Natural Language Processing Journal 3 (2023) 100017

Contents lists available at ScienceDirect

Natural Language Processing Journal

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

A survey on Named Entity Recognition — datasets, tools, and methodologies

Basra Jehangir *, Saravanan Radhakrishnan, Rahul Agarwal

EA-Process Automation, Cognizant Technology Solutions, Chennai, 600119, Tamil Nadu, India

ARTICLE INFO

Keywords:
Natural language processing
Named Entity Recognition
Deep Learning
Convolutional Neural Network
Bidirectional Long Short Term Memory
Recurrent Neural Networks

ABSTRACT

Natural language processing (NLP) is crucial in the current processing of data because it takes into account many sources, formats, and purposes of data as well as information from various sectors of our economy, government, and private and public lives. We perform a variety of NLP operations on the text in order to complete certain tasks. One of them is NER (Named Entity Recognition). An act of recognizing and categorizing named entities that are presented in a text document is known as named entity recognition. The purpose of NER is to find references of rigid designators in the text which belong to established semantic kinds like a person, place, organization, etc. It acts as a cornerstone for many Information Extraction-related activities. In this work, we present a thorough analysis of several methodologies for NER ranging from unsupervised learning, rule-based, supervised learning, and various Deep Learning based approaches. We examine the most relevant datasets, tools, and deep learning approaches like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Bidirectional Long Short Term Memory, Transfer learning approaches, and numerous other approaches currently being used in present-day NER problem environments and their applications. Finally, we outline the difficulties NER systems encounter and future directions.

1. Introduction

Currently, the data is typically presented in its raw form (unstructured, native language, confusing), from various sectors of our economy, government, and private and public lives so summarizing, searching, drawing conclusions, and doing statistical analysis are all challenging tasks for humans. We perform various NLP operations on the text to complete the abovementioned tasks. This makes NLP crucial in today's data processing. One of the operations is Named Entity Recognition. The task of recognizing and categorizing named entities that are presented in a text document is known as named entity recognition.

In 1997, the phrase "Named Entity Recognition" was first used by Chinchor and Robinson (1997) in a 6th Message Understanding Conference (MUC) conference to identify all expressions. NER has a vast range of applications in various diverse domains. A few of them are:

- Social media (Kim et al., 2022) generates a huge amount of data daily. Microblogs are becoming increasingly popular, generating a lot of user-generated data. Each day, Twitter produces 500 million tweets. The insights are concealed in the less organized forms of social media posts because social media writings do not strictly adhere to syntactic principles. Thus, NER is a crucial step for identifying appropriate entities in texts and offering assistance for following NLP activities. Applications such as sentiment analysis are performed using various models such as BERT (Nemes and Kiss, 2021) (Bidirectional encoder representations from transformers).
- Domain-specific NER such as Biomedical NER (Veysel and David, 2022), in which we try to identify biological elements in provided texts, such as chemicals, illnesses, and proteins. The requirement for obtaining biomedical knowledge stored in texts that are not structured and converting them into structured representations is the ability to recognize biological items effectively (Song et al., 2018).
- Information Retrieval by Petkova and Croft (2007) is retrieving important information from textual documents. Obtaining detailed information from retained data which is structured or unstructured, and analyzing it is one of the main application areas of NER. For this purpose, several NER-based approaches such as Bidirectional LSTM-CNN, Rule-based approach, and conditional random fields can be used (Aliwy et al., 2021).
- Semantic annotations (Rahman and Bowles, 2020) in which documents are tagged with pertinent ideas through the process of semantic annotation.

* Correspondence to: EA-Process Automation, Cognizant Technology Solutions, Cognizant Technology Solutions, Plot No-20 & 21, Sea view Developers Ltd, Building No-10, INFOSPACE, Sector 135, 201304 Bajidpur, Uttar Pradesh, India.
E-mail addresses: Basra.Jehangir@cognizant.com (B. Jehangir), Saravanan.Radhakrishnan@cognizant.com (S. Radhakrishnan), Rahul.Agarwal4@cognizant.com (R. Agarwal).

https://doi.org/10.1016/j.nlp.2023.100017
Received 28 February 2023; Received in revised form 17 May 2023; Accepted 19 May 2023

- Machine Translation (Yuval Marton, 2014), the process of automatically changing text from one natural language to another while retaining the text's meaning and creating good writing in the target language without human intervention. Using neural machine translation techniques based on several artificial neural network models, such as Long Short-Term Memory (Lee et al., 2021), has become prevalent in performing machine translation.
- Question and Answering systems (Raju et al., 2012) is where questions are asked, and answers are given automatically. It involves intelligently searching through various text documents to determine a response to a particular English-language query.
- Text summarization (Mukesh and Varun, 2022) has shown to be a fantastic resource for giving readers pertinent information in comparatively short amounts of time. Using selected phrases from a text to summarize it has both benefits and drawbacks. The advantage is that, despite the straightforward procedure, the summaries it generates will always be syntactically valid, even if they are not particularly effective. The drawback of extractive summarizers is that they can only anticipate so much from the original material.

In terms of NER methodologies, there are several technologies. The most common of them are

- Rule-Based approach
- Supervised approach
- Unsupervised approach

In this work, we have presented a detailed explanation of each of the datasets, tools, and detection methodologies with more focus on Deep Learning methodologies and, finally, the challenges. The timeline of the research articles used in this work spans from the initial usage of the technology to present-day novel techniques proposed in various domains such as biomedical NER, General purpose, News, and social media. These approaches are discussed further in detail.

2. Definition

Named Entity Recognition is the process of identifying numerous segments of information referenced in a text and then classifying them into pre-established categories. Entities like "person", "organization", "region", and many more might be considered categories. Named Entity Recognition is a broad category of NLP issues known as sequence tagging. Other sequences tagging tasks of NLP outside NER include chunking and Part of Speech tagging. When a model is used for sequence tagging, it is anticipated to produce labels for each word that appears in the sequence of words or tokens. Question-answering, information retrieval, machine translation, and all other applications all gain from the pre-processing that Named Entity Recognition provides. A NER system is given an order of tuples (s = ⟨w1⟩, ⟨w2⟩, , ⟨wN⟩); it will output a set of labels L1, L2, L3 and their corresponding tuples for each mentioned in the input text. Fig. 1 demonstrates a case where a system of NER correctly identifies two named things from a sentence.

3. Datasets

Data acquisition is a step of acquiring data online for model training and evaluation. Since annotations of the utmost quality are essential for model development and validation. We include popular datasets and their respective usage in various publications. We have established a common benchmark dataset for comparative analysis to objectively compare the capabilities of novel approaches. The research articles that have already utilized different models for NER and have been published in recent years are listed below. Each paper's summary is arranged to aid scholars in better comprehension and information retention. We have covered various study-related topics, including datasets, tools, pre-processing, traditional methods, deep learning models, and more.

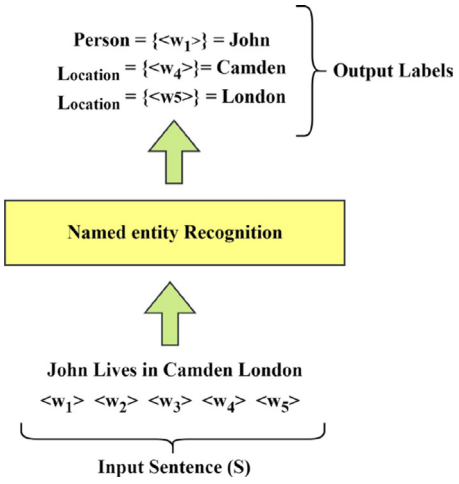

Fig. 1. Instancing of NER.

A corpus is a collection of authentic writing or sounds laid out into datasets. Here, the word "authentic" refers to text that has been put down or a voice that has been expressed by a person who is a common speaker of the dialect or language. Newspapers, recipes, books, television shows, motion pictures, radio broadcasts, and tweets can all be part of a corpus. The datasets available online are diverse depending on the usage and application area. Detailed description of each of the datasets is given in this section.

3.1. CoNLL datasets

- **CoNLL-2002 Dataset** The six files that make up the CoNLL-2002 named entity (NE) data "Tjong Kim Sang (2002)" are divided into two languages: Dutch and Spanish. The dataset focuses on four categories of named entities: people, organizations, places, and names of other entities that fail to fit into the former categories. Each language has a development file, training file, and testing file. The training data is utilized to train the algorithms. The learning methods' parameters can be adjusted using development data. Once the finest parameters have been identified, the model can be trained over the training data and further tested over the test data. To prevent systems from becoming tuned to the test data, development data, and test data have been divided.
- **CoNLL-2003 Dataset** There are eight items total within CoNLL-2003 NE data (Tjong Kim Sang and De Meulder, 2003), and they include two different languages: German and English. The dataset focuses on several categories of named entities: people, places, organizations, and names of other entities that fail to fit into the former categories. Each of these languages has a development file, a training file, and a testing file, along with big, unannotated data. The functionality of the former three files is the same as in the CoNLL-2002 dataset. However, the difficult part of this collaborative work was finding a means to use the unannotated data within the learning process.

