International Journal of Computational Intelligence Systems (2024) 17:71
https://doi.org/10.1007/s44196-024-00456-1

REVIEW ARTICLE

Named Entity Recognition Datasets: A Classification Framework

Ying Zhang[1] · Gang Xiao[1]

Received: 2 August 2023 / Accepted: 12 March 2024

Abstract

Named entity recognition as a fundamental task plays a crucial role in accomplishing some of the tasks and applications in natural language processing. In the age of Internet information, as far as computer applications are concerned, a huge proportion of information is stored in structured and unstructured forms and used for language and text processing. Before neural networks were widely used in natural language processing tasks, research in the field of named entity recognition usually focused on leveraging lexical and syntactic knowledge to improve the performance of models or methods. To promote the development of named entity recognition, researchers have been creating named entity recognition datasets through conferences, projects, and competitions for many years, based on various research goals, and training entity recognition models with increasing accuracy on this basis. However, there has not been much exploration of named entity recognition datasets. Particularly, there have been many datasets available since the introduction of the named entity recognition task, but there is no clear framework to summarize the development of these seemingly independent datasets. A closer look at the context of the development of each dataset and the features it contains reveals that these datasets share some common features to varying degrees. In this thesis, we review the development of named entity recognition datasets over the years and describe them in terms of the language of the dataset, the domain of research, the type of entity, the granularity of the entity, and the annotation of the entity. Finally, we provide an idea for the creation of subsequent named entity recognition datasets.

Keywords Named entity recognition · Recognition dataset · Classification framework · Entity description

1 Introduction

Named Entity Recognition (NER) aims to identify names of entities in the text that resemble predefined categories such as names of people, Location, and organizations [1]. This concept has been widely used in the field of natural language processing since its introduction at the 6th Message Understanding Conference (MUC-6) [2]. As a core fundamental task in the field of natural language processing, the improvement of recognition accuracy plays an important role in the effectiveness of downstream task implementation. Specifically, named entity recognition is often used as the first step in tasks [3–5] such as information retrieval [6–8], question answering system [9, 10], machine translation [11], text understanding [12, 13], automatic text summarization [14, 15], relation extraction [16–18], and co-reference resolution [19, 20]. The promotion of named entity recognition thus makes a self-evidently significant contribution to the ongoing exploration of the field of natural language processing.

In the nearly 30 years since the development of Named Entity Recognition, both the creation of NER datasets and the comprehensive study of NER systems have undergone many changes. On the one hand, the increasing variety of research objectives has led to the design and creation of suitable NER datasets in response to developments. Currently, there are roughly ten well known conferences or projects that include named entity recognition tasks and whose proposed datasets are often used to train NER models. Examples include, in chronological order, MUC [2], MET [21], IREX [22], CoNLL [23], ACE [24], GENIA [25, 26], StemNet [27], OntoNotes [28], BioCreative V [29], WNUT [30], SemEval [31, 32]. In addition, there are several independently proposed NER datasets, such as the GENE- TAG dataset [33, 34], created to evaluate gene/protein annotators. The BBN dataset [35], which can provide a fine-grained

✉ Ying Zhang
yingzhang199608@foxmail.com

Gang Xiao
searchware@qq.com

[1] Institute of Systems Engineering, Academy of Military Sciences (AMS), Beijing 100107, China

