming these features and recognizing similar information from unseen data. The block diagram of supervised NERC system is given in Fig. 1.

Training data is prerequisite to supervised learning algorithms. They are labeled instances of named entities, annotated by domain experts manually, making this task time consuming and labor intensive. The labeled data are then used to train the learning model which is further used to recognize and classify named entities out of unannotated or test data.

Appropriate selection of features is a crucial task in supervised learning based NERC systems. Features are the properties and attributes of textual objects in a computational model. Features play an important role to represent a multidimensional aspect of text forms which are further used by the learning methods for generating a model. This model is able to recognize patterns that find similar data and classifies positive and negative examples. Features in NERC task are well explained by [7]. The authors have classified the feature space into three groups namely list lookup features, document and corpus features and word-based features. List look-up features are based on linguistic resources such as lexicons, dictionaries, gazetteers, etc. These features determine whether a word is a member of any of these resources or not. Document and corpus related features are designed based on both document structure and content. Word-based features include orthographical, contextual and morphological features.

The choice of the learning algorithm is as well important as that of the feature selection. Various learning techniques have been used by different researchers for Named Entity Recognition Systems. Examples of these systems are: Hidden Markov Model (HMM) based systems [48], Support Vector Machine (SVM) based systems [43], Conditional Random Field (CRF) based systems [45,48], Maximum Entropy Markov Model (MEMM) based systems [59], Logistic Expression based systems [44], etc. Some classifier ensemble techniques [33] are also available in the literature.

5.2.2. Semi-supervised learning

Semi-supervised machine learning is a special form of learning. Traditional classifiers require a considerable amount of annotated training data. Annotation of training data is an expensive, difficult and time taking task because it requires the efforts of experienced human annotators. Semi-supervised learning addresses this problem by using both labeled and unlabeled corpus to make their hypothesis. These methods use a small number of training examples called "seed" for tagging unlabeled data. The results are then used to re-train the system to generate more labeled examples. This process continues to several times to make the learning decisions refined. The most popular method is "bootstrapping" [51] used by many researchers and gained popularity.

Semi-supervised pattern based bootstrapping approach has been used for identifying named entities from English and Tamil data [100]. In this approach, a small set of tagged training data is used to extract word and context features to define a five word window context pattern for each named entity category. The identified patterns are used as seed patterns. These seed patterns are used to identify the entities as an exact match in the test set. Two parameters are used to decide the modification needed to generate new patterns. These parameters are the pattern scoring and the tuple value scoring. The pattern score determines which set of patterns are used for the next iteration. The tuple value scoring provides which set of tuple contributes to the named entity and decides the window movement that is a shift to the left or to the right and masks one tuple which generates new patterns that are used to learn new context to identify named entities.

5.2.3. Unsupervised learning

Unsupervised learning is an algorithm that uses information which is neither classified nor labeled. These methods purely use unlabeled data to make their decisions. The goal of unsupervised learning is to generate a model that considers the structural and distributional features of data to find more learning about the data.

The typical unsupervised approach is clustering and association rules-based approach. Clustering based approach uses distributional statistics to extract named entities out of unlabeled data by making use of context similarity. Association rules-based technique is concerned with finding associations amongst items within large databases. To deal with lack of annotated text across domains and languages, unsupervised techniques for NERC have been proposed [51].

5.3. Hybrid approaches

Hybrid approaches hold the advantages of both learning-based and rule-based techniques. It finds the final results by combining the results of two or more machine learning techniques or handcrafted rules. Various hybrid Named Entity Recognition Systems have been introduced by different researchers.

Biomedical entities are extracted using hybridization of machine learning technique (CRF) and post-processing algorithms [55]. Chinese Named Entity Recognition System has been developed using the combination of CRF, transformation-based learning and rules [101]. Classifier ensemble technique has been used to extract named entities from Hindi, Bengali and Telugu data [102]. Combination of lexical resources and patterns and rote-based learning has been used for Turkish Named Entity Recognition [103]. Hybridization of supervised (Linear CRF) and unsupervised approach (Cluster based approach) is used for recognizing entities from English tweets [47].

Hybrid systems are found more accurate than individual systems as observed in reviewing the literature.

6. NERC systems

Named Entity Recognition and Classification is considered as an important research area by the research community. This is proven by the notable publications in this field. This section presents a brief summary of different NERC systems dependent on rule-based techniques, learning-based techniques and hybrid techniques.

6.1. Rule-based NERC systems

Shaalan and Raza (2009) [104] have introduced an Arabic rule-based Named Entity Recognition System (NERA) which involves the usage of a local grammar, a dictionary of names and a filtering technique. The purpose of the filter is to refine the system output by rejecting the incorrect named entities. The output obtained by NERA is the F-score of 87.7% for person, 85.9% for location, 83.15% for organization, and 91.6% for date entity.

Riaz (2010) [105] has proposed a successful rule-based Urdu Named Entity Recognition System while facing a number of challenges like no capitalization, agglutinative nature, ambiguity, spell ing variations, word order, nested entities, conjunction ambiguity, lack of resources, etc. and claimed that their rule-based system is more robust than CRF based NER proposed in IJCNLP 2008 shared task. The authors have also proved that NER for the Hindi language cannot be ported into the Urdu language even if both are closely related languages. Several hand-crafted rules are applied to 2262 documents collected from Becker Riaz corpus containing short news articles for identifying the named classes like person, person of influence, organization, location, date, and number. Experiment, conducted on 600 documents, results in Precision of 91.5%, Recall of 90.7%, and F-score of 91.1% by identifying 171 correct entities out of 187 named entities. The same rule-based system is also evaluated on the dataset of 36,000 Urdu tokens used in IJCNLP 2008 task and showed the F-score of 81.6% which is far better than earlier Urdu NERC systems presented in IJCNLP workshop.

Automatic text summarization is one of the applications of Named Entity Recognition. Gupta and Lehal (2011) [106] have developed a Punjabi language NERC System which is further embedded into Automatic Text Summarization System. As Punjabi is a resource-poor language so various gazetteer lists namely, proper name list, prefix list, middle name list, last name list and suffix list, are manually prepared by the authors from the Punjabi corpus. These lists are used as linguistic resources to perform the conditional rule-based technique for extracting named entities. The results obtained by the system is Precision of 89.32%, Recall of 83.4% and F-score of 86.25%. The error rate of 13.75% is also reported by the system which is due to the inconsideration of proper nouns as general nouns as well as non-availability of some words in gazetteers.

Singh et al. (2012) [107] have built a rule-based Urdu language Named Entity Recognition System. The system extracts 13 named entities by applying various ruling patterns out of which 12 NEs are the same discussed at IJCNLP-08 workshop and one is izaafats. Various dictionaries are looked up to find different named entities out of two datasets of different domains. Dataset 1 includes news information related to politics, short stories and articles and

dataset 2 consists of news related to business and science. The results are evaluated using intrinsic measures including Precision, Recall and F-score. Dataset 1 scores Precision of 86.17%, Recall of 90.40% and F-score of 88.1% while dataset 2 attains Precision of 58.15%, Recall of 62.05%, and F-score of 60.09%. Accuracy is mainly affected by less occurrence of four tags i.e. terms, brand name, izaafats and title object in the data. The combined accuracy of the whole system including 13 tags is 74.09%.

Gődény (2012) [108] has introduced a rule-based product name recognition and disambiguation system which identified product names from the user-generated online content. The records of products from two domains are considered for evaluation: the consumer electronics and automotive domains. Several rules and filtering methods such as chunking, type pattern of token sequences, custom sanity checks, etc. are applied to detect and disambiguate the product names. The system has shown the F-score value of 22.13% for publicly available test data.

Zaghouani (2012) [109] has developed a Named Entity Recognition System (RENAR) for Arabic language using rule-based techniques. Several handwritten local patterns, linguistic dictionaries, language independent as well as language-specific rules are applied to identify four named entities namely, person, organization, location and miscellaneous, out of the dataset named ArabiCorpus [110]. The corpus consists of 68,943,447 Arabic words from various resources such as newspapers, the Quran and Arabic literature. The results of the system are compared with Benajiba's ANerCorp [111] and Lingpipe [112]. The system has shown the best Precision value of 71.18% for person entity and best F-score value of 87.63% for location entity.

Alfred et al. (2014) [113] have developed a rule-based NERC System for Malay articles which are retrieved from two local Malay websites [114,115]. The articles are consisting of four different categories i.e. 155 articles from general category, 143 articles from the economic category, 35 articles in a politic category and 30 articles from sports category. The system is implemented by applying rule-based POS tagging process for the Malay language as well as contextual feature rules. When the token is detected as a proper noun then a specific rule is applied to determine whether the token is named entity or not. Besides that, several manually constructed dictionaries are used to detect three named classes such as person, location and organization. Evaluation using standard performance metrics has shown Recall of 94.44%, Precision of 85% and F-score of 89.47%.

Rahem and Omar (2015) [116] have proposed a rule-based NERC System for drug-related crime news documents. Several heuristics and grammatical rules are used to detect five different named classes such as types of drugs, price of drugs, amount of drugs, drug hiding methods, the nationality of the suspect. The dataset including 200 online crime news documents is collected from national news agency of Malaysia (BERNAMA) and 30 documents are used for testing. Two experiments are conducted on this system. The first experiment is based on heuristics only and second experiment is based on the combination of heuristics and grammatical rules. The second experiment has outperformed the first experiment by resulting Precision of 86%, Recall of 87% and F-score of 87%.

Quimbaya et al. (2016) [50] have proposed a dictionary lookup approach for identifying named entities from electronic health records. This task is quite challenging due to various issues like inclusion of images, test results, narrative text, a variety of notes, diversity of language, etc. in the text. The authors experiment dictionary lookup approach for exact matching, fuzzy matching and stemmed matching for extracting relevant named entities (diagnosis, treatment) from the i2b2 NLP Dataset #7b: 'Heart Disease Risk Factors Challenge Dataset'. Evaluation of various combinations shows that combination of exact, fuzzy, stemmed (efs) method has improved the performance indicating rising Recall value but a little impact on Precision.

Table 2 presents the comparison among various rule-based systems in terms of domain and language used, entities extracted, the dataset used and results obtained.

6.2. Learning based NERC systems

Learning based systems include supervised, semi-supervised and unsupervised learning systems. Comparison among different learning based NERC systems is given in Table 3.

6.2.1. Supervised NERC systems

Benajiba et al. (2009) [117] have performed Named Entity Recognition task on Arabic dataset by exploiting different contextual, morphological and lexical features. Various machine learning models such as SVM, CRF and ME models are developed and analyzed on 9 different genres of datasets separately and collectively. The results are evaluated on Broadcast News data and found that CRF based approach along with 15 top-ranked features has outperformed showing F-score of 83.34%.

Selection of appropriate features plays a key role in the performance of machine learning algorithm. Saha et al. (2009) [59] found that some features in the biomedical domain have great dimensionality which degrades the performance of NER task. These features can be surrounding words feature and affixes information. Here, the authors present an approach using feature selection and word clustering algorithm which reduced the dimensions of these features. The task is performed using MaxEnt classifier on JNLPBA-2004 data [42] consisting of five biomedical named entities. Data is annotated using BIO labeling scheme before training and testing. This approach outperforms the NERC task by improving F-score of 1.42% than the baseline system and other biomedical NERC systems.

Saha et al. (2010) [43] have developed a NERC system which is based on SVM classifier and works for Hindi and Biomedical data. The authors have proposed a novel composite kernel function which combines the properties of hierarchical word clustering kernel function and class association function. Brown clustering algorithm [136] is used for Hierarchical word clustering kernel. The basic idea behind the composite kernel function is to detect the semantic and contextual similarities among the words. Three named entities named as person, organization and location from Hindi domain and five naming words namely, protein, DNA, RNA, cell line, cell type from the biomedical domain are extracted. Hindi dataset is obtained from Hindi newspaper "Dainik Jagran" and the biomedical dataset is obtained from GENIA corpus version 3.02 [42]. The results obtained by Hindi and biomedical NERC show that SVM composite kernel outperforms with 83.56% F-score for Hindi and 67.89% F-score for biomedical domain as compared to MaxEnt based system [59].

Most of the studies in NER task are related to extracting named entities out of the structured text. Here Ek et al. (2011) [44] presents the study of extracting named entities from the unstructured text of short text messages (SMSes) running on the Android phone and written in Sweden. Five named entities i.e. date, time, location, person names, telephone numbers are considered for extraction. The authors have found it very challenging task due to certain constraints like limited computing power and memory. The dataset used for this task consists of 4500 text messages containing 60,000 tokens which are collected from incoming and outgoing messages of 11 participants. Tagging of the dataset is done using IOB2 format [9]. Regular expressions are first applied to the text for detecting NEs which have proved effective in the identification of numerical entities like date, time, and telephone numbers. Logistic Expression classifier is also used for identifying letter based entities

Table 2
Rule-based NERC systems.