3.2. WNUT-2017

The dataset by Derczynski et al. (2017) focuses on categories of named entities: location (including GPE, facility), product, corporation, person, creative work (movie, song, books, and many more), and group. The collection includes comments that were longer than 140 words because these reflect distinct writing styles and characteristics and because Twitter is a good source of noisy user-generated data. Twitter is also used as a resource to align part of the test and development data with training data, but extra remarks were also mined from Reddit,

Table 1
Datasets and their corresponding domains.

S.No.	Dataset	Type	Link	Scale (Datasize)			
				Train set (MB)	[Test set] (MB)	Dev. set (MB)	Total (MB)
1	CoNLL-2002	Multilingual	https://www.clips.uantwerpen.be/conll2002/ner/	–	–	–	14.8
2	CoNLL-2003	Multilingual	https://www.clips.uantwerpen.be/conll2003/ner/	–	–	–	16.7
3	WNUT-2017	Social-Media	http://noisy-text.github.io/2022/	–	–	–	1.74
4	OntoNotes	Multilingual	https://catalog.ldc.upenn.edu/LDC2013T19	–	–	–	947.4
5	NCBI Disease	Biomedical	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/	0.938	0.171	0.162	–
6	BioCreative	Biomedical	https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/	0.2461	0.2650	0.2840	–
7	Genia Corpus	Biomedical	http://www.geniaproject.org/home	1.75	0.491	0.366	–
8	Wiki Gold	Wikipedia	https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500	–	–	–	102
9	Wiki Coref	Wikipedia	http://rali.iro.umontreal.ca/rali/?q=en/wikicoref	–	–	–	172
10	Hyena	Wikipedia	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena	–	–	–	–
11	BC4CHEMD	Biomedical	https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/	–	–	–	855
12	JNLPBA	Biomedical	http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004	9.44	3.11	–	–
13	BC2GM	Biomedical	https://biocreative.bioinformatics.udel.edu/accounts/login/?next=/resources/corpora/biocreative-ii-corpus/	3.55	1.35	–	–

Stack Exchange, and YouTube. Because of their size and the ability to mine samples along several dimensions, such as texts from and about geo-specific places, and specific topics and events, these sources were chosen.

3.3. OntoNotes

The aim of this dataset (Weischedel et al., 2013) was always to annotate a large textual corpus in 3 languages which are Arabic, English, and Chinese, that includes a variety of text genres (weblogs, telephone speech, talk shows broadcast, news, newsgroups), with information that is structural (predicate-argument structure along with syntax) and shallow semantics. It has a 2.9-million-word count in all three languages, as mentioned before.

3.4. NCBI disease corpus

The NCBI corpus by Doğan et al. (2014) is a domain-specific dataset. The public release dataset contains 6892 mentions of diseases that are assigned to 790 different illness categories. 88% of them contain an OMIM (Online Mendelian Inheritance in Man) identity, and the remaining relate to a MeSH (Medical Subject Headings) identification. The collection is divided into a development file, a training file, and a testing file (see Table 2).

3.5. BioCreative V chemical disease relation

The dataset (Wei et al., 2016) comprises Chemical and related articles on diseases. The dataset comprises 500 for the development, training, and test sets for 1500 PubMed items. Of the 1500 publications, 1400 were chosen from an existing dataset associated with the CTD-Pfizer partnership. The other 100 items, which constituted an entirely new selection, were added to the test set.

3.6. Genia corpus

Prof. Jun'ichi Tsujii established the GENIA Project housed at the University of Tokyo's Tsujii Laboratory from 1998 until 2012. The main repository of literature consisting of biomedical assembled and annotated for the GENIA project is the GENIA corpus by Kim et al. (2003). The corpus was developed to aid in the creation and assessment of systems for the extraction of text mining and information for the field of molecular biology. Utilizing a PubMed search for 3 MeSH terms, "blood cells", "transcription factors", and "human", the corpus' 1999 Medline abstracts were chosen. Different language and semantic layers of data have been annotated into the corpus.

3.7. Wikipedia dataset

Several Wikipedia datasets are employed for NER. In this work, we have studied a few datasets that have used Wikipedia as the source.

- **WikiGold:** The Wikipedia Gold corpus (Joel et al., 2013) is made by annotating 39 000 words manually from English Wikipedia along with coarse-grained NE tags.
- **WikiCoref:** (Ghaddar and Langlais, 2016) It is a corpus of papers in English that has been annotated for anaphoric relationships, all taken from the English Wikipedia. With a few exceptions, the annotation format is similar to OntoNotes'. The researchers annotated each markable with the corresponding Freebase subject, coreference type, and mention type. Few tools are available for overused Wikipedia texts since similar annotation projects tend to focus on relatively specialized written content, like newswires.
- **HYENA:** In the dataset (Yosef et al., 2012), around million entities are mentioned in the 50,000 randomly chosen Wikipedia pages that make up the collection. 92% of the relevant entities are members of at least one of our top-level kinds, and 11% are members of two or more. Ten thousand randomly chosen Wikipedia articles from a different edition of the same Wikipedia make up the testing data, separate from the data used for training.

Apart from the datasets mentioned before, there are several other diverse datasets depending on the domain they are used in. Some of them are BC4CHEMD (Krallinger et al., 2015), JNLPBA (Collier et al., 2004), and BC2GM (Smith et al., 2008). These datasets are arranged in a tabular form in Table 1 along with their corresponding links and domains.

4. Tools used for NER

In this section, we have discussed various off-the-shelf tools that are utilized for named entity recognition.

4.1. SpaCy

SpaCy (Honnibal and Montani, 2017) is a package in Python utilized for high-level NLP and is open source. It helps us generate programs that "understand" and process large amounts of text as it was created specifically for utilization in production environments. Systems for extracting information or natural language comprehension can be produced using it. It supports over 72 languages and 80 trained pipelines for almost 24 languages. It also encourages multitask learning with transformers like BERT. It has SOTA (State of the art) speed. It also supports custom models in Tensorflow, Pytorch, and other libraries. Named entity recognition has a pipeline component called entity recognizer.

The NER module (Vychegzhanin and Kotelnikov, 2019) is built on a model that makes use of residual convolutional neural networks and Bloom embeddings. There are models for the English language that were previously trained on the OntoNotes Release 5.0 package, including tagged text taken from news articles, blogs, and phone conversations.

4.2. Natural Language Tool Kit (NLTK)

The well-known platform for making applications in Python that use human linguistic data is called NLTK (Bird et al., 2009). In addition to a compilation of text processing libraries for parsing, categorization, tagging, stemming, tokenization, and semantic reasoning, named entity recognition and an active discussion forum, it is frequently employed when performing research and training students. The NER module employs an Automatic Content Extraction corpustrained Maximum Entropy Classifier.

4.3. Apache openNLP

A machine learning (ML) established toolset (Nanavati and Ghodasara, 2015) for processing NL text is the Apache OpenNLP library. Parsing, sentence segmentation, tokenization, chunking, part-of-speech (POS) tagging, coreference resolution, and named entity extraction are only a few of the most popular NLP activities that are supported. In order to provide more sophisticated text processing solutions, several activities are often necessary. Perceptron-based and Maximum entropy machine learning are also included in OpenNLP. Both named items and numbers may be found in text with the Name Finder. The Name Finder requires a model in order to recognize entities. The model relies on the language and entity class for which it was developed.

4.4. Tensorflow

Tensorflow (Abadi et al., 2016) is an open-source math library. Python, C++, and CUDA are three languages in which TensorFlow is written. Google created it for use with neural network models or other machine learning applications. It is a data mining learning technique, and instead of using text data, it accepts input in numbers or hot encoding. Google Translate, text summarization, and NER are just some of the numerous applications that utilize Tensorflow.

4.5. Pytorch

Facebook created Pytorch (Paszke et al., 2019), another open-source deep-learning package designed purely for Python use. It serves as a premier platform for both industry and education purposes. Pytorch can be mostly used for image recognition and language processing by making machine learning more scalable several AI models can be built quickly and efficiently.