Published online: 28 March 2024

entity annotation reference for general domain NER tasks. The WikiGold dataset [36] and the WiNER dataset [37] for training models to identify named entities in Wikipedia. NCBI-Disease dataset for identifying disease mentions in the biomedical domain [38]. N^3 is a corpus for named entity recognition and disambiguation [39]. SCI-ERC dataset for scientific information extraction [40]. And NNE, a fine-grained nested named entity recognition dataset based on the BBN dataset [41]. The CoNLL + + dataset was created to modify entity annotation errors in the test set to re-evaluate the NER system accurately [42]. The CrossNER dataset is a multi-domain dataset designed to facilitate NER adaptation [79]. The FEW-NERD dataset, as the first few-shot entities dataset [80], has been proposed to significantly advance named entity recognition techniques for these entities. RadGraph is a dataset for the medical field chest X-ray radiology reports dataset [81]. In addition, some researchers have created datasets that can be used for named entity recognition for their own research purposes. For example, Jain et al. [43] observed that there are no named entity recognition datasets for the art domain and therefore created a dataset for artwork recognition based on the extensive digitized art historical documents provided by the Wildenstein Plattner Institute (WPI). Similarly, Sahin et al. [44] found that the current datasets for named entity recognition and text classification tasks are mainly in English and very few in Turkish, and created the largest dataset available in Turkish for named entity recognition and text classification based on the reference to previous datasets. Fu et al. [45] created an automatically generated Chinese NER training dataset based on a bilingual parallel corpus to address the limitations of the development of Chinese named entity recognition due to data shortage and domain overfitting problems. As can be seen, as named entity recognition has evolved over the years, more and more datasets have been created for a variety of different purposes. The diversity of datasets makes it difficult to capture the patterns of their development and it is difficult to systematically provide research ideas for the development of future data sets. There is therefore an urgent need for a framework to collate work related to NER datasets to further provide a systematic description of the development of datasets over the years.

On the other hand, most researchers are keen to train models on existing standard datasets to achieve breakthroughs in the accuracy of the models. For example, for the CoNLL 2003 (English) dataset [23], Wang et al. [46] found through research related work that combining different types of embeddings in appropriate combinations could lead to better word representations and inspired by previous related work, proposed the Automatic Embedding Technique (ACE), which aims to automatically find better embedding connections for structured prediction tasks. For the ACE 2004 [47] and ACE 2005 [48] datasets, Zhong et al. [49] proposed a simple pipeline approach for entity and relationship extraction, built on two separate encoders, respectively. The entity model is built on a span level representation and the relationship model is built on a contextual representation specific to a given span pair. For the biological domain dataset GENIA [25, 26], version 3.0.2 of GENIA was used by Yu et al. [50], who proposed a Biaffine model aimed at reconstructing the NER task as a structured prediction task and using the Biaffine model to explore all possible spans and assign scores to them, leading to accurate prediction of named entities. For NCBI-Disease [38], Lee et al. [51] first proposed a domain-based BERT model—Bio-BERT pre-trained language model. BERT is a contextualized word representation model that uses the masked language model and is pre-trained using bidirectional transformers. The Bio-BERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a domain-specific language representation model which is pre-trained on large-scale biomedical corpora. Through a series of experiments, it was demonstrated that the pre-trained and fine-tuned Bio-BERT could identify biomedical named entities that were not recognized by BERT as well as find the exact boundaries of named entities. In general, researchers are more interested in NER techniques and creating NER datasets based on different research purposes, but few researchers have focused on investigating the history of the development of NER datasets. The study in [77] proposed a two staged fine tuning method for named entity recognition in geological text based on GeoBERT. The study used a bidirectional encoder representation from the transformers language model using the geological domain knowledge on a BERT model. In the second stage, smaller number of samples was used to complete the NER tasks in the geological report on the basis of GeoBERT. The proposed model achieved higher F1 score in comparison to the traditional approaches. The study in [78] used a conditional random field and long-short-term memory technique for named entity recognition in case of English texts. The proposed approach included three stages namely the pre-processing, feature extraction, and NER phase. The dataset was collected online. Then the URL was removed, special characters were removed, username was removed, tokenization was performed, and stop word removal was performed as part of the pre-processing phase. The essential features were extracted and then the data were subjected to the model for the purpose of training. The arithmetic optimization algorithm in association to CRF and LSTM was implemented for training the parameters of the model. The proposed model was validated using statistical measurements and also compared with the traditional convolutional neural networks which justified the superiority of the proposed approach.