Authors & year	Language/domain	NEs found	Technique used	Dataset used	Evaluation results
Shaalan and Raza, 2009 [104]	Arabic	Person, date, company, location, time, phone no., measurement, price, filename , ISBN	Use of dictionaries and local grammar	Annotated corpora is collected from the ACE and the ATB	F-score: person: 87.7% date: 91.6% company: 83.15% location: 85.9% time: 95.4% phone no: 91.3% measurement: 97.2% price: 98.6% filename: 96.4% ISBN: 95.3%
Riaz, 2010 [105]	Urdu	Person, person of influence, location, organization, date, number	Use of handcrafted rules	2262 documents from Becker Riaz corpus containing short news articles	Precision: 91.5% Recall: 90.7% F-score: 91.1%
Gupta and Lehal, 2011 [106]	Punjabi	Proper nouns	Use of manually created gazetteers and condition based technique	50 Punjabi news documents	Precision: 89.32% Recall: 83.4% F-score: 86.25%
Singh et al. 2012 [107]	Urdu	Person name, organization, designation, location, title objects, number, measure, terms, date/time, title person, brand, abbreviation, izaafats	Use of gazetteer lookup and hand-crafted rules	Two datasets consisting of 12,032 tokens and 150,243 tokens related to news articles are collected from BBC website	Combined F-score: 74.09%
Gődény, 2012 [108]	The consumer electronics and automotive domains	Product names	Use of rules and filtering methods	User-generated web-based corpus containing blogs, forums, product review sites etc.	F-score: 22.13%
Zaghouani, 2012 [109]	Arabic	Person, location, organization, miscellaneous	Use of handwritten local pattern, language independent and language specific rules	68,943,447 Arabic words retrieved from ArabiCorpus	Best Precision value of 71.18% for person entity and best F-score value of 87.63% for location entity.
Alfred et al. 2014 [113]	Malay	Person, location, organization	Use of POS tagging, gazetteers lookup and contextual feature rules	Malay articles retrieved from two local Malay websites	Precision: 85% Recall: 94.44% F-score: 89.47%
Rahem and Omar, 2015 [116]	Drug-related crime news documents	Types of drugs, price of drugs, amount of drugs, drug hiding methods, nationality of the suspect	Use of heuristics and grammatical rules	200 online crime news documents from national news agency of Malaysia (BERNAMA)	Precision: 86% Recall: 87% F-score: 87%
Quimbaya et al. 2016 [50]	Electronic health records	Diagnosis, treatment	Use of dictionary lookup for exact, fuzzy and stemming matching	The i2b2 NLP Dataset #7b: 'Heart Disease Risk Factors Challenge Dataset'	Precision: 63.0% Recall: 57.3% F-score: 60.0%

like location, person names. For evaluation, Recall, Precision and F-score are calculated using 10-fold cross validation which shows 86.44% F-score for strict matches and 88.85% F-score for partial matches.

Lee et al. (2011) [118] have proposed a Modified Pegasos Algorithm for structural SVM, a stochastic gradient descent method, which is used for Named Entity Recognition task in TV domain and Food domain and compared its performance with baseline CRFs and 1-slack S-SVMs [137]. The NERC task is performed in Korean dataset for 15 NE classes: person, location, organization, artifacts, study fields, theory, civilization, date, time, quantity, event, animal, plant, material and term which includes 105,265 and 3719 sentences for training and testing in TV domain and 27,628 and 3719 sentences for training and testing in Food domain. Comparing the performance and training time, it is found out that pegasos algorithm has significantly outperformed CRFs with L-BFGS algorithm. There is a less difference between the outcome of Pegasos Algorithm and 1-slack S-SVMS but on account of training time, pegasos has outperformed 1-slack S-SVM.

Saha et al. (2012) [119] have focused on feature reduction approaches because high dimensionality of features lowers the performance of NERC system due to over-fitting of data with the small amount of training data. The authors have applied these approaches to data collected from Hindi, Bengali and biomedical domain. Features dimensionality can be reduced using feature Extraction [138] as well as feature selection techniques. NE class association metric [139] is used to reduce the word-level features. For feature selection, filter based selection method along with decision tree (C4.5) and sequential forward selection algorithm is used. Graph clustering [140] and hierarchical clustering algorithms are also used for extracting the features. Hindi data is obtained from Hindi Newspaper "Dainik Jagran" which is further annotated manually. For Bengali NERC, the corpus is collected from IJCNLP 2008 shared task [141]. The experiments are also conducted on JNLPBA 2004 data which is taken from the GENIA corpus version 3.02 [42]. Training and testing are done using two machine learning algorithms i.e. Conditional Random Field and Maximum Entropy along with word-level features such as tag information,

Table 3
Machine learning based NERC systems.

Authors & year	Language/domain	NEs found	Technique used	Dataset used	Evaluation results
Benajiba et al. 2009 [117]	Arabic	Arabic proper names	Supervised approach: SVM, ME and CRF	Different sizes of the datasetare collected from NEL corpus, ACE 2003, 2004, 2005 corpus in four different genres: broadcast news, newswire, Arabic treebank, weblogs	CRF approach outperformed with 83.34% F-score in broadcast news
Saha et al. 2009 [59]	Biomedical domain	Protein, RNA, DNA, cell_line, cell_type	Supervised approach : MaxEnt classifier	Data is collected from GENIA corpus	F-score: 67.41%
Saha et al. 2010 [43]	Hindi, Biomedical domain	Protein, RNA, DNA, cell_line, cell_type	Supervised approach: MaxEnt, CRF, SVM	Hindi data is collected from Hindi newspaper "Dainik Jagran" and biomedical data is collected from GENIA corpus	SVM based kernel outperformed. F-score for Hindi:83.56%, F-score for Biomedical domain: 67.89%
Ek et al. 2011 [44]	The unstructured text of SMSes in Sweden	Date, time, location, person names, telephone numbers	Supervised approach: Logistic Expression Classifier	60,000 tokens which are collected from incoming and outgoing messages of 11 participants	10-fold cross-validation showed F-score of 86.44% (strict matches) and 88.85% (partial matches)
Lee et al. 2011 [118]	TV domain and Food domain in Korean	Person, location, organization, artifacts, study fields, theory, civilization, date, time, quantity, event, animal, plant, material and term	Supervised approach: Modified Pegasos Algorithm	TV domain: Training and testing (105,265 and 3719 sentences) Food domain: Training and testing (27,628 and 3719 sentences)	TV domain: F-score: 89.46% Food domain: F-score: 85.43%
Saha et al., 2012 [119]	Hindi, Bengali, Biomedical domain	Person, location, organization	Supervised approach: CRF and MaxEnt classifier	200k words are collected from Hindi Newspaper "Dainik Jagran". 100k Bengali words are collected from the IJCNLP-08 website. About 500k words are collected from GENIA corpus	F-score: Hindi: 85.31% (CRF) 80.2% (MaxEnt) Bengali: 70.75% (CRF) 67.54% (MaxEnt) Biomedical data: 71.56% (CRF) 67.24% (MaxEnt)
Jung, 2012 [46]	Online streaming microtext on Twitter	Person, location, organization, digital ids	Supervised approach: Maximum Entropy	Dataset is collected from social website i.e. Twitter	F-score: 90.3%
Freire et al. 2012 [120]	Poorly structured data in bibliographic contents	Person, location, organization	Supervised approach: CRF	Bibliographic contents of cultural interest are collected from Europeana	Maximum Precision of 0.91 at 0.55 Recall and a maximum Recall of 0.82 at 0.77 Precision
Nothman et al. 2013 [121]	English, Spanish, German, Dutch, Russian	Person, location, organization	Supervised approach: Logistic regression classifier	4800 English, 870 German, and 1500 other language articles are collected from Wikipedia	F-score for 5 languages: English — 88.7 Spanish — 87.6 German — 88.4 Dutch — 87.4 Russian — 87.4
Wang et al. 2014 [48]	Traditional Chinese Medicine domain	Symptom name	Supervised approach: HMM, MEMM, CRF	11,613 clinical records are collected from traditional Chinese medicine from April, 2006 to June, 2008	highest F-score: 95.12% using CRF

(*continued on next page*)

n-gram, suffix, POS information, prefix, etc. The F-score values for Hindi NER, Bengali NER and Biomedical NER are 85.31%, 70.75% and 71.56% respectively with the use of CRF and 80.2%, 67.54% and 67.24% respectively with the use of MaxEnt. These systems have outperformed due to the usage of combined version of both clustering and selection based reduction methods.

Jung (2012) [46] has performed named entity extraction task on the microtext which streams online on social networking sites. Finding the contextual information of the microtext is quite difficult due to its small size. The authors have used the concept of microtext merging or clustering to resolve this problem. Microtext clustering is the group of microtexts which are contextually associated. To find the contextual relationship between the entities, three different heuristics are applied by the authors such as semantic association, social association and temporal association. Semantic association determines the similarity between the two sets of word features. Social association looks for the digital IDs of corresponding users. Temporal association determines the closeness between two microtexts with reference to time. In this study, the authors have experimented the NERC task on microtexts running on Twitter using Maximum Entropy (ME) approach [142]. The evaluation results have shown the higher accuracy of 90.3% with microtext clusters method as compared to the single microtext method.

Table 3 (*continued*)

Authors & year	Language/domain	NEs found	Technique used	Dataset used	Evaluation results
Fersini et al. 2014 [122]	4 different genres: US50 (postal addresses), CoNLL-2003 (general domain), Cora (citations on research papers), Advertisements (announcements for apartment rentals)	Different entities as per different genres such as street, city, state, zip code, person, location, organization, author, publisher, data, journal, address, photo etc.	Supervised approach: CRF-Soft (with all constraints)	Datasets including US50 (a set of postal addresses), CoNLL-2003 (consists of Reuters news stories), Cora (citations on research papers) and Advertisements (announcements for apartment rentals)	**US50:** Macro-average F-score: 88.75 Micro-average F-score: 83.90 **CoNLL-2003:** Macro-average F-score: 76.92 Micro-average F-score: 78.78 **Cora:** Macro-average F-score: 87.61 Micro-average F-score: 92.95 **Advertisement:** Macro-average F-score: 68.98 Micro-average F-score: 79.12
Bam and Shahi 2014 [123]	Nepali	Person, location, organization, miscellaneous	Supervised approach : SVM	Three different sizes of training data (5000 tokens, 15,000 tokens and 29,298 tokens)	With training size 29,298 tokens, maximum Precision: 86.85%, Recall: 98.53%, F-score:9 2.31%
Banerjee et al., 2014 [39]	Bengali	Person, organization, location, brand, term, measure, title person, designation, title object, abbreviation, number, time	Supervised approach: MIRA	Bengali dataset is collected from IJCNLP-08 web site	Precision: 89.26% Recall: 82.99% F-score: 86.01%
Keretna et al. 2015 [63]	Biomedical domain	treatment, test, and problem	Supervised approach: Segment representation with Naive Bayes, k-NN, CRF, Random Tree, ME, C4.5, Random Forest Ada-Boost	I2B2 2010 medical challenge dataset	Average F-score: 82.3%
Chen et al. (2015) [124]	Clinical text	Treatment, problem, and test	Supervised approach: CRF	Annotated corpus on 2010 i2b2/VA NLP challenge	F-score: 80%
Korkontzelos et al. 2015 [125]	Drug domain	Drug names	Supervised approach: Maximum entropy and multinomial. Logistic regression classifiers	Dataset is collected from the pharmacokinetic corpus, UKPMC corpus , a dictionary from DrugBank	F-score: 95%
Yan and Zhu, 2015 [126]	Scientific publications domain	Entities from publications	Supervised approach: Vocabulary based and CRF based methods	Dataset of 7262 research papers is collected from five leading computer science journals	Best performance with CRF + keyword-based dictionary Precision: 57% Recall: 68% Acc: 90% AUC: 80% AUP: 44%
Konkol and Konopík, 2015 [127]	English, Spanish, Dutch, Czech	Time, person, geography, address, institution, media, and other	Supervised approach: MaxEnt and CRF	English, Spanish and Dutch datasets are collected from CoNLL-2002 and CoNLL-2003 shared task, and Czech dataset is collected from format version of Czech named entity corpus	F-score: English: 83.24 Spanish: 81.39 Dutch: 75.97 Czech: 74.08
Kaur and Josan, 2015 [128]	Punjabi	Person, organization, location, facility, relationship, event, time, designation, date, number, title-person, abbreviation, measure, and artifact	Supervised approach: CRF	Raw data is collected from online Punjabi newspapers and annotated manually	Precision: 90.99% Recall: 84.19% F-score: 87.46%

(*continued on next page*)

Table 3 (*continued*)

Authors & year	Language/domain	NEs found	Technique used	Dataset used	Evaluation results
Bhasuran et al. 2016 [129]	Biomedical domain	Disease name	Supervised approach: CRF	Dataset is collected from NCBI disease corpus and BioCreative V CDR Corpus	F-score: NCBI corpus — 94.66% Biocreative corpus — 84.10%,
Wibawa and Purwarianti, 2016 [130]	Indonesian news articles	Person, god, organization, location, facility, product, event, natural-object, disease, color, timex, periodx, numex, countx, measurement	Supervised approach: Naïve Bayes, SVM, Simple Logistic classifier	Dataset of 457 Indonesian news articles is collected from different websites	Best F-score with simple logistic classifier is 52.8%
Adak et al. 2016 [52]	Offline unstructured handwritten document images	Person, location, organization	Supervised approach: BLSTM neural network classifier	20 and 66 pages of historical text collected from GWdb and QSAdb and 1539 handwritten documents of modern text collected from IAM database	F-score: 74.59%.
Guo et al. 2009 [131]	Queries	Named entities considered under 4 topics: movie, book, game, music	Semi-supervised approach: WS-LDA (Weakly Supervised Latent Dirichlet Allocation)	930 million unique queries collected from search log of commercial web search engine namely Amazon, GameSpot and Lyrics were considered for annotation	WS-LDA proved best than baselines
Majumdar et al. 2012 [45]	Homeopathic domain	drug names and disease names	Semi-supervised approach: CRF	100 k words collected from a homeopathic discussion forum	F-score: 84.35%
Ekbal et al. 2012 [132]	Hindi and Bengali	Person, location, organization, miscellaneous	Semi-supervised approach : SVM	Hindi data is collected from NERSSEAL 2008 shared task and Bengali data is collected from Bengali news corpus	Hindi: Precision: 90.22% Recall: 89.41% F-score: 89.81% Bengali: Precision: 91.65% Recall: 91.66% F-score: 91.65%
Küçük, 2015 [133]	Turkish	Enamex, numex and timex entities	Semi-supervised approach: k-NN classifier	A dataset of about 20 Wikipedia article titles is considered for manual annotation	Precision 91.25%.
Bhagavatula et al. 2012 [51]	Hindi and Marathi	General NEs	Unsupervised approach:Bootstrapping method	2935 Hindi and 2622 Marathi Wikipedia articles	F-score: Hindi 80.42% Marathi 81.25%
Zhang and Elhadad, 2013 [134]	Biomedical domain	Problem, treatment, test and protein, DNA, RNA, cell_type, cell_line	Unsupervised approach: Use of noun phrase Chunker, IDF filter, distributional semantic similarity based classification	I2B2 Challenge and GENIA Corpus	69.5% accuracy on I2B2 corpus and 53.8% accuracy on GENIA corpus
Konkol et al. 2015 [135]	English, Spanish, Dutch, and Czech	Person, location, organization, miscellaneous	Unsupervised approach: Use of Word similarity-based algorithms	Approximately 25,000 tokens are collected from English, Spanish, Dutch and Czech CoNLL corpora	F-score for English, Spanish, Dutch and Czech 89.44%, 83.08%, 83.01%, 74.08%.