Some of the other tools offered for industry-based projects are LingPipe, AllenNLP (Gardner et al., 2018), IBM Watson (Ferrucci, 2012), Intellexer (Intellexer), ParallelDots (Jain et al., 2019), and Dandelion API (SpazioDati).

5. Evaluation

In their evaluation of NER performance, Grishman and Sundheim (1996) used two criteria: type, which measured whether the predicted label was accurate regardless of entity borders, and text, which measured whether the predicted entity boundaries remained accurate regardless of label. For each score category, precision was defined as the proportion of correctly predicted entities by the system divided by the total number of entities predicted by the system, recall as the proportion of correctly predicted entities by the system divided by the total number of entities identified by the human annotators, and (micro) F-score as the harmonic mean of precision and recall regarding both type and text.

Only when the projected label for the entire predicted word matches the label's precise wording does the exact match metrics deem a prediction to be accurate (Segura-Bedmar et al., 2013). If a portion of the predicted entity is accurately recognized, Relaxed F1 deems the prediction to be correct. Strict F1 demands that a prediction's character offsets match perfectly with the input annotation's.

For example the words "in London" and "on August 8" this can be predicted as "London" and "August 8" respectively which is also acceptable as important data is extracted by the system.

The mathematical interpretation of the metrices mentioned can be done by using True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These are utilized for the calculation of precision (represented by Eq. (1)), F1-score (represented by Eq. (3)), and Recall (represented by Eq. (2)). These evaluation metrics constitute the exact-match evaluation which means the exactness of the model to identify its type and boundary simultaneously.

$$Precision = \frac{J}{J + L} \tag{1}$$

$$Recall = \frac{J}{J + M} \tag{2}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

where J = TP, K = TN, L = FP, M = FN. True Positive (TP) means an entity which a NER system offers and which is present in the real data. True Negative (TN) means an entity which a NER system does not offer and which is present in the real data. False Positive (FP) means an entity which a NER system offers and which is absent from the real data. False Negative (FN) means an entity which a NER system does not offer and is absent from the real data.

6. NER using rule based approach

Rule-based strategies include a collection of rules which specialists have carefully designed. These rules are accurate and are based on language, syntactic-lexical, and domain-specific expertise. These rules are necessarily designed to be domain-specific. The reason for this is for the network to be more accurate and more efficient. Rule-based approaches are not highly generalized because these rules are restricted to a specific domain only. These systems do not transcend domain boundaries. Additionally, these systems require human intervention to design the rules and require higher programming abilities.

Regular expressions may be used to match complex strings using basic or complicated patterns. It may be used to discover and retrieve patterns and swap out matching patterns in a string for other patterns. Using the same approach Korkontzelos et al. (2015) examined strategies for getting excellent drug name recognition performance in situations with little or no standard training data available. The technique uses a voting mechanism that may aggregate predictions from various recognizers. Additionally, string-similarity established on frequent suffixes of single-token medication names found in the Drug Bank database is evolved using genetic programming. In order to broaden the coverage and tag accuracy of dictionary items, these patterns are then utilized to

construct regular expressions (Regex). In work (Eftimov et al., 2017), the NER method is proposed for extracting dietary information based on evidence which is the first of its kind. The first step is finding and identifying the entities mentioned, and the next entails selecting and obtaining the entities. The approach is tested using text corpora from various sources like scientific websites and research publications.

A rule-based approach can also be used with other techniques as a two-step process. Any rule-based approach does the initial step of extracting the data, and further classification is done by either a supervised, semi-supervised, or un-supervised algorithm Sari et al. (2010) In order to seamlessly create pattern extraction resulting out of a small corpus and identify the already defined entities in a set of accident records, the study provides a Rule-based extraction of patterns and a Semi-Supervised Named Entity Recognition technique. Stanford Part-of-Speech (POS) tagger and Grammar Parser are used. The extractor uses a speech tagger to locate the designated object and create the pattern extraction. In order to classify the entities into certain specified categories, the pattern extraction is further fed to a semi-supervised NER. Exact Match assessment is used to gauge performance on the two distinct entities of DATE and LOCATION.

There are several other rule-based NER systems, such as FATUS (Hobbs et al., 1993) which, instead of comprehensive text interpretation, information extraction tasks are more suitable. In other words, it works best for text-scanning tasks when just a small portion of the text is significant, the information must be mapped into a pre-defined, straightforward target representation, and the writer's intended audience is irrelevant. LaSIE (Humphreys et al., 1998), which stands for Large Scale Information Extraction system, LTG or Language technology group (Mikheev et al., 1998), is used for information extraction in various domains. The tools consist of capabilities for text annotation (tokenizers, lemmatizers, and taggers), data collection tools, and general utilities and FACILE (Black et al., 1998), which categorizes the text into several languages such as German, English, Spanish, and Italian.

Though these rules are highly accurate and are established on language, syntactic-lexical, and domain-specific expertise, they may have certain drawbacks. These rules only apply to a single domain; hence they are not generalizable. These systems do not transcend domain boundaries. These systems also need more advanced programming skills and human involvement to develop the rules.

7. NER using unsupervised learning

Unsupervised learning approaches are used for the data that lack labeling. The common approaches for unsupervised learning are association and clustering. In a social network NER, to address the issue of noisy, insufficient information and short tweets, the work by Liu and Zhou (2013) provides a unique technique that uses tweet redundancy by doing two-stage named entity recognition for several similar tweets. Specifically, it pre-labels every tweet with a sequential labeler built on the linear Conditional Random Fields model first. After then, tweets are clustered to gather tweets having comparable content together. In the end, for every cluster, it fine-tunes the labels of each tweet using an improved Conditional Random Field model that considers the cluster-level data, i.e., the labels of the current word and its nearby terms across all tweets in the cluster.

In biomedical NER (Śniegula et al., 2019), the method by Zhang and Elhadad (2013) is easily adaptable to other semantic categories and text genres since it does not depend on manually created instances or rules of entities that are annotated. Entity recognition relies on terminology, shallow syntactic understanding, and corpus statistics rather than supervision. According to experimental results, the technique produces good results on two well-known datasets of clinical notes, distinct genres, biomedical data, and varied related entity types.

Peng et al. (2021) in their work addressed problems such as domain distribution shift and dependency on external resources and suggested a cross-domain model unsupervised for NER utilizing adversarial training with entity-aware attention. To alleviate the incorrect alignment of entity characteristics throughout the training, adversarial training is proposed to decrease domain distribution shift. The model was evaluated over several datasets, and the experimental finding demonstrates that the suggested model surpasses other recent approaches.

The study by Dutta and Gupta (2022) investigates how noise affects the unsupervised entity ranking with names by initially gathering a naming list with content-based features using NER operating systems that come pre-installed, features extraction for both the entities that are identified, and then ranking them with a brand-new unsupervised Kernel Density Estimation (KDE) centered ranking algorithm.

A significant and prominent research topic is the development of unsupervised learning techniques for NLP applications. These techniques' key highlight is their ability to build models only from unlabeled data; this is a significant benefit because labeled datasets are typically expensive to create, whereas the unlabeled text in digital form is widely available (Vlachos, 2011).

The assessment of unsupervised learning methods is more difficult than their supervised counterparts due to the benefit of needing just unlabeled input to train a model. The main reason for this is that unsupervised techniques' output lacks labels that would be present in a manually created gold standard. Defined, "no labels" for model learning refers to the absence of labels in the result. As a result, it is impossible to evaluate unsupervised algorithms using the conventional evaluation paradigm of benchmarking against a gold standard using a performance metric like accuracy or F-score.

8. NER using supervised learning

Supervised learning approaches are used for the data that is labeled. The machines are trained over the labeled data. Then either prediction are made, or data is classified based on the problem statement. For NER systems, choosing the right learning algorithm is crucial. The most popular approaches for doing NLP using supervised learning are typically the Hidden Markov Model (HMM) (Zhou and Su, 2002), Conditional Random Field (CRF) (Patil et al., 2020), and Support Vector Machines (SVM) (Singh et al., 2009). For models to complete the work asked of them, feature templates are needed. Even though these feature templates have a high degree of accuracy, they could be more generalizable because NLP is used on a diverse range of text. In the biomedical domain, Zhang et al. (2004) performed HMM-based NER for cascaded and abbreviation NER. The suggested system incorporates several characteristics (morphology, head noun trigger features, orthography, and part-of-speech) and techniques (including two cascaded NER and an abbreviation recognition algorithm method). The experiment demonstrates that the system performs noticeably better than the prior best methods.