Freire et al. (2012) [120] have presented a Named Entity Recognition approach for the structured data in free text using Conditional Random Field technique for person, location, organization entities using specialized features without having lexical evidence and using semantic context pattern given by the structure of data. The evaluation is performed on the datasets obtained from Europeana [143], which consisted of bibliographic contents of cultural interest including titles, table of contents, subjects, authors, and publications, etc. The model has achieved maximum Precision of 0.91 at 0.55 Recall and a maximum Recall of 0.82 at 0.77 Precision and outperformed Stanford Named Entity Recognizer, a model developed for well-structured text. The authors have claimed that their model with this kind of accuracy can be easily adapted to any other poorly structured data.

The performance of supervised NERC system is highly dependent on a large annotated dataset which is required for training and testing to extract people names, organization, location and other entities. It is very time consuming and expensive task. To overcome this problem, Nothman et al. (2013) [121] have created a large, free and multilingual silver standard annotated corpora for NER from Wikipedia by exploiting the text and its structure in five languages namely, German, English Dutch, Russian and Spanish. In this study, the authors have classified all Wikipedia articles into NE types manually by labeling each link with the entity class of that article. For modeling and evaluating, 870 German, 4800 English and 1500 other language articles are annotated with fine-grained entities which include 3.5 million tokens to train the model in each language. Logistic regression classifier has used both textual and

document structure features to evaluate the system. The proposed method does not outperform the traditional models presented in CONLL 2002–2003 shared task but English and German Wikipedia NER boost the performance than gold models. Besides, the silver standard annotations have outperformed traditional training scheme of manual annotation of Wikipedia articles [144] by 10%–12% F-score. This approach is most suitable for the resource-poor languages where training data is not available. This approach has used various Wikipedia categories such as infoboxes, bag-of-words contents in article classification, outgoing links, incoming link texts, interlingual links for extracting training data which make the model enormous and valuable.

Wang et al. (2014) [48] have proposed a supervised Symptom Name Recognition (SNR) System in Free Text Clinical Records (FCRs) of Traditional Chinese Medicine (TCM). Recognition of symp tom name is a new and challenging task. As there is unstructured content in FCRs so the models used for well-structured data cannot be adopted to Symptom Name Recognition (SNR) in chief complaints. SNR task is considered as a sequence labeling task and domain adaptation scheme applied to the labeling units enhances the performance of the task. Three supervised models including HMM, MEMM and CRF are constructed to evaluate the system. MEMM and CRF have explored character N-Gram and position feature to get the better results. The dataset collected from April 2006 to June 2008 contains 11,613 clinical records out of which 3483 chief complaints are treated as training data and rest of 8130 chief complaints are included in the test dataset. The system is evaluated on two metric groups. The first group finds Precision (P_{rec}), Recall (R_{rec}) and F-score (F_{rec}) of symptom name recognition and the second group includes Precision(P_{lab}), Recall(R_{lab}) and F-score (F_{lab}) for symptom name labeling. On Evaluating, it is found out that CRF and MEMM have outperformed HMM model and in comparison to CRF and MEMM, CRF has proved to be the best by giving 95.12% F_{rec}.

While labeling named entities, dependencies among variables are modeled through supervised techniques but these dependencies among variables should be long ranged to make the system effective. In this study, Fersini et al. (2014) [122] have proposed a novel inference method which works on Integer Linear Programming (ILP) formulation of the problem. ILP considered long-distance dependencies among variables by exploiting non-deterministic soft constraints which are obtained by learning declarative rules from data. The application of these constraints requires extra knowledge to find the complex relationships among variables which enhances the labeling output of CRF model by correcting mistakes of local predictions. Soft constraints include Adjacency, Precedence, State Change, Begin–End, Presence and Precedence. The performance is evaluated using four metrics, i.e. Precision, Recall, F-score and accuracy on four different datasets including US50 (a set of postal addresses), CoNLL-2003 (consists of Reuters news stories), Cora (citations on research papers) and Advertisements (announcements for apartment rentals) for different label distribution in each dataset. The proposed system has been compared with two state-of-the-art approaches, i.e. Viterbi algorithm (CRF-Viterbi) and SemiCRF algorithm (SemiCRF-Viterbi) out of which CRF-Soft with all constraints has outperformed both in terms of macro-average and micro-average across all considered datasets.

Bam and Shahi (2014) [123] have proposed Nepali language Named Entity Recognition System for extracting four named entities such as location, person, organization and miscellaneous. This system exploits the use of Support Vector Machine along with features like word features, digit features and list look-up features, etc. The experiments are done on three different training sizes with 5000 tokens, 15,000 tokens and 29,298 tokens. 10 different datasets with sizes of 1000 tokens to 5500 tokens are tested. Experiment 1 with training size of 5000 tokens shows the results of an average Precision of 65.93%, Recall of 80.42% and F-score of 72.44%. Average Precision of 82.66%, Recall of 97.27% and F-score of 89.36% are obtained in experiment 2 with training size of 15,000 tokens. Experiment 3 shows an average Precision of 86.85%, Recall of 98.53% and F-score of 92.31% with training size of 29,298 tokens.

Banerjee et al. (2014) [39] have developed Named Entity Recognition System for Bengali language. The authors have used Margin Infused Relaxed Algorithm (MIRA) for Bengali NERC. This algorithm works for multiclass classification problem as reported by Crammer and Singer [145]. The dataset in Bengali has been collected from South and South East Asian Languages (SSEAL) shared task for training and testing which is annotated with 12 different named entities. Various experiments are done by the authors using several language independent and dependent features. The results given by the system with language-independent features are Precision of 89.26%, F-score of 86.01% and Recall of 82.99%. The system has improved the performance with F-score of 3.12% after integrating both language independent and dependent features. It is found to be the best model after comparing with CRF, ME, HMM, SVM based systems due to the better optimization technique of MIRA.

Segment Representation is the technique to tag the named entities in the corpus in order to get the multi-word entities. SR techniques play an important role in Named Entity Extraction task as it affects the results effectively. A number of different segment representation techniques including IO, IOB, IOE, IOBE, and IOBES have been effectively used in various applications as given in Table 1. Here, Keretna et al. (2015) [63] have proposed an extension to IOBES segment representation technique by including Ambiguous (A) class to the words which appear named entities in some contexts and non-named entities in other contexts. Named Entity Recognition task for three medical named entities named treatment, test and problem have been accomplished using 8 different classifiers including CRF, Naive Bayes, k-NN, ME, C4.5, Random Tree, Random forest, Ada-Boost on I2B2 2010 medical challenge dataset. An average F-score of results of three experiments is calculated by the authors which have shown that an extended SR technique improves the performance of seven out of eight classifiers. Only k-NN classifier has increased the average error rate by 0.18% across three different experiments.

The supervised NER models require a large annotated corpus for training purpose enabling large annotation cost in terms of time and efforts. In this study, Chen et al. (2015) [124] have proposed an active learning (AL) method and evaluated existing AL methods which minimize the annotation cost and maximize the performance of supervised models for three entities named as treatment, problem and test in clinical text. Different active learning algorithms are categorized into three different groups including baseline sampling-based techniques, diversity-based techniques and uncertainty based techniques. CRF based machine learning system is modeled using optimized features and 5-fold cross-validation is conducted to evaluate the performance of the system. Learning curves are plotted to represent the cost of annotation for evaluating different active learning algorithms and passive algorithm based on random sampling and Area under the Learning Curve (ALC) score is also computed. Out of all active learning algorithms, uncertainty based algorithms have outperformed all other methods giving F-score of 80% in case of sentence based and as well word based evaluation. Uncertainty based methods are found to save 66% cost of annotations in sentences and 42% cost of annotations in words in comparison to random sampling.

Manual annotation of data is quite tedious and time-taking task for NERC system so Korkontzelos et al. (2015) [125] have presented an approach that improves the performance of drug name recognition without using manual annotation or with limited annotation. The authors have performed Drug Named Entity Recognition by

using a gold standard corpus i.e. the pharmacokinetic corpus, a silver annotated dataset i.e. a part of UKPMC database, a dictionary of drug names named DrugBank. Several heterogeneous models are constructed using Maximum entropy and multinomial logistic regression classifiers along with gold standard corpora, dictionary knowledge and silver annotations. Besides it, 11 regular-expression patterns based on genetic programming are also evolved which has refined the drug recognition task by improving the Recall. Aggregation of Named Entity Recognition methods based on gold standard annotation, patterns and dictionary knowledge have achieved the highest performance of 95% F-score as well as models based on silver standard annotation, dictionary and patterns have achieved similar or comparable performance to models based on the exclusively gold standard dataset.

Entity extraction from scientific publications data is a novel task performed by Yan and Zhu (2015) [126]. The study analyzes five extraction methods out of which two are vocabulary-based methods (a keyword-based and a Wikipedia-based) and three are model-based methods including CRF, CRF with a keyword-based dictionary and CRF with a Wikipedia-based dictionary, on a dataset of 7262 research papers collected from five leading computer science journals through stratified sampling. The system is evaluated using five indicators named as Recall, Precision, Accuracy, Area under the ROC Curve (ARC) and Area under the Precision-Recall Curve (AUP) which has found that the statistical model based methods outperform vocabulary based methods and out of three statistical based methods, CRF with keyword-based dictionary is showing the best performance. The major objective of the study is to find the entities from publications in computer science journals and to analyze informetric research at a more granular level.

Mining named entities from tweets is a challenging task because of short, dynamic, context-dependent and noisy nature of tweets that is why Named Entity Recognition methods have shown only 30%–50% accuracy on tweets data [146,147]. In this study, Derczynski et al. (2015) [148] have focused on the performance of various recent Named Entity Recognition methods on tweets and found out the problems caused in dropping the Recall and suggested some solutions to handle these problems. On evaluation, the authors have found various error sources specific to this genre namely capitalization issue, typographic errors, shortening of language, lack of contextual information, a mixture of languages, etc. The authors have suggested that removal of microblog noises can make the tweets NERC improved by adopting some pre-processing algorithms including language identification to handle the multilingualism, microblog trained part of speech tagging, normalization of tweets data, etc. Besides it, creating large annotated tweet corpora is allowing several machine learning algorithms to bring the various performance metrics up.

Konkol and Konopík (2015) [127] have shown the impact of segment representation methods on Named Entity Recognition task. The authors have experimented with ten different segment representation techniques namely IO, BIO-1, BIO-2, IOU, BIOU, IEO-1, BIEO,IEO-2, IEOU, and BILOU on the corpora of four different languages including English, Dutch, Czech and Spanish using two supervised learning techniques: Maximum Entropy and Conditional Random Field. Corpora for English, Spanish and Dutch are collected from CoNLL-2002 [9] and CoNLL-2003 [10] shared task for identifying four different entities — Person, Organization, Location and Miscellaneous. For Czech language, format version of Czech named entity corpus [149,150] consisting of 1,50,000 tokens are used for identifying seven classes — geography (G), time (T), address (A), person (P), institution (I), media (M) and other (O). The experiments are performed in two ways in which the first way includes the standard partitioning of data and the second way includes 10-fold cross-validation. Paired student's t-test is used for evaluation. On evaluation, it is found that the second test proves best by giving more accurate results and also indicates that IOE-1 and IOE-2 are the most promising segment representation techniques. Besides, the authors have cleared that the choice of optimal segment representation method depends on the language, approach and feature set used for the NER task.

Kaur and Josan (2015) [128] have developed a CRF based Punjabi language Named Entity Recognition System by exploiting various language non-dependent and dependent features. Language non-dependent features include infrequent word feature, context word feature, word length feature and several digit features. Raw data is obtained from Punjabi newspaper which is available on sites i.e. www.ajitjalandhar.com and www.ajitweekly.com. Further, the data is annotated manually with 14 named classes such as organization, person, relationship, location, facility, date, event, number, time, title-person, designation, measure, artifact and abbreviation. Training is done on a dataset of 1,70,000 words and testing is done on a test dataset of 30,000 words. CRF model including features of 5-word context window, infrequent word, and digit and word length has shown its excellence by giving Recall of 84.19%, Precision of 90.99%, and F-score of 87.46%.

Biomedical NER is most focused research area in Biomedicine text mining. Bhasuran et al. (2016) [129] have presented a new stacked ensemble approach combined with fuzzy matching for extracting disease named entities. Two separate CRF based sequence labeling models consisting of forward and backward labeling approach are trained and tested on two different corpora named NCBI disease corpus [151] and BioCreative V CDR Corpus [152] and results of base classifiers are stacked over by the results of second level meta-classifier. A number of domain-specific, morphological, orthographical and contextual features along with two fuzzy matching algorithms named Rabin Karp [153] and Tuned Boyer Moore [154] are used to improve the performance of the model by tagging rare disease named entities with the help of in-house disease dictionary. Some post-processing measures such as abbreviation resolution, disambiguation and parenthesis mismatching are applied to enhance the Recall and consistency of the model. Cross-validation and error analysis is also performed to find the robustness of the system which shows promising results with F-score of 94.66%, 89.12%, 84.10%, and 76.71% on training and testing set of both NCBI disease and BioCreative V CDR Corpora.

Wibawa and Purwarianti (2016) [130] have proposed a NERC System for classifying 15 named categories namely, person, god, organization, location, facility, product, event, natural-object, disease, color, timex, periodx, numex, countx, measurement, out of Indonesian newspaper articles. The authors have used several gazetteer features, sentence-level features, word-level features, contextual features and combination of these for training and testing. The pre-processed dataset of 457 Indonesian news articles has been collected from different web sources. Half of the dataset is used for training and another half is used for testing on three classifiers including Naïve Bayes, SVM and Simple Logistic. The systems are evaluated on a direct basis and incremental basis out of which Simple Logistic classifier has shown the best F-score value of 52.8% using feature combination of word-level, sentence-level and list lookup as well as direct scheme.

Adak et al. (2016) [52] have developed a method for identifying named entities from offline unstructured handwritten document images without using any linguistic resources and any explicit word/character recognition. Three handwritten offline datasets namely George Washington database (GWdb) [155], Queensland State Archives database (QSAdb) [156] and IAM database (IAMdb) [157] are considered for NERC task. GWdb and QSAdb contain 20 and 66 pages of historical text and IAMdb contain 1539 handwritten documents of modern text. The datasets are first pre-processed using binarization of images, word segmentation and

slant/skew/baseline correction of words. Then structural and positional properties of NEs are analyzed and extracted various features from the pre-processed word images. Bidirectional Long-Short Term Memory (BLSTM) neural network classifier is used for the recognition task. Some post-processing heuristics are also applied to improve the outcome of NERC system which results in an average F-score of 74.59%.

6.2.2. Semi-supervised NERC systems

Named Entity Recognition in the query is another issue solved by Guo et al. (2009) [131] by employing weakly supervised learning method named WS-LDA (Weakly Supervised Latent Dirichlet Allocation) in which few seed queries are manually annotated and trained and further used for classification. The authors found it as a difficult task because of the short size and informal nature of queries. NERQ, a useful application in web search, is performed on single named entity queries by representing it in triples (named entity (e), context of named entity (t), class of named entity (c)). The queries dataset is collected from search log of commercial web search engine namely Amazon, GameSpot and Lyrics which consists of 930 million unique queries out of more than 6 billion queries. Out of total 180 named entities under four Topics — "Movie", "Book", "Game", "Music",120 entities are considered for training and 60 entities are considered for testing. The evaluation results of WS-LDA is compared with two baselines — Determ [158] and conventional LDA. WS-LDA has been found superior in terms of accuracy and execution speed.

Majumdar et al. (2012) [45] have presented a CRF based NER system which identifies two named entities i.e. drug names and disease names from homeopathic diagnosis discussion forum. Tra ining of 100k words is done on a manually annotated dataset collected from ABC homeopathic discussion forum while 12k words are considered for testing. Feature set used in this task includes word feature, affixes, capitalization and numerical feature. Usually, capitalization feature is considered as best identification feature but in this study, it is found that capitalization feature brought down the accuracy low due to great noise in the data. Instead, the numerical feature has improved the accuracy a lot because most of the drug names contain a numeric value for e.g. Belladonna 30C, Arnica 10 m, etc. Due to small annotated data, the system provides low Recall (77.80%) and F-score (83.29%) value. So the authors have applied semi-supervised learning technique on large unannotated data which raises the Recall value to 79.80% and F-score value to 84.35%.

Ekbal et al. (2012) [132] have proposed a new machine learning technique i.e. an active annotation technique for automatic annotation of data. As manual annotation of data is very time taking and costly so usage of active learning technique has been found effective in this manner. It can extract a big amount of meaningful sentences from a pool of unlabeled documents, annotate it and add it to the training dataset in every iteration. This process runs until the outputs of two consecutive steps become same.SVM classifier is used for active annotation. Bengali and Hindi datasets are tagged with five named classes such as person, organization, location, miscellaneous and others through active annotation technique. Training and testing are done on a dataset collected from Bengali News Corpus [159] for Bengali and from IJCNLP-08 NERSSEAL shared task for Hindi. The result shows Precision of 91.65%, F-score of 91.65% and Recall of 91.66% for Bengali and Precision of 90.22%, F-score of 89.81% and Recall of 89.41% for Hindi.

Küçük (2015) [133] has presented a novel approach of an automatic compilation of language resources from Wikipedia article titles in Turkish. Rule-based NER [24] in Turkish utilizes the various lexical resources and pattern bases for the extraction task so the author has extended these resource sets in view of improving the performance. The author first randomly selects one-twentieth of downloaded Wikipedia articles (about 20 article titles) for manual annotation. These manually annotated titles are used as a training dataset for further classification of named entities out of remaining article titles using k-NN classifier. This automated procedure produces moderate sized language resource set for NER in Turkish which is providing a Precision rate of 91.25%. Several experiments using the extended resource set are performed on three datasets namely, news set, financial news set and historical text set for classifying ENAMEX, NUMEX and TIMEX entities and results obtained outperform rule-based recognizer.

6.2.3. Unsupervised NERC systems

Wikipedia, a major source of data, is frequently used in recent research issues. Bhagavatula et al. (2012) [51] have exploited the whole structure of Wikipedia to extract the named entities by employing co-occurrence frequency of words between English and Indian languages. The method used by the authors has explored English Wikipedia data to bootstrap the identification of NEs in Hindi and Marathi which aids to improve multilingual entity filling, the task of mapping between closely related multilingual content. The task is performed on 3853 English, 2935 Hindi and 2622 Marathi Wikipedia articles. The objective of this paper is to extract the list of cricketers of various countries in different languages. Experimental results have shown that the complete structure of Wikipedia (including unstructured pages, subtitles, abstracts, infoboxes) is best for extracting named entities because Hindi and Marathi have already limited amount of data in Wikipedia articles. The proposed system also outperforms Hindi NERC system developed by LTRC (Language Technology Research Center, IIIT Hyderabad) [141] by giving 80.42% F-score for Hindi NER and 81.25% F-score for Marathi NER. Through the multilingual entity filling, the multilingual content has been increased by 64% which makes this approach very important.

Most of the studies in Biomedical NERC are rule-based and supervised tools based which make the system portable to different genres of text. Here Zhang and Elhadad (2013) [134] have presented an unsupervised approach which includes the stepwise solution from seed term extraction to noun phrase chunker, IDF filter, distributional semantic similarity based classification for Biomedical Named Entity Detection. The authors have also exploited shallow syntactic analysis and lexical semantics in different phases. Experiments are done on two main biomedical datasets named I2B2 Challenge [160] and GENIA corpus [161]. Three named entities from I2B2 corpus are extracted: problem, treatment and test and five named entities from GENIA corpus are explored:RNA, DNA, protein, cell_line, and cell_type. The authors have shown that the quality of seed term and usage of noun phrases for boundary detection are very important to make the system effective. IDF filter based candidate filtering and classification based on distributional similarity have brought the competitive performance on entity classification. Evaluation results include 69.5% accuracy on I2B2 corpus and 53.8% accuracy on GENIA Corpus which outperforms MetaMap system, another unsupervised approach.

Konkol et al. (2015) [135] have designed a new approach for Named Entity Recognition which considers latent semantic features including an automatic creation of gazetteers exploiting word similarity features. Word similarity-based algorithms used for local context are Hyperspace Analogue to Language (HAL) [162], Correlated Occurrence Analogue to Lexical Semantic (COALS) [163], (RI) [164], Bound Encoding of the AggreGate Language Environment (BEAGLE) [165], Purandare and Pedersen (P&P) [166] and for finding global contextual similarity, Latent Dirichlet Allocation (LDA) [167] is used. Besides it, high Precision stemmer [168], an unsupervised stemmer is used to explore stem related features. Both word based and stem-based features yield the best performance. English, Spanish, Dutch and Czech CoNLL corpora, including

25,000 tokens approximately, is used for detecting person, location, organization and miscellaneous entities using BIO format. The best achieved F-score for English, Dutch, Spanish and Czech is 89.44%, 83.01%, 83.08% and 74.08% respectively. So the authors have proved that the use of latent semantic features and LDA, which is never used before, make the NERC system fully language independent.

6.3. Hybrid systems

Li et al. (2009) [55] have presented a CRF based two-phase Named Entity Recognition System. In the first phase, all biomedical entities (DNA, RNA, protein, cell_type, cell_line) on JNLPBA-2004 [42] data are detected as one type using relevant features. The second phase classifies the one type of entities into their respective classes using BIO and BIOEW label model. Post-processing algorithms (Paired Punctuation Marks Expansion, Part of Speech Expansion, Forward Maximum Match) have been applied to the output of both phases which boosts the performance giving the F-score value of 74.31%.

Guanming et al. (2009) [101] have proposed a Chinese Named Entity Recognition System using a hybrid technique. The authors have generated the CRF model trained on SIGHAN2007 MSRA corpus and detected three named entities i.e. person, organization and location using Template-3. The entities are labeled with five tags including B1, B2, I, E, O. As the performance of the system is found unsatisfactory, so the authors perform post-processing in the next step which includes some rules and transformation-based learning. The current system is compared with the system which uses Template-5 with tag set of four labels including B, I, E, O. The authors find that the current system with Template-3 is giving more promising results with F-score of 93.49% and uses less time and fewer system resources for training.

Ekbal and Saha (2011) [102] have proposed a novel classifier ensemble technique based on Archived Multiobjective Simulated Annealing (AMOSA) algorithm for multi-objective optimization. The authors have experimented with several objective functions such as overall Precision and Recall values (MOO1), F-score of each NE class (MOO2), Precision and Recall of each NE class (MOO3), F-score value of NE boundary detection (MOO4) to model Hindi, Telugu and Bengali NERC System. Dataset used for Hindi and Telugu is collected from IJCNLP-08 workshop [141] and for Bengali, it is collected from Bengali News Corpus [159]. Different models are produced using different feature sets and three classifiers named as CRF, ME and SVM. The best classifier is selected based on highest Recall, Precision and F-score value for each class. The best F-score value is 92.80% for Hindi, 94.55% for Bengali, 89.85% for Telugu and is obtained using aMOO4 objective function which highlights that proper boundary identification plays a key role in entity identification.

A multi-strategy approach to recognize the biomedical entities as proposed by Etkinson and Bull (2012) [169] is a quite promising task without using any external lexical resources (dictionary, ontologies) and post-processing tasks. The main focus is given on preprocessing such as tokenization, POS tagging, lemmatization, stop words removal which enhances the performance by combining the results of two classification methods (SVM and HMM). The training and testing are done on standard Biocreative Corpus [170] containing 7500 sentences of the GENETAG corpus, giving a total of 8876 gene/protein names. The best results i.e. Precision 91%, Recall 80.1%, F-score 85.14% has been obtained by using 5-gram context feature and assembling multi-classification methods.

Küçük and Yazici (2012) [103] have proposed a hybrid Named Entity Recognizer in Turkish text. It is called hybrid as it is using the same information sources such as lexical resources and pattern bases used by the rule-based recognizer [171] for the recognition task as well as it has the ability to extend the information resources by rote learning from annotated data. This is the first Turkish hybrid system which can be easily ported to other genres of text provided the annotated dataset should be available. The performance of the system is evaluated on four genres of text i.e. news text, financial news text, child stories text and historical text through 10-fold cross-validation. On evaluation, it is found that hybrid system has outperformed the rule-based system by giving the best F-score value of 92.47% in case child stories text, 90.13% in case of news text, and 80.66% in case of historical text and 76.80% in case of financial news text when the capitalization feature is on. Besides it, various application areas are mentioned where the hybrid NER can be used such as event extraction, video indexing system, etc. In case of event extraction, it can be used for filling event slots of the agent, the object, the location and date/time of the event. In addition to this, the extracted entities can be used as an index term in information retrieval systems for Turkish [172]. The authors have proved the utilization of this hybrid recognizer in video indexing system [173] where it shows 82.78% accuracy in case of sliding video (noisy) text and 87.93% accuracy in case of perfect video transcript.

Ekbal et al. (2012) [174] have proposed multi-objective optimization genetic algorithm NSGA-II (non-dominated sorting genetic algorithm) [175] based classifier ensemble of CRF and SVM for Biomedical Named Entity Recognition. Various orthographical, contextual features are extracted from benchmark dataset of JNLPBA 2004 shared task [42] which is converted into BIO format. The training dataset includes 2000 abstracts of about 500k word forms and testing dataset includes 404 abstracts of around 100k words. The entities extracted are protein, RNA, DNA, cell_type, cell_line. On evaluation, it is found that the proposed approach outperforms best individual model, Baseline1, Baseline2, and Baseline3 by 0.90%, 2.48%, 2.21% and 1.88% respectively. The proposed approach is also evaluated on other datasets like AIMed and GENETAG. Evaluation results with AIMed dataset are Precision, Recall and F-score values of 94.81%, 96.08%, 95.44%, respectively and with GENETAG dataset are 98.45%, 98.05% and 98.25%, respectively.