For the categorization and identification of the named entities included in the Marathi language text by Patil et al. (2020), a statistical NER system using machine learning as a basis was developed. The system uses conditional random fields to identify and categorize named items. Since Marathi is a morphologically rich language, a machine learning method is used for the identification as well as categorization of named entities that are present in the language. Statistical techniques produce accurate NE identification and classification results. Using CRF, named entities in the system are located and post that categorized. Statistical methods successfully identify and classify NE in this morphologically complex language, although further information might improve accuracy. Studies on the FIRE-2010 corpus demonstrate that more knowledge may improve accuracy.

The authors of the study Lee et al. (2004) suggested a double-phase NER technique based on Support Vector Machines that comprises a phase for identifying NE boundaries and a phase for semantic categorization, which comprises selecting a certain classification technique and the features important to each subtask using the two-phase NER approach. Additionally, this resolves the issue of imbalanced class distribution that a discriminative learner like SVM frequently encounters.

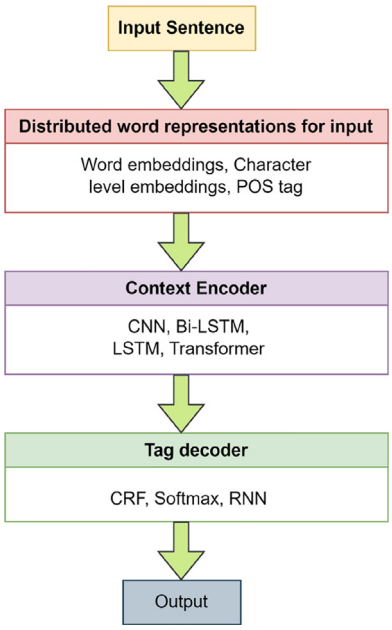

Fig. 2. NER using Deep Learning.

Additionally, by recommending a hierarchical semantic classification approach based on ontology, performance and a significant reduction in computing complexity are improved.

The researchers Saha and Ekbal (2013) hypothesize that each classifier's prediction accuracy varies depending on the output classes. Therefore, in an ensemble method, either the eligible classes for which a classifier is best suited to vote (i.e., binary vote-based ensemble) or to quantify the quantity of voting for each class in a specific classifier are required (i.e., real vote-based ensemble). Researchers build a range of models using seven different models, including Decision Tree (DT), Naive Bayes, HMM, Maximum Entropy (ME), Memory Based Learner (MBL), SVM, and CRF, depending on the different feature representations that are available and are identified and selected without using any language-specific resources or domain knowledge.

8.1. NER using deep learning

The initial work in deep learning for NER was done by Collobert and Weston (2008) first used Deep Neural Network for performing the NER techniques on the text data. In their work, they defined a simple convolutional network architecture for performing Natural Language Processing tasks like NER, semantic role labeling, parts of speech tagging, and chunking. These tasks were, for the first time, done automatically without hand-extracted features by a deep neural network. Fig. 2 shows the process to be followed when using deep learning in NER.

Mostly used methods for NER with deep learning are recurrent neural networks such as BiLSTM, convolutional Neural networks, their combination (hybrid) with each other, and combination with other supervised learning algorithms like SVM, Bootstrapped neural networks, and many more. A thorough study of many of them is as below. In this section, we discuss and summarize various basic Deep Learning architectures that are used in NLP.

8.1.1. Text representations

These are representations wherein terms from a bigger corpus or corpora are indicated by their associated indices to their place in a dictionary. The word representations usually used are One-Hot encoding, Count Vectorizer, TF-IDF, Word2vec methods like skip gram, and the second Continuous Bag of Words (CBOW).

- One Hot Encoding in NER uses the Automatic Construction of Training Data, for example, from social media messaging applications (Lee and Ko, 2020). It is very easy to understand and use. However, it cannot prioritize the words based on importance and can result in the formation of a sparse matrix that is computationally and memory expensive.
- The Count Vectorizer determines how frequently a term appears in a document. It creates a matrix of reviews and words from the corpus of sentences, filling with the frequency of every term within every phrase and condensing a sentence into a single vector. TF-IDF stands for term-frequency times inverse document-frequency (IDF) (Upendraa and Sudheer, 2016), which is a product of 2 factors, TF and IDF, mathematically represented by Eqs. (4) and (5):

$$TFIDF = TF(w,k) \times IDF \tag{4}$$

TF is the frequency of the word (w) in a document (k)

$$IDF = \log(\frac{N}{df(w)}) \tag{5}$$

Here N is the number of documents and df is the frequency of documents containing word (w)

- Word2Vec is a class of shallow, two-layer neural networks (NN) trained to recover word contexts from linguistic data. A huge corpus of words is used to retrieve pairs of context/target words for training, where the target word is context words, and the randomly selected terms fall inside a specified window surrounding the target word. In this work, we study two word2vec methods specified above as CBOW (Carreras et al., 2003) and Skip Grams (Hsieh et al., 2017). In the case of a continuous bag of words, the context words are input into the model in their embedding form (first initialized randomly), and the model is anticipated to produce the target word utilizing softmax. While skip-gram performs the exact opposite, producing the context words together with the target word is required.

8.1.2. Convolutional neural networks

The Convolutional Neural Network (Alzubaidi et al., 2021) is a type of widely accepted neural network (NN) design that has been extensively researched for handling the named entity relationships classification. CNN extracts the features by performing a convolutional operation on the embedding input data. The embedding input representation is expressed as s = w1, w2, w3, …, wn given a sentence = e1, e2, e3, …, en. Typically, CNNs include a collection of convolutional filters. To create a feature based on m-grams, each convolutional filter executes convolution operations to l continuous words specified by the weight matrix. A convolution filter produces the following feature (ci) for a given window of words can be seen in (6):

$$c_i = f(W.e_{i:i+m-1} + b) \tag{6}$$

where W is used for weight and b stands for the bias.

Originally used in image processing, CNNs are also applied to NER tasks. To efficiently create high-level combinatorial feature embedding in biomedical NER, the proposed model by Cho et al. (2020) combines convolutional NN and bidirectional – Long Short Term Memory for wordlevel and character-level representations. The token format includes both the global and local characteristics at the character level of the embedding by combining character-level features from the convolutional neural networks and bidirectional-LSTM. In order to learn the combination of each representation, the representations — word-level CNN, character-level, and character-level bidirectional-LSTM are then chained and input into a fully connected (FC) network. In order to focus on tokens that are related in the sentence and forecast the tag of the token that is current, the work also uses an attention technique to the bidirectional-long short-term memory-CRF model to calculate the resemblance between the tokens that are input. The model was further validated over two standardized datasets, NCBI Disease, and JNLPBA.

Other variations of CNN, which are Dilated Convolutional Neural Networks, gated CNNs, and recurrent neural networks (RNN), are also utilized in several NER applications. For example, in biomedical NER Zhang et al. (2022) used BiLSTM and dilated convolutional neural network (DCNN) for hierarchical encoding and used DCNN's benefits to record massive amounts of data efficiently. To enhance the performance of medical NER, several feature words are simultaneously put into the medical text data. On three real-world datasets, extensive tests have been conducted that show their approach to be better than the comparison models. Similarly, the work (Wang et al., 2020) proposed an adversarial trained LSTM -convolutional neural networks (ASTRAL) system to perform the NER techniques on unstructured data. In their work, a gated CNN was introduced into the NER task; this enhanced the feature extraction process. This is followed by adversarial training, in which perturbations are added to arbitrary variables to create the ability of generalization in the model. This is then constrained by the target variable so that it can be applied to any variable. Finally, the work is evaluated on benchmarks, OntoNotes, CoNLL-03, and WNUT-17, attaining SOTA results.

CNN, which is used in obtaining text-level representations, can extract features inside sentences and pays attention to the sequential link among sentences; this has demonstrated this Chang and Han (2023) using a 3D convolutional neural network for document-level feature extraction instead of a BiLSTM. Since a BiLSTM has an incapability to identify multiple sentences at once, making it impossible to obtain the complete information. This makes it necessary for a CNN to be employed. Additionally, they research a layer- by-layer residual structure to optimize each Bi-directional LSTM block of the model, which can address the deterioration issue that arises as the count of model layers rises.

Out of vocabulary problem, which usually occurs in the neural machine translation (NMT) problem, is addressed in the model used by Na et al. (2019). By handling the OOV (Out of vocabulary) problem, compositional techniques for Korean NER problems were addressed. The work developed a unique hybrid representation incorporating compositional morpheme vectors based on LSTM and ConvNet (Convolutional Neural Network). The input character vectors are first individually subjected to Long Short Term Memory and ConvNet-based compositions, and the resultant compositional morpheme vectors are then concatenated to provide the hybrid representation of morpheme.