Finding named entities out of the tweets is quite challenging due to its short and noisy nature, lack of information in a single tweet, slang expressions and informal abbreviations used in tweets, etc. Liu and Zhou (2013) [47] have proposed a novel hierarchical two-stage approach for NER in tweets. In the first stage, the tweets are pre-labeled based on linear CRF model [176]; then tweets with similar contents are put in the clusters enabling the enhanced CRF model to refine the labels of each tweet in each cluster using cluster level information as well as using conventional features and features derived from the clusters of pre-labeled results. Total 12,245 English tweets are manually annotated using BILOU (Beginning, Inside, Last token of the multi-token word, Outside, Unit word) schema which is furthered clustered into 1815 groups. Features used in this task are orthographic features, gazetteer related features and lexical features adopted from [177]. The system shows the 5-fold cross-validation results using Precision, Recall and F-score values of 84.8%, 80.4%, and 82.5% respectively which outperforms the dictionary based tweet NER and baseline model without using the second stage. The authors also report that the main errors occur due to slang expressions, data sparseness and high error-prone nature of tweets which can be further removed by using tweet normalization technology.

Saha and Ekbal (2013) [178] have presented a classifier ensemble based Named Entity Recognition System. Several supervised learning algorithms such as Maximum Entropy, Naive Bayes, Conditional Random field, Memory Based Learner, Decision Tree, Support Vector Machine and Hidden Markov Model are used to produce different models to get the best results. Besides it, two classifier ensemble methods i.e. binary vote based ensemble and real vote based ensemble have also been applied. Single Objective

Optimization (SOO) function based on GA (Genetic Algorithm) and Multi-objective Optimization (MOO) function based on NSGA-II (Non-dominated Sorting Genetic Algorithm) are used to quantify the weights of votes for each class. The experiments are done on Hindi, Telugu and Bengali data which is taken from NERSSEAL shared task and Bengali News Corpus [159]. The system is evaluated using Recall, Precision and F-score for each language and comparisons are made with the best individual classifier, different baseline classifiers, SOO with a binary vote based classifier ensemble, MOO with a binary vote based classifier ensemble. On comparison, it is found that SOO and MOObased classifier ensemble using real voting outperforms binary vote based approaches as well as baseline methods. Results also show that MOO with real voting is more effective than SOO based classifier ensemble. The results given by MOO with real voting are Recall of 94.21%, Precision of 94.72% and F-score of 94.74% for Bengali, Recall of 99.07%, Precision of 90.63% and F-score of 94.66% for Hindi and Recall of 82.79%, Precision of 95.18% and F-score of 88.55% for Telugu.

Li et al. (2013) [179] have developed a biomedical Named Entity Recognition System which works in two phases. In the first phase, the entities are detected using two-layer stacking method. In layer 1, results of six different classifiers are integrated together to get the better results. These results are treated as a feature vector by layer 2. Layer 2 uses this feature vector to construct new training and testing data. This training data is further used to generate a model using the best learning algorithm to get the desired results. The classifiers used in this study are CRF++, FMallet, BMallet, Yamcha1vs1 and Yamcha1vsall, ME. In the second phase, five agents are constructed which extracts five different named entities out of the pool of entities extracted at first phase using CRF++ technique which is found as the best technique. The result obtained by the system is the F-score value of 76.06% which is much better than other two-layer stack based systems.

Munkhjargal et al. (2015) [33] have proposed a classifier ensemble based Named Entity Recognition approach for Mongolian language which is a quite challenging task due to agglutinative morphology and complex structure of Mongolian language. A number of models are constructed using different combinations of features by exploiting the properties of three classifiers namely Maximum Entropy, SVM and CRF. The performance of machine learning approaches is enhanced by the use of gazetteers as well as the string matching patterns in order to handle the out-of-vocabulary words. A Mongolian POS-tagged corpus [180] is manually annotated which includes 310 articles, 14,837 sentences, about 277,000 tokens, 4932 location names, 4382 personal names and 3366 organization names. Classifier ensemble of five models are done by using genetic algorithm [181] with majority voting mechanism which has shown the best performance by giving 82.72% F-score for ME, 87.36% F-score for CRF and 86.43% F-score for SVM.

Comparison among different Hybrid NERC systems is given in Table 4.

6.4. Analysis of results

In this subsection, our findings on the basis of experimental results are presented. Different NERC systems including rule-based to machine learning based to hybridized NERC systems have been introduced so far.

Table 2 presents rule-based NERC systems. Generally, Rule-based NERC systems perform well because they follow language-specific rules and language specific resources like gazetteers, POS tagger, morphological analyzer, stemmer, etc. The performance of Arabic NERC system [109] is found to be average than another Arabic NERC system [105] because the former focuses more on language-independent rules instead of using dictionaries and local grammar. Urdu rule-based NERC system [107] is showing low results due to the recognition of specific, highly inflectional entity "izaafats". Rule-based NERC system in consumer, electronic and automotive domain [108] is showing inferior results due to improper handling of unstructured data and non-usage of NLP artillery such as sentence parsing, part of speech tagging, etc.

Table 3 presents supervised, semi-supervised and unsupervised NERC systems. These machine learning based NERC systems are showing better performance to rule-based systems. A comparison between features of supervised and rule-based NERC systems is shown in Table 5. The performance of supervised NERC systems is highly dependent on corpus specific information and classifiers used for training and testing. In this study, the performance of several NERC systems using different classifiers namely CRF, SVM, MaxEnt, HMM, Naïve Bayes, etc. is shown. CRF based NERC systems [117,119,120] are proved to be good because CRF takes contextual information into account while predicting named entities' class. Context-related features are helpful in handling ambiguous word forms found in data.

Table 4 presents hybrid NERC systems. These systems are made with the combination of two or more techniques. Hybrid NERC systems are highly concerned with improving their accuracy rate by developing new techniques using strongest points of each one. Classifier ensemble method is one of the best methods for hybridization in which results of two or more classifiers are joined together on the basis of some optimization technique. Saha and Ekbal [178] have introduced a multi-objective based classifier ensemble NERC system by combining predictions of seven base classifiers namely, Naïve Bayes, HMM, Memory Based Learner, Decision Tree, CRF, MaxEnt, and SVM. This system shows superior classification accuracy than baseline ensembles.

Comparison of the performance of these systems is not possible only on the basis of one or two factors. Different factors affect the outcome of NERC systems which includes language and domain factor, a number of entities extracted, a technique used, resources available, features used, etc. English and other European languages provide capitalization clue which is a great feature aiding in recognition of named entities while Asian and other languages like Hindi, Bengali, Telugu, Arabic, Chinese, etc. suffer due to lack of this feature. Domain factor also affects performance metrics. Recognition of named entities in the biomedical domain or unstructured data such as SMSes, tweets, microtext, handwritten data requires extraction of more complex features thus increasing the computational task of the system.

Entity type factor is another reason for the variation in performance metrics. Coarse-grained NERC systems are more accurate than fine-grained NERC systems. Three main challenges are faced by fine-grained NERC systems: selection of optimal tag set, preparation of training and testing dataset and development of a fast and accurate multi-class labeling method [62]. Considering the boundaries of named entities' class also affects the performance of NERC systems. Different segment representation techniques such as BIO, BIEO, BILOU, etc. are presented by different researchers which are responsible for mapping multi-word entities. More complex segment representation scheme increases the recognition performance [131].

Selection of appropriate technique is a big challenge for recognizing named entities. Indian and other languages like Hindi, Bengali, Urdu, Oriya, Telugu, Arabic, Chinese, Malay, etc. are resource-scarce languages so supervised learning methods along with language-independent features or hybrid methods are a good substitute to handcrafted rules which provide a tremendous output at a high system engineering cost. Semi-supervised and unsupervised NERC systems are found effective for the languages which suffer from a scarcity of tagged dataset. Bootstrapping method [51] and word similarity-based algorithms [135] are used for recognizing named entities from data belonging to the biomedical, unstructured and open domain.

Table 4
Hybrid NERC systems.

Authors & year	Language/domain	NEs found	Technique used	Dataset used	Evaluation results
Li et al. 2009 [55]	Biomedical domain	Protein, RNA, DNA, cell_line, cell_type	CRF + post processing algorithms	Dataset is collected from GENIA Corpus	F-score: 74.31%
Guanming et al., 2009 [101]	Chinese	Person, location, organization	CRF + transformation based learning + rules	Dataset is collected from SIGHAN2007 MSRA Corpus	F-score 93.49%
Ekbal and Saha, 2011 [102]	Hindi, Bengali and Telugu	Person, organization, location, event, facility, time, date, relationship, title-person, designation, number, measure, artifact and abbreviation	Classifier ensemble technique	The dataset for Hindi and Telugu is collected from IJCNLP-08 and Bengali from Bengali news corpus	F-score: Hindi 92.80% Bengali 94.55% Telugu 89.85%
Etkinson and Bull, 2012 [169]	Biomedical domain	Gene and protein names	SVM+HMM	7500 sentences from BioCreative corpus	Precision: 91% Recall: 80.1% F-score: 85.14%
KuKuk and Yazici, 2012 [103]	Turkish	Person, location, organization, date/time, money/percent	Lexical resources + pattern base + rote learning	Text, financial news text, child stories text and historical text collected from METU Turkish corpus	F-score: child stories: 92.47% news text: 90.13% historical text: 80.66% financial text: 76.80%
Ekbal et al. 2012 [174]	Biomedical domain	Protein, RNA, DNA, cell_type, cell_line	Classifier ensemble of CRF and SVM	Dataset is collected from GENIA Corpus and system is evaluated against AIMed and GENETAG dataset	AIMed dataset: Precision: 94.81% Recall: 96.08% F-score: 95.44% GENETAG dataset: Precision: 98.45% Recall: 98.05% F-score: 98.25%
Liu and Zhou, 2013 [47]	English tweets	Person, location, organization, product, other	Linear CRF + Cluster based approach	12,245 randomly sampled tweets are collected from Twitter	5-fold cross-validation showed Precision: 84.8% Recall: 80.4% F-score: 82.5%
Saha and Ekbal, 2013 [178]	Hindi, Bengali, Telugu	Person, location, organization, miscellaneous	Classifier ensemble using MOO	Dataset for Hindi and Telugu is collected from IJCNLP-08 and Bengali from Bengali news corpus	Hindi: Precision: 90.63% Recall: 99.07% F-score : 94.66% Bengali: Precision: 94.72% Recall: 94.21% F-score 94.74% Telugu: Precision: 95.18% Recall: 82.79% F-score: 88.55%
Li et al. 2013 [179]	Biomedical domain	Protein, DNA, RNA, cell_type, cell_line	Combination of FMallet, BMallet, CRF++, Yamcha1vs1 and Yamcha1vsall, ME results	Dataset is collected from GENIA Corpus	F-score: 76.06%
Munkhjargal et al. 2015 [33]	Mongolian	Proper names	Classifier ensemble of Maximum Entropy, SVM and CRF.	310 articles from Mongolian POS tagged corpus	F-score: ME: 82.72% CRF:87.36% SVM: 87.43%

Table 5
Comparison of rule-based and Supervised NERC systems.

Features	Rule-based NERC systems	Supervised NERC systems
Need of tagged corpus	✗	✓
Use of linguistic resources (POS tagger, Gazetteers)	✓	✗
Involvement of linguistic experts	✓	✗
Involve computations	✗	✓
Need of big data	✗	✓

7. Base classifiers for Named Entity Recognition

Different machine learning classifiers are used by different researchers for the NERC task. A brief description of some of the classifiers is given below.

7.1. Naïve Bayes

Naive Bayes [182] is a simple and probabilistic classifier. It is mainly used in resolving issues of various streams including Natural Language Processing, Information Extraction, Pattern

Recognition, etc. [183–186]. This classifier follows Bayes' theorem with a strong assumption that the attributes used for classification are conditionally independent. Naïve Bayes algorithm computes the mean value and variance of the attributes of each class for classification purpose. The classification performance of the system is highly dependent on the calculated variance. The performance will increase as the calculated variance goes down.

Let t be the possible target or class and x_i be a vector. Here the final objective is to calculate the probability that the vector x_i belongs to the class t [182].

$$P(t|x_i) = P(t) * \frac{P(x_i|t)}{P(x_i)}. \quad (1)$$

Here, $P(t)$ is the priori means prior probability of class (t) before the feature vector is given. $P(x_i)$ is the prior probability of the feature vector (x_i). $P(x_i|t)$ is the likelihood which is the probability of vector given class. Finally, $P(t|x_i)$ is the posterior probability of class (target) when the feature vector is given.

7.2. Conditional Random Field (CRF)

Conditional Random Field [176] is a probabilistic graphical model which is frequently used for sequence labeling tasks like Part of Speech (POS) Tagging, Object Recognition, Named Entity Recognition, etc. CRF is conditionally trained model which can easily work with a vast amount of non-independent features. Unlike discrete classifiers, CRF has a special property of considering neighboring examples. It takes into account the contextual features while predicting the sequence of labels for a sequence of input samples.

CRF is used to calculate the conditional probability of a class sequence $c = \langle c_1, c_2, \ldots, c_N \rangle$ given an observation sequence $o = \langle o_1, o_2, \ldots, o_N \rangle$ which is calculated as follows:

$$P \wedge (c|o) = \frac{1}{Z_0} \exp\left(\sum_{n=1}^{N}\sum_{k=1}^{K} \lambda_k * f_k(n, o, c_{n-1}, c_n)\right). \quad (2)$$

Here, $f_k(n, o, c_{n-1}, c_n)$ is a feature function which considers previous and current class label to compute the probability. λ_k is the learning weight which is calculated via training. Z_0 is the normalization factor which makes conditional probabilities sum up to 1.

7.3. Support Vector Machine (SVM)

Support Vector Machine is a supervised and discriminative learning model which outputs a linear hyperplane that separates the underlying data on either side according to the positive and negative category. SVM is a non-probabilistic linear classifier which achieves higher accuracy even dealing with a large number of features without falling into over-fitting [187]. SVM is mainly used to deal with classification and regression problems. SVM works well with a binary-class problem: $\{(P_1, T_1), \ldots, (P_N, T_N)\}$ where $Pi \in Fv$ is the feature vector of the ith instance in the training data and $Ti \in \{0, 1\}$ is the target of the ith instance.

Basically, the main aim of SVM classifier is to divide the instances into a positive or negative group with maximum margin.

$$(w.p) + m = 0 \quad w \in Fv,\ m \in F. \quad (3)$$

Here, the linear separator is defined by two elements: a weight vector w (with one component for each feature), and a maximum margin m which stands for the distance of the hyperplane to the origin.

7.4. Hidden Markov Model (HMM)

Hidden Markov Model is a generative statistical model which assigns probable target sequence to each word sequence following the Viterbi algorithm [188]. HMM is able to capture the locality of phenomena which enhances its evaluation performance. Formally, HMM can be defined as follows:

$\lambda = (X, Y, pi)$. Here, X represents transition probability. Y represents emission probability and pi represents start probability [189].

$X = xij =$ (Number of transitions from state si to sj)/(Number of transitions from state si) [189].

$Y = yj\,(m) =$ (Number of times in state j and observing symbol m)/(expected number of times in state j) [189].

The Viterbi algorithm considers only the most probable state sequences out of all the state sequences to find the most appropriate tag sequence in linear time from the pool of the possible tag distribution based on the state transition probabilities [41].

7.5. Maximum Entropy (MaxEnt)

Maximum Entropy classifier [190] is a probabilistic exponential classifier which can be used to solve text classification problems such as language detection, sentiment analysis, named entity recognition and many more. The MaxEnt works on the principle of Maximum Entropy which states that from all the models that fit our training data, it selects the one which has the largest entropy. The MaxEnt does not assume that the features are conditionally independent of each other. The exponential form of Maximum Entropy is as follows:

$$P(c|m) = \frac{1}{Z(m)} \exp(\sum_{j=1}^{n} \lambda_j f_j(m, c)) \quad (4)$$

where, c is the class or tag, m is the context, $f_j(m, c)$ are the features along with associated weight λ_j and $Z(m)$ is a normalization function.

8. Evaluation measures of NERC

Different evaluation measures for examining the performance of NERC systems are discussed in the literature. The evaluation is done basically to check the ability of tool on finding correct entity types and their boundaries. For this, the NERC system's predictions are compared with the predictions made by human annotators. The intrinsic evaluation metrics used for the comparison are Precision, Recall and F-score.

8.1. Precision, recall and F-score

Precision, Recall and F-score are calculated on the basis of true positives (TP), false positives (FP) and false negatives (FN). True positives are the correctly labeled instances. False positives are the incorrectly labeled instances and false negatives are the missed out instances by the system. F-score is the weighted mean of Precision and Recall. These metrics are formulated as given below.

$$Precision = \frac{number\ of\ instances\ correctly\ labeled\ by\ the\ system}{total\ number\ of\ instances\ labeled\ by\ the\ system} \quad (5)$$

$$Recall = \frac{number\ of\ instances\ correctly\ labeled\ by\ the\ system}{total\ number\ of\ relevant\ instances\ labeled\ by\ the\ system} \quad (6)$$

$$F\text{-}score = 2 * \frac{Precision \cdot Recall}{Precision + Recall}. \quad (7)$$

In simple words, Precision is the ratio of correctly classified entities over total detected NEs. Recall is the ratio of relevant NEs over total detected entities by the system.

8.2. Matching predictions against Gold standard

Named Entity Detection involves finding the correct entity boundaries as well as a correct entity type. Most of the systems

require an exact match on both entity type and boundary. The shared task for CoNLL-2003 [10] is one of the examples of exact matching.

The MUC-6 [5] events have defined a more loosened scheme which allows a partial credit for the systems finding correct boundaries regardless of the type as well as finding correct type regardless of boundaries.

ACE [11] has proposed the most complex form of evaluation, which is based on the fact that each NE class is assigned a weighted parameter that contributes up to a maximum proportion of the final score. This measurement resolves the issues like a partial match, wrong type, etc. as well as considers the sub-types of NEs.

8.3. Macro-and micro averaged F-score

As most of the NERC systems involve multiple entities types, so it is often required to assess the performance of the system for all entity classes. Two measures are considered for this: Macro-averaged F-score and Micro-averaged F-score. Macro-averaged F-score is the average of F-scores of all entity classes in the corpus while micro-averaged F-score is calculated by adding the number of labeled entities together and then calculating Precision, Recall and F-score. The difference is that the micro-averaged measure can be badly affected by the larger classes in the corpus suppressing the performance of the system on smaller classes. However, MUC's final score is calculated using micro-averaged F-score.

8.4. Cross-validation

Cross-validation is the measure used by different researchers [47,124,127]. This technique is the balanced version of evaluation, normally used for supervised learning methods. Cross-validation is based on the idea of dividing the dataset into n chunks and treating all the chunks except one for modeling the system. This process is usually repeated for k iterations that is why it is known as k-fold cross-validation. In each iteration, a different chunk is left and used for testing. The final score is calculated by averaging the results obtained in all iterations. Usually, 10-fold cross-validation is widely used for NERC task.

9. Future directions in NERC

Tremendous research has been done in the field of NERC over the past 20 years. A number of state-of-the-art approaches have been proposed by different researchers, facing different issues and challenges and resolving them to meet the actual purpose of the NERC system i.e. proper recognition and classification of named entities in different NE types. The NERC system has been proved to be an excellent pre-processing tool in numerous natural language applications like Question Answering, Automatic Text Summarization, Machine Translation, etc. Marreo et al. (2013) [191] have thrown a light on the general belief that NER is a solved task. But the authors presented different evaluation forums in NERC task indicates that NERC research should be considered back by the research community. This research field is improving continuously. Therefore, in this section, the focus is emphasized on different issues arising in this field which needs to be addressed by the research community.

Existing Named Entity Recognition Systems are incorporating new techniques like new machine learning methods are being employed to build NERC systems. The performance of machine learning methods is highly dependent on the selection of appropriate features that explore the sensitivity of each entity type enabling generation of the reliable model which further promotes actual classification. But there is not much change in the features (list lookup, pattern matching) required to extract named entity classes. Therefore, some new and language independent features need to be discovered that can detect relevant named entities out of the text.

Large annotated corpora are the foremost requirement of supervised machine learning methods for training and testing which is a big challenge to obtain for many resource-poor languages, especially for Indian languages. Semi-supervised or unsupervised techniques require a fewer amount of annotated data or no annotated data for supervision. Therefore, future researchers can explore more semi-supervised and unsupervised methods which can categorize semantically related word forms into named entity classes.

Earlier studies have focused mainly on the extraction of coarse-grained entities like ENAMEX, TIMEX, and NUMEX. So future research can enhance the NERC task by considering fine-grained entities. Fine-grained entities will be able to broad up the scope of NERC in many other NLP applications along with robust information retrieval.

Many new approaches focusing on dealing with linguistic features improve the quality of NERC system but at the cost of more computational time and memory space because they need more linguistic knowledge and difficult linguistic techniques. Moreover, these systems require the employment of linguistic resources which is a big challenge to obtain for resource-poor languages. Therefore, there is a need to develop statistical Named Entity Recognition Systems that can detect relevant named entities with good quality.

Rule-based approaches are found domain dependent, language specific, require high maintenance cost and cannot be adapted to new languages and domains easily. On the other hand, machine learning approaches are robust, portable in different domains and languages, cost-effective but highly dependent on large annotated corpora which are a cumbersome task to obtain especially for resource-poor languages like most of the Indian languages. Therefore, the hybrid approaches need to be improved more. Hybrid approaches can be developed to produce a good quality NERC system by combining rule-based and machine learning based techniques together.

Besides above all, certain issues like ambiguity in the text, nested entities, suitable segment representation which is helpful in determining the entities' boundary, should be efficiently handled by each NERC system being developed in future. To remove ambiguity in the text, contextual features can be incorporated in our feature vector.

10. Conclusion

Named Entity Recognition is an emerging field and is continuously improving due to its major contribution in many natural language applications. The purpose of this article is to make researchers' aware of some important information related to the history of Named Entity Recognition, current state-of-the-art and some future possibilities. This survey will help novice researchers' to gain insight into issues and challenges related to the categorization of named entities. This article presents an introduction of Named Entity Recognition approaches in a systematic manner. In this article, an exhaustive review of rule-based, learning based and hybrid NERC systems has been presented. In addition, all these systems have been compared in a tabular form, providing some more useful information about these approaches. Certain factors affecting the performance of the NERC task like language factor, entity type factor, textual genres/domain factors, etc. have also been highlighted. The results presented by different NERC systems are analyzed and discussed in detail. A brief description of various classifiers is also provided. There is well-established practice for evaluating NERC systems. So evaluation measures are also discussed in detail. Finally, some good future directions are also indicated that will help researchers in enhancing NERC techniques so that this research field grows continuously.

References

[1] C. Nobata, S. Sekine, H. Isahara, R. Grishman, Summarization system integrated with named entity tagging and ie pattern discovery, in: LREC, 2002 pp. 1742–1745.

[2] B. Babych, A. Hartley, Improving machine translation quality with automatic named entity recognition, in: Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools: Resources and Tools for Building MT, Association for Computational Linguistics, 2003, pp. 1–8.

[3] T. Mandl, C. Womser-Hacker, The effect of named entities on effectiveness in cross-language information retrieval evaluation, in: Proceedings of the 2005 ACM Symposium on Applied Computing, ACM, 2005, pp. 1059–1064.

[4] L.A. Pizzato, D. Molla, C. Paris, Pseudo relevance feedback using named entities for question answering, in: Proceedings of the 2006 Australian Language Technology Workshop, ALTW-2006, 2006, pp. 89–90.

[5] R. Grishman, B. Sundheim, Message understanding conference-6: A brief history, in: COLING, 96, 1996, pp. 466–471.

[6] M. Pasca, D. Lin, J. Bigham, A. Lifchits, A. Jain, Organizing and searching the World Wide Web of facts-step one: The one-million fact extraction challenge, in: AAAI, 6, 2006, pp. 1400–1405.

[7] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvist. Investig. 30 (1) (2007) 3–26.

[8] S. Satoshi, I. Hitoshi, IREX: IR and IE evaluation project in Japanese, in: Proceedings of the 2nd International Conference on Language Resources & Evaluation, 2000.

[9] E.F.T.K. Sang, Introduction to the CoNLL-2002 shared task: Language-indepen dent named entity recognition, in: Proceedings of Natural language learning, Association for Computational Linguistics, 2002, pp. 155–158.

[10] E.F.T.K. Sang, F.D. Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, 4, Association for Computational Linguistics, 2003, pp. 142–147.

[11] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weische del, The Automatic Content Extraction (ACE) program — tasks, data, and evaluation, in: LREC, 2, 2004, p. 1.

[12] D. Santos, N. Seco, N. Cardoso, R. Vilela, HAREM: An advanced NER evaluation contest for Portuguese, in: LREC, 2006, pp. 1986–1991.

[13] L. Coheur, A. Guimaraes, N. Mamede, Supporting named entity recognition and syntactic analysis with full-text queries, in: International Conference on Application of Natural Language to Information Systems, Springer, Berlin, Heidelberg, 2008, pp. 341–342.

[14] Z. Liu, C. Zhu, T. Zhao, Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words? in: Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, Springer, Berlin, Heidelberg, 2010, pp. 634–640.

[15] Y. Miao, L. Yajuan, L. Qun, S. Jinsong, X. Hao, Chinese named entity recognition and disambiguation based on Wikipedia, in: Natural Language Processing and Chinese Computing, Springer, Berlin, Heidelberg, 2012, pp. 272–283.

[16] Z. Liao, Z. Zhang, Y. Liu, Chinese named entity recognition based on hierarchical hybrid model, in: Pacific Rim International Conference on Artificial Intelligence, Springer, Berlin Heidelberg, 2010, pp. 620–624.

[17] C.W. Wu, R.T.H. Tsai, W.L. Hsu, Semi-joint labeling for Chinese named entity recognition, in: Asia Information Retrieval Symposium, Springer, Berlin, Heidelberg, 2008, pp. 107–116.

[18] Z. Wu, Z. Yu, J. Guo, C. Mao, Y. Zhang, Fusion of long distance dependency features for Chinese named entity recognition based on Markov logic networks, in: Natural Language Processing and Chinese Computing, Springer, Berlin, Heidelberg, 2012, pp. 132–142.

[19] A.L.F. Han, D.F. Wong, L.S. Chao, Chinese named entity recognition with conditional random fields in the light of Chinese characteristics, in: Language Processing and Intelligent Information Systems, Springer, Berlin, Heidelberg, 2013, pp. 57–68.

[20] J. Lu, M. Ye, Z. Tang, X.J. Huang, J.L. Ma, A novel method for Chinese named entity recognition based on character vector, in: International Conference on Collaborative Computing: Networking, Applications and Worksharing, Springer International Publishing, 2015, pp. 141–150.

[21] L. Buitinck, M. Marx, Two-stage named-entity recognition using averaged perceptrons, in: International Conference on Application of Natural Language to Information Systems, Springer, Berlin, Heidelberg, 2012, pp. 171–176.

[22] A. Azpeitia, M. Cuadros, S. Gaines, G. Rigau, NERC-fr: Supervised named entity recognition for French, in: International Conference on Text, Speech, and Dialogue, Springer International Publishing, 2014, pp. 158–165.

[23] Y. Mosallam, A. Abi-Haidar, J.G. Ganascia, Unsupervised named entity recognition and disambiguation: an application to Old French journals, in: Industrial Conference on Data Mining, Springer International Publishing, 2014, pp. 12–23.