Data obtained from an OCR (Optical character recognition) is often used for NER purposes. Zhou et al. (2020b) proposed a NN architecture for the information extraction from packages of a drug. The model utilizes the data obtained from an OCR as a data source. The model comprises of three different layers. The first layer is used to correct data as a pre-processing step consisting of the language model and seq2seq model. The next layer has a CNN, and finally, the last layer determines the sentence present in the text. On the evaluation of the model, it was found that it had 4%–6% more F1-Score.

As mentioned earlier CNNs are originally used in image processing, and local structure is important in images because nearby pixels often carry meaningful semantic information. Nevertheless, words are frequently linked in sentences without necessarily being close to each other.

8.1.3. Recurrent neural networks

RNNs are NN that use internal memory to store lengthy input data. The vanishing gradient problem mainly affects the basic RNN model, making it unable to represent long-term relationships in real-world situations accurately. LSTMs (Long Short Term Memory) and gated recurrent units (GRUs) have now been developed to address this.

- **LSTM**:
  Regarding NER, LSTMs have emerged as the most widely used model (Gasmi et al., 2018) and have attained cutting-edge performance. Its important feature is the gate mechanism of the LSTM model (Hochreiter and Schmidhuber, 1997), this includes the input gate, forget gate, and output gate and learns the long-term dependence. Based on the present input word representation n_t, the previously hidden state $h_{(t-1)}$, and the prior memory cell $c_{(t-1)}$, the presently hidden state h_t and memory cell c_t at time step t are created. The extracted feature vector $\tilde{c}_t$, the forget gate f_t, the output gate O_t, and the input gate i_t are defined in Eqs. (7)–(12) as:

$$i_t = Sigmoid(W_i n_t + U_i h_{(t-1)} + b_i) \tag{7}$$

$$f_t = Sigmoid(W_f n_t + U_f h_{(t-1)} + b_f) \tag{8}$$

$$O_t = Sigmoid(W_o n_t + U_o h_{(t-1)} + b_o) \tag{9}$$

$$\tilde{c}_t = \tanh(W_c n_t + U_c h_{(t-1)} + b_c) \tag{10}$$

where W and U are weights and b is the bias. The hidden state h_t and current state c_t are calculated as

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{11}$$

$$h_t = O_t \odot \tanh(c_t) \tag{12}$$

In work by Jin et al. (2019), a new Gated Convolutional Recurrent Neural Network with Attention (GCRA) for chinese NER job is character-based. In particular, a hybrid CNN with a gating filter mechanism to collect local context information and a highway neural network following LSTM to pick characters of interest are introduced. An extra gated self-attention method is utilized to collect the global dependencies of numerous sub-spaces and random nearby characters.

- **Bi-LSTM:**
  Bi-LSTMs are frequently employed in NLP applications to extract contextual characteristics from the input text in both the backward and forward directions. Let and stand for the outputs of the forward and backward LSTM. The concatenation is Bi-Long Short-Term Memory's output. The feature representation is preceded by a fully connected layer (FCL), much like in the CNN model, and a softmax function is employed to carry out the final classification. For a given period, bidirectional RNNs effectively use past data (through forward states) and future data (through backward states). One of the initial works that utilized Bi-LSTM-CRF for sequence tagging application is by Huang et al. (2015), which proposes various models, including combinations of LSTM, CRF, and Bi-LSTM, and gave some SOTA results.
  Xu et al. (2019) Proposed a NN and named it Dic-Att-BiLSTM-CRF. The method applies a string-matching mechanism to match the entities with a dictionary of diseases. It constructs a document-level attention mechanism and dictionary-matching method. Then a Bi-LSTM network and CRF combine the disease dictionary to develop a NER. On evaluation, it was found that DABLC outperformed state-of-the-art methods by accomplishing the highest F1 scores (NCBI: F1 = 88.6%, Recall = 89%, Precision = 88.3%; BioCreative V CDR: F1 = 88.3%, Recall = 87.5%, Precision = 89.1%). Gajendran et al. (2020) suggested a unique neural network architecture that consists of a dynamic recurrent neural network (DRNN), CRF, and bidirectional LSTM that employs character- and word-level embeddings as the sole characteristics to identify chemical substances. To capture the character and word representation, the author first used a 3-level BLSTM-DRNN model, then the LSTM-CRF model.
  Many diverse datasets require named entity recognition to be done on them, such as the work Rizou et al. (2022), which is performed over the famous ATIS, which stands for the Airline Travel Information Systems dataset. The researchers compared the outcomes of experiments that were carried out to solve the IC (Item categorization) and NER tasks and assessed numerous previously used state-of-the-art architectures. A Bidirectional-LSTM network is used as the foundation of the network for the NER

task, and the authors investigate modifications of the model by experimenting with various character and word embedding. An SVM architecture is used for the independent IC work. The aforementioned tasks are addressed for the second set of tests under the assumption that they are highly reliant on one another, meaning that the NE in a sentence is closely related to its intent and that the intent is significantly determined by things that make up the sentence. Transformer- based architectures, more notably BERT, are recommended for this joint model.

Similarly, work by Zhou et al. (2020a) proposed a deep neural network consisting of a Bi-LSTM with CRF decoding algorithm for bug-specific entity recognition, also called DBNER. The data employed in the work was extracted from two source application software projects, Apache and Mozella, and System software, such as the kernel. The proposed method extracts feature from the data containing information on bugs and utilize the attention method to enhance the quality of tag tags. The model gave an F1-score of 91.19%. Multi-lingual data such as Jin and Yu (2021) used the best sequence labeling model, Bi-LSTM-CRF, which is also the foundation of the model proposed in work. Introducing a morphological-level NE tagger and a masked self-attention mechanism for extracting the contextual text and lexical texts from the Korean sentence was performed.

Other applications, such as in Biomedical NER (An et al., 2022), to increase the diversity and specificity of feature representations, provide a better character-level representation strategy that incorporates character embedding and character label embedding. Next, the MUSA-BiLSTM-CRF model for multi-head self-attention-based bi-directional LSTM is presented. The algorithm can more accurately obtain the weight association between multi-level semantic feature information and characters by combining a medical dictionary and multi-head self-attention, which is anticipated to significantly increase the performance of Chinese clinical NER. The model was tested by using two standard datasets from the CCKS challenge. Similarly, Nath et al. (2022) proposed NEAR, which stands for a Named entity, and attributes the recognition of clinical concepts, which works with three models to contribute to the multi-label tagging problem. For that purpose, three models were designed BiLSTM n-CRF-TF, BiLSTM n-CRF, and BiLSTM-CRF-Smax- TF. These models were evaluated over two shared datasets, i2b2 2012 and the 2010 i2b2/VA; the models got the best scores of 80.8% and 90.3% over the two datasets, respectively.

Because long sentences only have one latent vector as an output and the final LSTM unit might not fully capture the meaning of the phrase, LSTM Networks or LSTM-based Encoder–Decoder models may perform poorly for long sentences. Since a single vector represents every word of the lengthy sentence, the simple LSTM-based encoder–decoder model must give each word the attention it deserves. Hence, the Attention Based Model was created to address these problems.

8.1.4. Transfer learning

Transfer learning is another method that can be applied when we have a relatively smaller dataset. It is a tool that is yet to be explored completely. Particularly language models are frequently utilized for transfer learning via pre-training. Their significance may be due to some factors: first, these are unsupervised models which train on raw textual data, which is a plentiful resource, and second, they train to build an intuitive knowledge of syntax and semantics. On enormous corpora such as the complete English Wikipedia, models may pre-train. After having been trained, language models such as Elmo (Affi and Latiri, 2021) and BERT's (Devlin et al., 2018) network topologies may be readily modified for tasks of classification of text by (Li et al., 2021b) like emotion detection through text classification (Li et al., 2023), Information Extraction (Li et al., 2021c; Wei et al., 2019), sequence

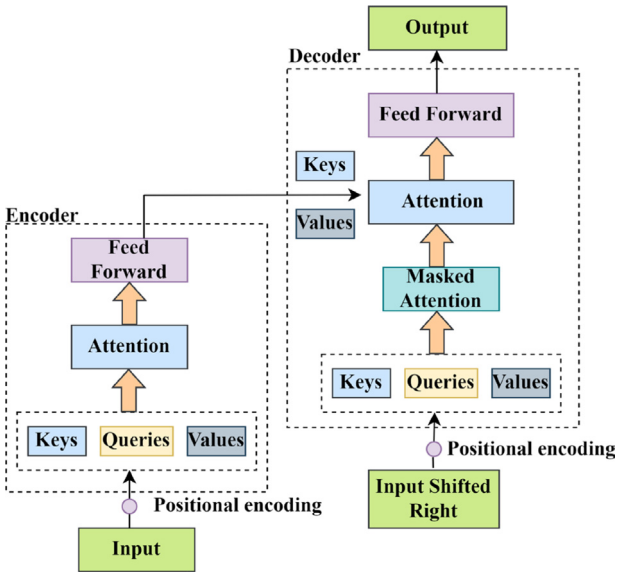

Fig. 3. Architecture of a transformer.

tagging activities such as named entity identification, generative tasks such as abstractive summarization, spam detection, and other types of tasks (see Table 2).

Transformers, a deep learning model where each output element is linked to each input element and the weightings among them are dynamically determined relying upon their correlation, is the foundation for BERT, which stands for Bidirectional Encoder Representations from Transformers.

The Transformer architecture is built on an attention mechanism that enables the model to generate predictions by focusing on particular segments of the input sequence. This contrasts with conventional sequence models that analyze input sequences sequentially and need help with long-term reliance, such as RNNs. Encoders and decoders make up the architecture. As the decoder creates the output sequence, the encoder analyses the input sequence. A multi-head self-attention mechanism and a feed-forward network are both sublayers of identical layers that comprise the encoder and decoder. Fig. 3 depicts the architectural representation of a transformer similar to one in the original paper Vaswani et al. (2017).

In biomedical NER, Sun et al. (2021) performed BERT over biomedical NER in a machine-reading compression framework. The work was performed over six datasets, namely BC4CHEMD, JNLPBA, NCBI-Disease, BC2GM, BC5CDR-Disease, and BC5CDR-Chem datasets. The work achieved SOTA performance on all the datasets. A comparative analysis of different models such as BioBERT-Softmax, BioBERT-MRC, BioBERT-BiLSTM-CRF, and BioBERT-CRF was performed in which the proposed method gave performed exceptionally. The evaluation scores used in the work were Recall, F1-Score, and Precision.

Several hybridized works containing BERT and Bi-LSTM, such as by Liu et al. (2021), used several techniques for complex biochemical NER. The hybrid approach includes BERT, BILSTM, Multi-Head Attention (MHATT), and Conditional Random Fields (CRF). The accomplishment of this hybrid deep neural network model is evaluated over a publicly accessible dataset, and the four fundamental methodologies mentioned above are effectively combined to meet the various complex parts of NER. A more effective representation model is utilized in place of the widely used word2vec library, which creates a word vector, to train on high-level abstract data and handle polysemy (BERT). The multi-head attention process from cognitive neuroscience is creatively consolidated into the Bi-directional LSTM model to extract chapter-level properties. The softmax deep learning results are coupled to the CRF layer to increase the recognition rate by leveraging the reliance between tags.

Table 2
Comparative analysis of works that have used the NCBI dataset.

S.No.	Work	Year	Algorithm used	F1-score (%)
1	Fries et al. (2017)	2017	Swell Shark	64.2
2	Beltagy et al. (2019)	2019	SciBERT	88.57
3	Neumann et al. (2019)	2019	ScispaCy	81.65
4	Xu et al. (2019)	2019	Dic-Att-BiLSTM-CRF (DABLC)	88.6
5	Sun et al. (2021)	2020	BioBERT-MRC	90.04
6	Gajendran et al. (2020)	2020	Bi-LSTM-DRNN	90.84
7	Cho et al. (2020)	2020	CNN-LSTM	86.93
8	Fan et al. (2022)	2022	CNN-LSTM	86.93

Compared with recognition accuracies for lesser frequency entities, this hybrid approach achieved when applied to two entity datasets (IDENTIFIER and MULTIPLE) gave an improvement of 80% and 21.69%, respectively. Similarly, Li et al. (2021a) designed a model for bridge inspection NER. The suggested methodology enhances context understandability by concurrently identifying nested and flat items in reports of Chinese bridge inspection and integrating existing domain knowledge through sample words in inquiry inquiries. The model includes lexical data and bigram embedding via a Bi-directional LSTM-based fea- ture fusion component in addition to the fundamental features, which are character-level encoded by the Chinese BERT from transformers pre-trained model, which enhances the power of contextual feature modeling. The corpus is built using extensive genuine bridge inspection information. The outcome of the experiments demonstrates that the recommended model performs surpasses the popular flat and nested NER models.

Other transfer learning methods, such as one used in Mehmood et al. (2020), a unique method that makes use of transfer learning in order to enhance the performance of both the single task model (STM) and multi-task model (MTM). Thereafter, researchers modified the fine-tuned strategy by optimizing the pre-trained multi-task model for a selected target dataset. Furthermore, modification of the fine-tuned strategy entails training the multi-task model over a range of epochs before using it to establish the weights of the Single Task Model for a specific dataset.

Transfer learning is preferred when we have a relatively smaller dataset. Transfer learning aims to improve the model's generalizability and minimize the training labeled data for the target task.

Transfer learning (Liu et al., 2019) can use the data from the source domain to supplement the data from the target domain, but during the transfer process, misclassification may result in a negative transfer, meaning that the knowledge learned in the source domain has a negative impact on learning on the target domain.

9. Challenges

- **Data Annotations:** Huge amounts of annotated training data are necessary for supervised NER systems, particularly DL-based NER. Annotating data is still time and money-consuming. Since domain specialists are required to complete annotation activities, it is a significant barrier for many languages and particular fields that need more resources (Jie et al., 2019).
  For example, to create relevant features in multilinguistic annotations, sufficient linguistic knowledge is essential for every single language of interest. As a result, creating multilingual NER annotators is timeconsuming, expensive, and difficult.
- **Complex Biomedical Texts:** In biomedical literature, lengthy and complex sentences are common. Sometimes, two complex entities may occasionally appear in two separate clauses. Accurately recog- nizing and classifying relevant biomedical relations inside a complex biomedical sentence is a significant difficulty in categorizing biomedical relations. It has been observed that the convoluted and unclear naming convention is a contributing factor in complications (Leaman et al., 2015). Such as in the sentence "Exon-intron structure of the human neuronal nicotinic acetylcholine receptor alpha 4 subunit (CHRNA4)" from the NCBI dataset mentioned in previous sections. Use of acronyms such as "recent tension PTX at OSH."
- **Ambuigity in language:** Natural language words can have many interpretations, making it challenging to determine if a term corresponds to a designated thing or otherwise. Example the sentence "He saw a bat," this simple sentence can have four distinct meanings because of the word "saw" and "bat." A saw can have two meanings past tense of the word see and cutting, and a bat can have two meanings a bird or sports equipment. As a result, the sentence can be interpreted in four different ways.
- **Informal texts:** Because of their conciseness and noise, NER over informal text (such as comments, tweets, and user forums) (Leaman et al., 2015) presents more of a challenge than formal text. NER systems must manipulate the user-generated text in various application settings, including customer care in e- commerce and banking. They have a weak structure in sentences, making it challenging to locate the named entity using generic characteristics.
- **Multilingual NER:** The increase in internet users worldwide, bringing many cultures, people, and languages together, makes it diverse linguistically. Nevertheless, more research must be done to address the growing language diversity of web material. Natural Language Processing techniques are constrained to few languages, typically just English, which does not keep up with the Internet's rapid rate of change. Accordingly, current multilingual text-based IR systems are limited to basic processing stages that utilize word frequency-based techniques and surface forms.
- **Domain Adaptation:** Cross-domain NER is a difficult but doable problem. Across domains, entity mentions can differ greatly. For instance, in botany, "orange" can refer to both a fruit and a color. Likewise, distinct types of entities can differ in their similarity to other types of entities in other disciplines. Diseases, for instance, are treated differently in medicine and biochemistry.
- **Named entity linking:** One of the challenges in NER is entity linking. An entity may have different meanings and is referred to differently in a document. For example, "Barcelona" can be referred to as a football team and a city in the same document.
- **Entity co-reference resolution:** Determining whether entities mentioned in a text or dialogue refer to the same real-world entity is known as entity coreference resolution.
- **Handling noisy and misspelled text:** Most of the documents have clarity issues example, the unclear scanned documents that are read from OCR because of some OCR issue misspelling is common alphabets that look alike are wrongly detected and result in misspelling. Human errors can also cause misspellings in the texts.

10. Future directions

- Since each model or approach has its methodological benefits and drawbacks, combining various models or approaches may

produce superior outcomes. In the classification of relationships, some hybrid models have demonstrated the possibility of merging CNNs with RNNs. Since RNNs cannot identify multiple sentences at once, it is impossible to obtain complete information. CNN, which is often used to extract text-level representations, not only has the ability to extract features inside sentences but pays attention to the sequential link among these sentences. Some of the works (Chang and Han, 2023; Wang et al., 2020; Zhang et al., 2022) have shown great potential in the hybridization of the previously mentioned models.