[24] D. Küçük, Named entity recognition experiments on Turkish texts, in: Interna tional Conference on Flexible Query Answering Systems, Springer, Berlin, Heidelberg, 2009, pp. 524–535.

[25] S.R. Yavuz, D. Küçük, Named entity recognition in Turkish with Bayesian learning and hybrid approaches, in: Information Sciences and Systems 2013, Springer International Publishing, 2013, pp. 129–138.

[26] H.M. Mo, K.T. Nwet, K.M. Soe, CRF-Based named entity recognition for Myanmar language, in: International Conference on Genetic and Evolutionary Computing, Springer International Publishing, 2016, pp. 204–211.

[27] D.B. Nguyen, S.H. Hoang, S.B. Pham, T.P. Nguyen, Named entity recognition for Vietnamese, in: Asian Conference on Intelligent Information and Database Systems, Springer, Berlin, Heidelberg, 2010, pp. 205–214.

[28] D.T. Vo, C.Y. Ock, A hybrid approach of pattern extraction and semi-supervised learning for Vietnamese named entity recognition, in: Interna tional Conference on Computational Collective Intelligence, Springer, Berlin, Heidelberg, 2012, pp. 83–93.

[29] V.H. Nguyen, H.T. Nguyen, V. Snasel, Named entity recognition in Vietnamese tweets, in: International Conference on Computational Social Networks, Springer International Publishing, 2015, pp. 205–215.

[30] K. Shaalan, H. Raza, Arabic named entity recognition from diverse text types, in: Advances in Natural Language Processing, Springer, Berlin, Heidelberg, 2008, pp. 440–451.

[31] S. Abdallah, K. Shaalan, M. Shoaib, Integrating rule-based system with classification for Arabic named entity recognition, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg, 2012, pp. 311–322.

[32] M.A. Meselhi, H.M.A. Bakr, I. Ziedan, K. Shaalan, A novel hybrid approach to Arabic named entity recognition, in: China Workshop on Machine Translation, Springer, Berlin, Heidelberg, 2014, pp. 93–103.

[33] Z. Munkhjargal, G. Bella, A. Chagnaa, F. Giunchiglia, Named entity recognition for Mongolian language, in: International Conference on Text, Speech, and Dialogue, Springer International Publishing, 2015, pp. 243–251.

[34] M. Marcińczuk, M. Piasecki, Study on named entity recognition for Polish based on hidden Markov models, in: International Conference on Text, Speech and Dialogue, Springer, Berlin, Heidelberg, 2010, pp. 142–149.

[35] R. Gareev, M. Tkachenko, V. Solovyev, A. Simanovsky, V. Ivanov, Introducing baselines for Russian named entity recognition, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg, 2013, pp. 329–342.

[36] A.K. Singh, Named entity recognition for south and South East Asian languages: Taking stock, in: IJCNLP, 2008, pp. 5–16.

[37] W. Li, A. McCallum, Rapid development of Hindi named entity recognition using conditional random fields and feature induction, ACM Trans. Asian Lang. Inf. Process. 2 (3) (2003) 290–294.

[38] V. Garg, N. Saraf, P. Majumder, Named entity recognition for Gujarati: A CRF based approach, in: Mining Intelligence and Knowledge Exploration, Springer International Publishing, 2013, pp. 761–768.

[39] S. Banerjee, S.K. Naskar, S. Bandyopadhyay, Bengali named entity recognition using margin infused relaxed algorithm, in: International Conference on Text, Speech, and Dialogue, Springer International Publishing, 2014, pp. 125–132.

[40] L. Jimmy, D. Kaur, Named entity recognition in Manipuri: A hybrid approach, in: Language Processing and Knowledge in the Web, Springer, Berlin, Heidelberg, 2013, pp. 104–110.

[41] D. Shen, J. Zhang, G. Zhou, J. Su, C.L. Tan, Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain, in: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, 13, Association for Computational Linguistics, 2003, pp. 49–56.

[42] The GENIA project. Online available at http://www.tsujii.is.s.u-tokyo.ac.jp/genia.

[43] S.K. Saha, S. Narayan, S. Sarkar, P. Mitra, A composite kernel for named entity recognition, Pattern Recognit. Lett. 31 (12) (2010) 1591–1597.

[44] T. Ek, C. Kirkegaard, H. Jonsson, P. Nugues, Named entity recognition for short text messages, Procedia Soc. Behav. Sci. 27 (2011) 178–187.

[45] M. Majumder, U. Barman, R. Prasad, K. Saurabh, S.K. Saha, A novel technique for name identification from homeopathy diagnosis discussion forum, Procedia Technol. 6 (2012) 379–386.

[46] J.J. Jung, Online named entity recognition method for microtexts in social networking services: A case study of Twitter, Expert Syst. Appl. 39 (9) (2012) 8066–8070.

[47] X. Liu, M. Zhou, Two-stage NER for tweets with clustering, Inf. Process. Manag. 49 (1) (2013) 264–273.

[48] Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu, Y. Jiang, Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study, J. Biomed. Inform. 47 (2014) 91–104.

[49] E. Yan, Y. Zhu, Identifying entities from scientific publications: A comparison of vocabulary-and model-based methods, J. Informetr. 9 (3) (2015) 455–465.

[50] A.P. Quimbaya, A.S. Múnera, R.A.G. Rivera, J.C.D. Rodríguez, O.M.M. Velandia, A.A.G. Peña, C. Labbé, Named entity recognition over electronic health records through a combined dictionary-based approach, Procedia Comput. Sci. 100 (2016) 55–61.

[51] M. Bhagavatula, S GSK, V. Varma, Named entity recognition an aid to improve multilingual entity filling in language-independent approach, in: Proceed ings of the First Workshop on Information and Knowledge Management for Developing Region, ACM, 2012, pp. 3–10.

[52] C. Adak, B.B. Chaudhuri, M. Blumenstein, Named entity recognition from unstructured handwritten document images, in: 12th IAPR Workshop on Document Analysis Systems, DAS, IEEE, 2016, pp. 375–380.
[53] Y. Wang, J. Patrick, Cascading classifiers for named entity recognition in clinical notes, in: Proceedings of the Workshop on Biomedical Information Extraction, Association for Computational Linguistics, 2009, pp. 42–49.
[54] D. Downey, M. Broadhead, O. Etzioni, Locating complex named entities in web text, in: IJCAI, 7, 2007, pp. 2733–2739.
[55] L. Li, R. Zhou, D. Huang, Two-phase biomedical named entity recognition using CRFs, Comput. Biol. Chem. 33 (4) (2009) 334–338.
[56] C. Meyer, H. Schramm, Boosting HMM acoustic models in large vocabulary speech recognition, Speech Commun. 48 (5) (2006) 532–548.
[57] M.J. Silva, B. Martins, M. Chaves, A.P. Afonso, N. Cardoso, Adding geographic scopes to web resources, Comput. Environ. Urban Syst. 30 (4) (2006) 378–399.
[58] N. Chinchor, P. Robinson, MUC-7 named entity task definition, in: Proceedings of the 7th Conference on Message Understanding, 1997, p. 29.
[59] S.K. Saha, S. Sarkar, P. Mitra, Feature selection techniques for maximum entropy based biomedical named entity recognition, J. Biomed. Inform. 42 (5) (2009) 905–911.
[60] R. Batista-Navarro, R. Rak, S. Ananiadou, Optimizing chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics, J. Cheminform. 7 (1) (2015) 1.
[61] C. Lee, Y.G. Hwang, H.J. Oh, S. Lim, J. Heo, C.H. Lee, J.H. Wang, M.G. Jang, Fine-grained named entity recognition using conditional random fields for question answering, in: Asia Information Retrieval Symposium, Springer, Berlin, Heidelberg, 2006, pp. 581–587.
[62] X. Ling, D.S. Weld, Fine-grained entity recognition, in: Proceedings of the Conference on Artificial Intelligence, AAAI, 2012.
[63] S. Keretna, C.P. Lim, D. Creighton, K.B. Shaban, Enhancing medical named entity recognition with an extended segment representation technique, Comput. Methods Programs Biomed. 119 (2) (2015) 88–100.
[64] L. Ramshaw, M.P. Marcus, Text chunking using transformation-based learning, in: Proceedings of the Third ACL Workshop on Very Large Corpora, Association for Computational Linguistics, 1995.
[65] A. Ratnaparkhi, Maximum Entropy Models for Natural Language Ambiguity Resolution (Doctoral dissertation), University of Pennsylvania, 1998.
[66] E.F. Sang, J. Veenstra, Representing text chunks, in: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1999, pp. 173–179.
[67] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, H. Isahara, Named entity extraction based on a maximum entropy model and transformation rules, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2000, pp. 326–335.
[68] R. Danger, F. Pla, A. Molina, P. Rosso, Towards a protein–protein interaction information extraction system: Recognizing named entities, Knowl.-Based Syst. 57 (2014) 104–118.
[69] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, D.S. Weld, Knowledge-based weak supervision for information extraction of overlapping relations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, Association for Computational Linguistics, 2011, pp. 541–550.
[70] S. Riedel, L. Yao, A. McCallum, Modeling relations and their mentions without labeled text, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 2010, pp. 148–163.
[71] N. Indurkhya, F.J. Damereau (Eds.), Handbook of Natural Language Processing, second ed., Chapman & Hall/CRC, Boca Raton, 2010.
[72] D. Mollá, M. Van Zaanen, D. Smith, Named entity recognition for question answering, in: Proceedings of the Australasian Language Technology Workshop, ALTW2006, 2006, pp. 51–58.
[73] R. Srihari, W. Li, Information extraction supported question answering, in: 8th Text Retrieval Conference, TREC-8, 500, 1999. pp. 185–196.
[74] Á. Rodrigo, J. Pérez-Iglesias, A. Peñas, G. Garrido, L. Araujo, Answering questions about european legislation, Expert Syst. Appl. 40 (15) (2013) 5811–5816.
[75] Y. Chen, C. Zong, K.Y. Su, A joint model to identify and align bilingual named entities, Comput. Linguist. 39 (2) (2013) 229–266.
[76] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A Method for Automatic Evaluation of Machine Translation. Technical Report RC22176, W0109-022, IBM Research Report, 2001.
[77] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, S. Shah, Multi-Document summarization based on the Yago ontology, Expert Syst. Appl. 40 (17) (2013) 6976–6984.
[78] M. Hassel, Exploitation of named entities in automatic text summarization for Swedish, in: NODALIDA'03, 14th Nordic Conference on Computational Linguistics, 2003, p. 9.
[79] T.H. Cao, T.M. Tang, C.K. Chau, Text clustering with named entities: A model, experimentation and realization, in: Data Mining: Foundations and Intelligent Paradigms, Springer, Berlin, Heidelberg, 2012, pp. 267–287.
[80] J. Zhang, Q. Dang, Y. Lu, S. Sun, Suffix tree clustering with named entity recognition, in: International Conference on Cloud Computing and Big Data, IEEE, 2013, pp. 549–556.
[81] C. Faloutsos, D.W. Oard, A Survey of Information Retrieval and Filtering Methods. Technical Report, 1998.
[82] T. Mandl, C. Womser-Hacker, The effect of named entities on effectiveness in cross-language information retrieval evaluation, in: Proceedings of the 2005 ACM Symposium on Applied Computing, ACM, 2005, pp. 1059–1064.
[83] J.B. Antony, G.S. Mahalakshmi, Content-based information retrieval by named entity recognition and verb semantic role labelling, J. UCS 21 (13) (2015) 1830–1848.
[84] H.J. Song, S.B. Park, S.Y. Park, An automatic ontology population with a machine learning technique from semi-structured documents, in: IEEE International Conference on Information and Automation, 2009, pp. 534–539.
[85] V. De Boer, M. Van Someren, B.J. Wielinga, Relation instantiation for ontology population using the web, in: Proceedings of the 29th Annual German Conference on AI, KI, 4314, 2007, pp. 202–213.
[86] O. Etzioni, M. Cafarella, D. Downey, A.M. Popescu, T. Shaked, S. Soderland, et al., Unsupervised named-entity extraction from the web: An experimental study, Artificial Intelligence 165 (1) (2005) 91–134.
[87] L. Zhang, B. Liu, Aspect and entity extraction for opinion mining, in: Data Mining and Knowledge Discovery for Big Data, 1, Springer, 2014, pp. 1–40.
[88] A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Natural Language Processing and Text Mining, Springer, 2007, pp. 9–28.
[89] I. Habernal, K.M. Konopí, SWSNL: Semantic web search using natural language, Expert Syst. Appl. 40 (9) (2013) 3649–3664.
[90] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920.
[91] R. Zhang, M.J. Cairelli, M. Fiszman, G. Rosemblat, H. Kilicoglu, T.C. Rindflesch, et al., Using semantic predications to uncover drug–drug interactions in clinical data, J. Biomed. Inform. 49 (2014) 134–147.
[92] A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, et al., Identifying potential adverse effects using the web: A new approach to medical hypothesis generation, J. Biomed. Inform. 44 (6) (2011) 989–996.
[93] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, et al., Biomedical text mining and its applications in cancer research, J. Biomed. Inform. 46 (2) (2013) 200–211.
[94] L.M. De Bruijn, A. Hasman, J.W. Arends, Automatic SNOMED classification — a corpus-based method, Comput. Methods Programs Biomed. 54 (1) (1997) 115–122.
[95] R.A.A. Seoud, M.S. Mabrouk, TMT-HCC: A tool for text mining the biomedical literature for Hepatocellular Carcinoma (HCC) biomarkers identification, Comput. Methods Programs Biomed. 112 (3) (2013) 640–648.
[96] J. Urbain, Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models, J. Biomed. Inform. 58 (2015) S143–S149.
[97] U. Leser, J. Hakenberg, What makes a gene name? Named entity recognition in the biomedical literature, Brief. Bioinform. 6 (2005) 4.
[98] K. Shaalan, Rule-based approach in Arabic natural language processing, Int. J. Inf. Commun. Technol. 3 (3) (2010) 11–19.
[99] S. Sarawagi, Information extraction, Found. Trends Databases 1 (3) (2008) 261–377.
[100] B. Jagan, S. Thenmalar, T.V. Geetha, Semi-supervised bootstrapping approach for named entity recognition, CoRR (2015). abs/1511.06833.
[101] Z. Guanming, Z. Chuang, X. Bo, L. Zhiqing, CRFS-based Chinese named entity recognition with improved tag set, in: 2009 WRI World Congress on Computer Science and Information Engineering, 2009.
[102] A. Ekbal, S. Saha, A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies, Expert Syst. Appl. 38 (12) (2011) 14760–14772.
[103] D. Küçük, A. Yazıcı, A hybrid named entity recognizer for Turkish, Expert Syst. Appl. 39 (3) (2012) 2733–2742.
[104] K. Shaalan, H. Raza, NERA: Named entity recognition for Arabic, J. Am. Soc. Inf. Sci. Technol. 60 (8) (2009) 1652–1663.
[105] K. Riaz, Rule-Based named entity recognition in Urdu, in: Proceedings of 2010 Named Entities Workshop, Association for Computational Linguistics, 2010, pp. 126–135.
[106] V. Gupta, G.S. Lehal, Named entity recognition for Punjabi language text summarization, Int. J. Comput. Appl. 33 (3) (2011) 28–32.
[107] U. Singh, V. Goyal, G.S. Lehal, Named entity recognition system for Urdu, in: COLING, 2012, pp. 2507–2518.
[108] B. Gődény, Rule based product name recognition and disambiguation, in: 2012 IEEE 12th International Conference on Data Mining Workshops, 2012.
[109] W. Zaghouani, RENAR: A rule-based arabic named entity recognition system, ACM Trans. Asian Lang. Inf. Process. 11 (1) (2012) 2.
[110] Arabic Corpus Search Tool. Online available at http://arabicorpus.byu.edu/.
[111] Benajiba's ANerCorp. Online available at http://www.dsic.upv.es/~ybenajiba.
[112] LingPipe ToolKit. Online available at http://alias-i.com/lingpipe/.
[113] R. Alfred, L.C. Leong, C.K. On, P. Anthony, Malay named entity recognition based on rule-based approach, Int. J. Mach. Learn. Comput. 4 (3) (2014) 300.
[114] Malay articles. Online available at http://www.bernama.com/bernama/v7/bm/.

[115] Malay articles. Online available at http://www.mstar.com.my/.
[116] K.R. Rahem, N. Omar, Rule-based named entity recognition for drug-related crime news documents, J. Theoret. Appl. Inf. Technol. 77 (2) (2015).
[117] Y. Benajiba, M. Diab, P. Rosso, Arabic named entity recognition: A feature-driven study, IEEE Trans. Audio Speech Lang. Process. 17 (5) (2009) 926–934.
[118] C. Lee, P.M. Ryu, H. Kim, Named entity recognition using a modified pegasos algorithm, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011, pp. 2337–2340.
[119] S.K. Saha, P. Mitra, S. Sarkar, A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition, Knowl.-Based Syst. 27 (2012) 322–332.
[120] N. Freire, J. Borbinha, P. Calado, An approach for named entity recognition in poorly structured data, in: Extended Semantic Web Conference, Springer, Berlin, Heidelberg, 2012, pp. 718–732.
[121] J. Nothman, N. Ringland, W. Radford, T. Murphy, J.R. Curran, Learning multilingual named entity recognition from Wikipedia, Artificial Intelligence 194 (2013) 151–175.
[122] E. Fersini, E. Messina, G. Felici, D. Roth, Soft-constrained inference for named entity recognition, Inf. Process. Manag. 50 (5) (2014) 807–819.
[123] S.B. Bam, T.B. Shahi, Named entity recognition for Nepali text using support vector machines, in: Intelligent Information Management, 2014.
[124] Y. Chen, T.A. Lasko, Q. Mei, J.C. Denny, H. Xu, A study of active learning methods for named entity recognition in clinical text, J. Biomed. Inform. 58 (2015) 11–18.
[125] I. Korkontzelos, D. Piliouras, A.W. Dowsey, S. Ananiadou, Boosting drug named entity recognition using an aggregate classifier, Artif. Intell. Med. 65 (2) (2015) 145–153.
[126] E. Yan, Y. Zhu, Identifying entities from scientific publications: A comparison of vocabulary-and model-based methods, J. Informetr. 9 (3) (2015) 455–465.
[127] M. Konkol, M. Konopík, Segment representations in named entity recognition, in: International Conference on Text, Speech, and Dialogue, Springer International Publishing, 2015, pp. 61–70.
[128] A. Kaur, G.S. Josan, Evaluation of named entity features for Punjabi language, Procedia Comput. Sci. 46 (2015) 159–166.
[129] B. Bhasuran, G. Murugesan, S. Abdulkadhar, J. Natarajan, Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases, J. Biomed. Inform. 64 (2016) 1–9.
[130] A.S. Wibawa, A. Purwarianti, Indonesian named-entity recognition for 15 classes using ensemble supervised learning, Procedia Comput. Sci. 81 (2016) 221–228.
[131] J. Guo, G. Xu, X. Cheng, H. Li, Named entity recognition in query, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2009, pp. 267–274.
[132] A. Ekbal, S. Saha, D. Singh, Active machine learning technique for named entity recognition, in: Proceedings of International Conference on Advances in Computing, Communications and Informatics, ICACCI, ACM, 2012, pp. 180–186.
[133] D. Küçük, Automatic compilation of language resources for named entity recognition in Turkish by utilizing Wikipedia article titles, Comput. Stand. Interfaces 41 (2015) 1–9.
[134] S. Zhang, N. Elhadad, Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts, J. Biomed. Inform. 46 (6) (2013) 1088–1098.
[135] M. Konkol, T. Brychcín, M. Konopík, Latent semantics in named entity recognition, Expert Syst. Appl. 42 (7) (2015) 3470–3479.
[136] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based N-gram models of natural language, Comput. Linguist. 18 (4) (1992) 467–479.
[137] T. Joachims, T. Finley, C.N.J. Yu, Cutting-plane training of structural SVMs, Mach. Learn. 77 (1) (2009) 27–59.
[138] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, 2012.
[139] S.K. Saha, P. Mitra, S. Sarkar, Word clustering and word selection based feature reduction for maxent based Hindi NER, in: ACL, 2008, pp. 488–495.
[140] C. Biemann, Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems, in: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, Association for Computational Linguistics, 2006, pp. 73–80.
[141] Indian languages annotated dataset. Online available at http://ltrc.iiit.ac.in/ner-ssea-08.
[142] A.L. Berger, V.J.D. Pietra, S.A.D. Pietra, A maximum entropy approach to natural language processing, Comput. Linguist. 22 (1) (1996) 39–71.
[143] European Collections. Online available at: http://www.europeana.eu/.
[144] D. Balasuriya, N. Ringland, J. Nothman, T. Murphy, J.R. Curran, Named entity recognition in Wikipedia, in: Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, Association for Computational Linguistics, 2009, pp. 10–18.
[145] K. Crammer, Y. Singer, Ultraconservative online algorithms for multiclass problems, J. Mach. Learn. Res. 3 (2003) 951–991.
[146] A. Ritter, S. Clark, O. Etzioni, Named entity recognition in tweets: An experimental study, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 1524–1534.
[147] X. Liu, M. Zhou, F. Wei, Z. Fu, X. Zhou, Joint inference of named entity recognition and normalization for tweets, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, 1, Association for Computational Linguistics, 2012, pp. 526–535.
[148] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, et al., Analysis of named entity recognition and linking for tweets, Inf. Process. Manag. 51 (2) (2015) 32–49.
[149] M. Konkol, M. Konopík, CRF-based Czech named entity recognizer and consolidation of Czech NER research, in: International Conference on Text, Speech and Dialogue, Springer, Berlin, Heidelberg, 2013, pp. 153–160.
[150] M. Ševčíková, Z. Žabokrtský, O. Krůza, Named entities in Czech: annotating data and developing NE tagger, in: International Conference on Text, Speech and Dialogue, Springer, Berlin, Heidelberg, 2007, pp. 188–195.
[151] R.I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: A resource for disease name recognition and concept normalization, J. Biomed. Inform. 47 (2014) 1–10.
[152] C.H. Wei, Y. Peng, R. Leaman, A.P. Davis, C.J. Mattingly, J. Li, et al. Overview of the biocreative V chemical disease relation (CDR) task, in: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 2015, pp. 154–166.
[153] R.M. Karp, M.O. Rabin, Efficient randomized pattern-matching algorithms, IBM J. Res. Dev. 31 (2) (1987) 249–260.
[154] R.S. Boyer, J.S. Moore, A fast string searching algorithm, Commun. ACM 20 (10) (1977) 762–772.
[155] G. Washington, The George Washington Papers, Dodd, Mead, 1964.
[156] Queensland State Archives. Australia-4113. Online available at: http://www.archivessearch.qld.gov.au/Search/BasicSearch.aspx.
[157] U.-V. Marti, H. Bunke, The IAM-database: An English sentence database for off-line handwriting recognition, Int. J. Doc. Anal. Recognit. 5 (2002) 39–46.
[158] M. Paşca, Weakly-supervised discovery of named entities using web search queries, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, 2007, pp. 683–690.
[159] A. Ekbal, S. Bandyopadhyay, A web-based Bengali news corpus for named entity recognition, Lang. Resour. Eval. 42 (2) (2008) 173–182.
[160] Ö. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 I2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (5) (2011) 552–556.
[161] J.D. Kim, T. Ohta, Y. Tateisi, J.I. Tsujii, GENIA corpus — a semantically annotated corpus for bio-text mining, Bioinformatics 19 (1) (2003) i180–i182.
[162] C.B.A.K. Lund, Modelling parsing constraints with high-dimensional context space, Lang. Cogn. Process. 12 (2–3) (1997) 177–210.
[163] D.L. Rohde, L.M. Gonnerman, D.C. Plaut, An improved method for deriving word meaning from lexical co-occurrence, Cogn. Psychol. 7 (2004) 573–605.
[164] M. Sahlgren, An introduction to random indexing, in: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE, 2005, p. 5.
[165] M.N. Jones, D.J. Mewhort, Representing word meaning and order information in a composite holographic lexicon, Psychol. Rev. 114 (1) (2007) 1.
[166] A. Purandare, T. Pedersen, Word sense discrimination by clustering contexts in vector and similarity spaces, in: CoNLL, 4, 2004. pp. 41–48.
[167] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[168] T. Brychcín, M. Konopík, HPS: High Precision Stemmer, Inform. Process. Manag. 51 (1) (2015) 68–91.
[169] J. Etkinson, V. Bull, A multi-strategy approach to biological named entity recognition, Expert Syst. Appl. 39 (17) (2012) 12968–12974.
[170] Biology datasets. Online available at http://www.biocreative.org/tasks/biocreative-i/first-task-gm/.
[171] D. Küçük, A. Yazici, Rule-based named entity recognition from turkish texts, in: Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, 2009.
[172] R. Mihalcea, D.I. Moldovan, Document indexing using named entities, Stud. Inform. Control 10 (1) (2001).
[173] D. Küçük, A. Yazici, A text-based fully automated architecture for the semantic annotation and retrieval of Turkish news videos, in: IEEE International Conference on Fuzzy Systems, FUZZ, IEEE, 2010, pp. 1–8.
[174] A. Ekbal, S. Saha, U.K. Sikdar, Multiobjective optimization for biomedical named entity recognition and classification, Procedia Technol. 6 (2012) 206–213.
[175] K. Deb, A. Pratap, S. Agarwal, T.A.M.T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.
[176] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML, 1, 2001, pp. 282–289.
[177] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, 2009, pp. 147–155.

[178] S. Saha, A. Ekbal, Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition, Data Knowl. Eng. 85 (2013) 15–39.
[179] L. Li, W. Fan, D. Huang, A two-phase bio-NER system based on integrated classifiers and multiagent strategy, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (4) (2013) 897–904.
[180] P. Jaimai, O. Chimeddorj, Part of speech tagging for Mongolian corpus, in: Proceedings of the 7th Workshop on Asian Language Resources, Association for Computational Linguistics, 2009, pp. 103–106.
[181] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, 1989, p. 102.
[182] Y. Tsuruoka, J. Tsujii, Training a Naive Bayes classifier via the EM algorithm with a class distribution constraint, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 4, Association for Computational Linguistics, 2003, pp. 127–134.
[183] G. Escudero, L. Arquez, G. Rigau, Naive Bayes and exemplar-based approaches to word sense disambiguation revisited, in: Proceedings of the 14th European Conference on Artificial Intelligence, 2000.
[184] D.D. Lewis, Naive Bayes at forty: The independence assumption in information retrieval, in: Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.
[185] A. McCallum, K. Nigam, A comparison of event models for Naive Bayes text classification, in: AAAI-98 Workshop on Learning for Text Categorization, 1998.
[186] T. Pedersen, A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation, in: Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, 2000.
[187] T. Joachims, Making Large Scale SVM Learning Practical, MIT Press, Cambridge, 1999, pp. 169–184.
[188] A.J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Trans. Inform. Theory 13 (2) (1967) 260–267.
[189] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[190] A. McCallum, D. Freitag, F. Pereira, Maximum entropy Markov Models for Information Extraction and Segmentation, in: Proceedings of the ICML, 2000, pp. 591–598.
[191] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J.M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, Comput. Stand. Interfaces 35 (5) (2013) 482–489.