- Creating models that recognize named items in languages for which they have not been explicitly trained is the goal of zero-shot cross-lingual NER (Eronen et al., 2023). With this method, the requirement for labeled data in low-resource languages may be greatly reduced, and NER could be effectively used in languages with little training data.
- New models and technologies could be combined with neural network-based methods to increase efficiency even more. Deep contextualized word representations, like ELMo (Affi and Latiri, 2021) and BERT's (Li et al., 2022), network topologies may be readily mod- ified for tasks of classification like sequence tagging activities such as named entity identification, productive tasks such as abstractive summarization, spam detection, and other types of tasks. Many other transfer learning approaches can also be utilized by integrating the existing models or as standalone for better results.
- Transfer learning is preferred when we have a relatively smaller dataset. Transfer learning aims to improve the model's generalizability and minimize the training labeled data for the target task.

11. Conclusion

In addition to producing scenario templates and relationship identification, named entity recognition is crucial for additional Information Extraction activities, like ontology population, semantic annotation, and opinion mining, to mention a few. We present a quick rundown of conventional methods, modern state-of-the-art, and challenges. Initially, we briefly introduced Named Entity Recognition, its beginning, definition, and application areas. Then we described the methodologies that are present for Named Entity Recognition. Starting with the Rule-based approach and how they are not highly generalizable because these rules are restricted to specific domains only, require human intervention to design the rules, and require higher programming abilities making them complex to implement. Then we move on to other methodologies, such as supervised and unsupervised learning and deep learning-based methodologies. Studying the advantages of several deep learning-based architectures over the former mentioned. These advantages include feature engineering, less implementation complexity, less human intervention, and faster execution. Finally, based on the literature review, we evaluated the current challenges and future directions. This review was done to aid further research and is a useful resource when developing deep learning-based Named Entity Recognition models.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Basra Jehangir reports a relationship with Cognizant Technology Solutions Pvt ltd that includes: employment. Saravanan Radhakrishnan reports a relationship with Cognizant Technology Solutions Pvt ltd that includes: employment. Rahul Agarwal reports a relationship with Cognizant Technology Solutions Pvt ltd that includes: employment.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning. In: Osdi, Savannah, GA, USA. pp. 265–283.

Affi, M., Latiri, C., 2021. Be-blc: Bert-elmo-based deep neural network architecture for english named entity recognition task. Procedia Comput. Sci. 192, 168–181.

Aliwy, A., Abbas, A., Alkhayyat, A., 2021. Nerws: Towards improving information retrieval of digital library management system using named entity recognition and word sense. Big Data Cogn. Comput. 5, 59.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L., 2021. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. J. Big Data 8, 1–74.

An, Y., Xia, X., Chen, X., Wu, F.X., Wang, J., 2022. Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. Artif. Intell. Med. 127, 102282.

Beltagy, I., Lo, K., Cohan, A., 2019. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.

Black, W.J., Rinaldi, F., Mowatt, D., 1998. Facile: Description of the ne system used for muc-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.

Carreras, X., Màrquez, L., Padró, L., 2003. A simple named entity extractor using adaboost. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 152–155.

Chang, J., Han, X., 2023. Multi-level context features extraction for named entity recognition. Comput. Speech Lang. 77, 101412.

Chinchor, N., Robinson, P., 1997. Muc-7 named entity task definition. In: Proceedings of the 7th Conference on Message Understanding. pp. 1–21.

Cho, M., Ha, J., Park, C., Park, S., 2020. Combinatorial feature embedding based on cnn and lstm for biomedical named entity recognition. J. Biomed. Inform. 103, 103381.

Collier, N., Ohta, T., Tsuruoka, Y., Tateisi, Y., Kim, J.D., 2004. Introduction to the bio-entity recognition task at JNLPBA. pp. 73–78, URL: https://aclanthology.org/W04-1213.

Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. pp. 160–167.

Derczynski, L., Nichols, E., van Erp, M., Limsopatham, N., 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. http://dx.doi.org/10.18653/v1/W17-4418, URL: https://aclanthology.org/W17-4418.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Doğan, R., Leaman, R., Lu, Z., 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. J. Biomed. Inform. 1–10, 47.

Dutta, H., Gupta, A., 2022. Pnrank: Unsupervised ranking of person name entities from noisy ocr text. Decis. Support Syst. 152, 113662.

Eftimov, T., Korošić Seljak, B., Korošec, P., 2017. A rule-based named- entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLoS One 12, e0179488.

Eronen, J., Ptaszynski, M., Masui, F., 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. Inf. Process. Manage. 60, 103250.

Fan, S., Yu, H., Cai, X., Geng, Y., Li, G., Xu, W., Wang, X., Yang, Y., 2022. Multi-attention deep neural network fusing character and word embedding for clinical and biomedical concept extraction. Inform. Sci. 608, 778–793.

Ferrucci, D.A., 2012. Introduction to this is watson. IBM J. Res. Dev. 56, 1.

Fries, J., Wu, S., Ratner, A., Ré, C., 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. arXiv preprint arXiv:1704.06360.

Gajendran, S., Manjula, D., Sugumaran, V., 2020. Character level and word level embedding with bidirectional lstm–dynamic recurrent neural network for biomedical named entity recognition from literature. J. Biomed. Inform. 112, 103609.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., Zettlemoyer, L., 2018. Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640.

Gasmi, H., Bouras, A., Laval, J., 2018. Lstm recurrent neural networks for cybersecurity named entity recognition. ICSEA 11, 2018.

Ghaddar, A., Langlais, P., 2016. WikiCoref: An english coreference- annotated corpus of wikipedia articles. pp. 136–142, URL: https://aclanthology.org/L16-1021.

Grishman, R., Sundheim, B.M., 1996. Message understanding conference- 6: A brief history. In: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

Hobbs, J.R., Appelt, D., Bear, J., Israel, D., Kameyalna, M., Tyson, M., 1993. Fastus: A system for extracting information from text. In: Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Honnibal, M., Montani, I., 2017. Spacy. *Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing*.

Hsieh, J.T., Li, C., Liu, W., 2017. Effective word representation for named entity recognition.

Huang, Z., Xu, W., Yu, K., 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cun-ningham, H., Wilks, Y., 1998. University of sheffield: Description of the lasie-ii system as used for muc-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.

Intellexer, https://www.intellexer.com/. Accessed: 2023-02-21.

Jain, A., Aggarwal, I., Singh, A., 2019. Paralleldots at semeval-2019 task 3: Do-main adaptation with feature embeddings for contextual emotion analysis. In: Proceedings of the 13th International Workshop on Seman- Tic Evaluation. pp. 185–189.

Jie, Z., Xie, P., Lu, W., Ding, R., Li, L., 2019. Better modeling of incomplete annotations for named entity recognition. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 729–734.

Jin, Y., Xie, J., Guo, W., Luo, C., Wu, D., Wang, R., 2019. Lstm-crf neural network with gated self attention for chinese ner. IEEE Access 7, 136694–136703.

Jin, G., Yu, Z., 2021. A korean named entity recognition method using bi-lstm-crf and masked self-attention. Comput. Speech Lang. 65, 101134.

Joel, N., Nicky, R., Will, R., Tara, M., James, R.C., 2013. Learning multi- lingual named entity recognition from wikipedia. Artificial Intelligence 194, 151–175.

Kim, J., Kim, Y., Kang, S., 2022. Weakly labeled data augmentation for social media named entity recognition. Expert Syst. Appl. 209, 118217.

Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J., 2003. Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19, i180– i182.

Korkontzelos, I., Piliouras, D., Dowsey, A.W., Ananiadou, S., 2015. Boosting drug named entity recognition using an aggregate classifier. Artif. Intell. Med. 65, 145–153.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., et al., 2015. The chemdner corpus of chemicals and drugs and its annotation principles. J. Cheminform. 7, 1–17.

Leaman, R., Khare, R., Lu, Z., 2015. Challenges in clinical natural language processing for automated disorder normalization. J. Biomed. Inform. 57, 28–37.

Lee, K.J., Hwang, Y.S., Kim, S., Rim, H.C., 2004. Biomedical named entity recognition using two-phase model based on svms. J. Biomed. Inform. 37, 436–447.

Lee, S., Ko, Y., 2020. Named-entity recognition using automatic construction of training data from social media messaging apps. IEEE Access 8, 222724–222732.

Lee, J., Lee, J., Lee, M., Jang, G.J., 2021. Named entity correction in neural machine translation using the attention alignment map. Appl. Sci. 11, 7026.

Li, W., Du, Y., Li, X., Chen, X., Xie, C., Li, H., Li, X., 2022. Ud_bbc: Named entity recognition in social network combined bert-bilstm-crf with active learning. Eng. Appl. Artif. Intell. 116, 105460.

Li, X., Li, Z., Xie, H., Li, Q., 2021b. Merging statistical feature via adaptive gate for improved text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 13288–13296.

Li, Z., Li, X., Xie, H., Wang, F.L., Leng, M., Li, Q., Tao, X., 2023. A novel dropout mechanism with label extension schema toward text emotion classification. Inf. Process. Manage. 60, 103173.

Li, X., Luo, X., Dong, C., Yang, D., Luan, B., He, Z., 2021c. Tdeer: An efficient translating decoding schema for joint extraction of entities and relations. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8055–8064.

Li, R., Mo, T., Yang, J., Li, D., Jiang, S., Wang, D., 2021a. Bridge inspection named entity recognition via bert and lexicon augmented machine reading comprehension neural model. Adv. Eng. Inform. 50, 101416.

Liu, J., Gao, L., Guo, S., Ding, R., Huang, X., Ye, L., Meng, Q., Nazari, A., Thiruvady, D., 2021. A hybrid deep-learning approach for complex biochemical named entity recognition. Knowl.-Based Syst. 221, 106958.

Liu, R., Shi, Y., Ji, C., Jia, M., 2019. A survey of sentiment analysis based on transfer learning. IEEE Access 7, 85401–85412.

Liu, X., Zhou, M., 2013. Two-stage ner for tweets with clustering. Inf. Process. Manage. 49, 264–273.

Mehmood, T., Gerevini, A.E., Lavelli, A., Serina, I., 2020. Combining multi-task learning with transfer learning for biomedical named entity recognition. Procedia Comput. Sci. 176, 848–857.

Mikheev, A., Grover, C., Moens, M., 1998. Description of the ltg system used for muc-7. In: Seventh Message Understanding Conference (MUC- 7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.

Mukesh, K.R., Varun, M., 2022. An exploratory study of automatic text summarization in biomedical and healthcare domain. Healthc. Anal. 2, 100058.

Na, S.H., Kim, H., Min, J., Kim, K., 2019. Improving lstm crfs using character-based compositions for korean named entity recognition. Comput. Speech Lang. 54, 106–121.

Nanavati, J., Ghodasara, Y., 2015. A comparative study of stanford nlp and apache open nlp in the view of pos tagging. Int. J. Soft Comput. Eng. 5, 57–60.

Nath, N., Lee, S.H., Lee, I., 2022. Near: Named entity and attribute recognition of clinical concepts. J. Biomed. Inform. 130, 104092.

Nemes, L., Kiss, A., 2021. Information extraction and named entity recognition supported social media sentiment analysis during the covid- 19 pandemic. Appl. Sci. 11, 11017.

Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Patil, N., Patil, A., Pawar, B., 2020. Named entity recognition using conditional random fields. Procedia Comput. Sci. 167, 1181–1188.

Peng, Q., Zheng, C., Cai, Y., Wang, T., Xie, H., Li, Q., 2021. Unsupervised cross-domain named entity recognition using entity-aware adversarial training. Neural Netw. 138, 68–77.

Petkova, D., Croft, W.B., 2007. Proximity-based document representation for named entity retrieval. Eur. Phys. J. B. 731–740, CIKM ' 07.

Rahman, F., Bowles, J., 2020. Semantic annotations in clinical guidelines. In: From Data to Models and Back: 9th International Symposium, Data- Mod 2020, Virtual Event, October 20, 2020, Revised Selected Papers. pp. 190–205.

Raju, B., Gulfishan2, F.A., Nepal, B., 2012. An approach for extracting exact answers to question answering (qa) system for english sentences. Procedia Eng. 30, 1187–1194.

Rizou, S., Paflioti, A., Theofilatos, A., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C., 2022. Multilingual name entity recognition and intent classification employing deep learning architectures. Simul. Model. Pract. Theory 120, 102620.

Saha, S., Ekbal, A., 2013. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. Data Knowl. Eng. 85, 15–39.

Sari, Y., Hassan, M.F., Zamin, N., 2010. Rule-based pattern extractor and named entity recognition: A hybrid approach. In: 2010 International Symposium on Information Technology. IEEE, pp. 563–568.

Segura-Bedmar, I., Martínez Fernández, P., Herrero Zazo, M., 2013. Semeval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (Ddiextraction 2013). Association for Computational Linguistics.

Singh, T.D., Nongmeikapam, K., Ekbal, A., Bandyopadhyay, S., 2009. Named entity recognition for manipuri using support vector machine. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2. pp. 811–818.

Smith, L., Tanabe, L.K., Kuo, C.J., Chung, I., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., Torii, M., et al., 2008. Overview of biocreative ii gene mention recognition. Genome Biol. 9, 1–19.

Śniegula, A., Poniszewska-Marańda, A., Chomątek, Ł., 2019. Study of named entity recognition methods in biomedical field. Procedia Comput. Sci. 160, 260–265.

Song, H.J., Byeong-Cheol, J., Park, C.Y., Kim, J.D., Kim, Y.S., 2018. Comparison of named entity recognition methodologies in biomedical documents. BioMed. Eng. Online 17, 158.

SpazioDati, 2023. Dandelion api. https://dandelion.eu/. Accessed: 2023-02- 21.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., Wang, J., 2021. Biomedical named entity recognition using bert in the machine reading comprehension framework. J. Biomed. Inform. 118, 103799.

Tjong Kim Sang, E.F., 2002. Introduction to the CoNLL-2002 shared. URL: https://aclanthology.org/W02-2024.

Tjong Kim Sang, E.F., De Meulder, F., 2003. Introduction to the CoNLL- 2003 shared task: Language-independent named entity recognition. pp. 142–147, URL: https://aclanthology.org/W03-0419.

Upendraa, B., Sudheer, B., 2016. Knn tfidf based named entity recognition. Int. J. Sci. Res. 1, 35–39.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Veysel, K., David, T., 2022. Accurate clinical and biomedical named entity recognition at scale. Softw. Impacts 13, 100373.

Vlachos, A., 2011. Evaluating unsupervised learning for natural language processing tasks. In: Proceedings of the First Workshop on Unsupervised Learning in NLP. pp. 35–42.

Vychegzhanin, S., Kotelnikov, E., 2019. Comparison of named entity recognition tools applied to news articles. In: 2019 Ivannikov Ispras Open Conference. ISPRAS, pp. 72–77.

Wang, J., Xu, W., Fu, X., Xu, G., Wu, Y., 2020. Astral: adversarial trained lstm-cnn for named entity recognition. Knowl.-Based Syst. 197, 105842.

Wei, C.H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C., Lu, Z., 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. In: Database 2016.

Wei, Z., Su, J., Wang, Y., Tian, Y., Chang, Y., 2019. A novel cascade binary tagging framework for relational triple extraction. arXiv preprint arXiv:1909.03227.

Weischedel, Ralph, et al., 2013. Ontonotes Release 5.0. Linguistic Data Consortium, Philadelphia.

Xu, K., Yang, Z., Kang, P., Wang, Q., Liu, W., 2019. Document-level attention-based bilstm-crf incorporating disease dictionary for disease named entity recognition. Comput. Biol. Med. 108, 122–132.

Yosef, M.A., Bauer, S., Hoffart, J., Spaniol, M., Weikum, G., 2012. Hyena: Hierarchical type classification for entity names. In: Proceedings of COLING 2012: Posters. pp. 1361–1370.

Yuval Marton, I.Z., 2014. Transliteration normalization for information extraction and machine translation. J. King Saud Univ. Comput. Inf. Sci. 26 (4), 379–387.

Zhang, S., Elhadad, N., 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. J. Biomed. Inform. 46, 1088–1098.

Zhang, J., Shen, D., Zhou, G., Su, J., Tan, C.L., 2004. Enhancing hmm- based biomedical named entity recognition by studying special phenomena. J. Biomed. Inform. 37, 411–422.

Zhang, R., Zhao, P., Guo, W., Wang, R., Lu, W., 2022. Medical named entity recognition based on dilated convolutional neural network. Cogn. Robot. 2, 13–20.

Zhou, R.G., Chang, S., Li, Y., 2020b. A neural network architecture for information extraction in chinese drug package insert. IEEE Access 8, 51256–51264.

Zhou, C., Li, B., Sun, X., 2020a. Improving software bug-specific named entity recognition with deep neural network. J. Syst. Softw. 165, 110572.

Zhou, G., Su, J., 2002. Named entity recognition using an hmm-based chunk tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 473–480.