This paper aims to systematically study NER datasets generated at different times, at different conferences, and in different mission contexts. Since the first English-only NER dataset was presented at the MUC-6 conference in 1995, the dataset has evolved to varying degrees along different dimensions depending on the research area, the goals of the conference, and the interests of the researchers. First, in the year following this, research on NER datasets in three languages—Chinese, Japanese and Spanish was introduced by the MET project, which marked the starting point of the multilingual NER task. Second, as the work on named entity recognition continued to advance, the news domain that was studied at the beginning could no longer meet the research needs, and researchers had to look farther into major domains as needed, and several major domains are now commonly covered, including news, biomedical, Wikipedia, scientific text, user text, etc. At the same time, differences in the formulation of entity categories are a direct result of the different fields of study. In addition, entity granularity is also changing to some extent, as named entity recognition is now performed as an underlying task in a variety of applications, and a shift from coarse-grained to fine-grained entities is inevitable due to the requirements of various applications for entity granularity [52]. In general, the NER dataset has developed in many aspects over the years, but there is no research work that has focused on the changes in the NER dataset over the years. Therefore, this thesis aims to dissect the potential development of NER datasets by collating information about the creation of NER datasets and their basic characteristics. Ultimately, a review of the pattern of development of NER datasets over the last 30 years can provide some insight into the creation of future datasets.

The contribution of this survey can be summarized as follows:

- Comprehensive review. We have conducted a comprehensive survey of the development of NER datasets over the years.
- New taxonomy. We have proposed a development framework by investigating many papers describing NER datasets. This development framework is based on the different evolutionary dimensions of the dataset. Further, research ideas are provided on possible future directions for the development of the dataset in each dimension.
- Future directions. Many NER datasets are discussed and analyzed and future research directions for NER datasets are proposed.

The rest of the paper is organized as follows: In Sect. 2, the development of NER datasets is outlined according to the language of the dataset, the research domain, the entity type, the entity granularity and the entity annotation, and the future direction of NER datasets is given in terms of different dimensions. In Sect. 3, an overview of common datasets is given in chronological order of their creation. In Sect. 4, a comprehensive discussion is presented concerning the possible linkages that exist between both the NER dataset and the mainstream NER techniques. In Sect. 5, all the above work is integrated to predict the future trends of datasets in general.

2 Taxonomy

As research on named entity recognition continues, an understanding of the development and evolution of NER datasets has become an integral part of this research. This section provides a chronological overview of the common NER datasets presented since the MUC-6 conference. It explores the development of NER datasets over the years in terms of language, research domain, entity type, entity granularity, entity annotation schema, and inferring how the creation of datasets may have changed since then. The most direct application of this review of NER dataset trends is to enable researchers to find the right NER dataset quickly and accurately for their research needs when training models. Second, researchers creating datasets can, to a certain extent, refer to the development process of existing datasets in a certain dimension to further create NER datasets that meet the expectations of research and meet the needs of technological development. In addition, understanding the contribution of NER datasets to a particular area of research at different times can provide insight into the focus of research at that time and can be a valuable reference for researchers working on related issues in the future. For example, the development of NER datasets in the biomedical field from scratch, both in terms of further development of NER datasets in this field and in other emerging areas of exploration, can provide informative examples in terms of data source selection, entity type formulation, entity annotation, etc. In short, the significance of the following work is to analyze relevant NER datasets in various dimensions and then to inform the creation of subsequent NER datasets in the light of their development over the years. The dimensions of NER presented in the thesis are shown in Fig. 1. Figure 1 divides the named entity recognition dataset from left to right in terms of language, research domain (entity types formulated based on the research domain), entity granularity, and entity annotation approach.

2.1 Language

Although the corpora used to create NER datasets over the years have largely been drawn from English texts, there has been a growing effort by researchers to create NER datasets using corpora from other languages than English. The MET conference held in 1996 marked the beginning

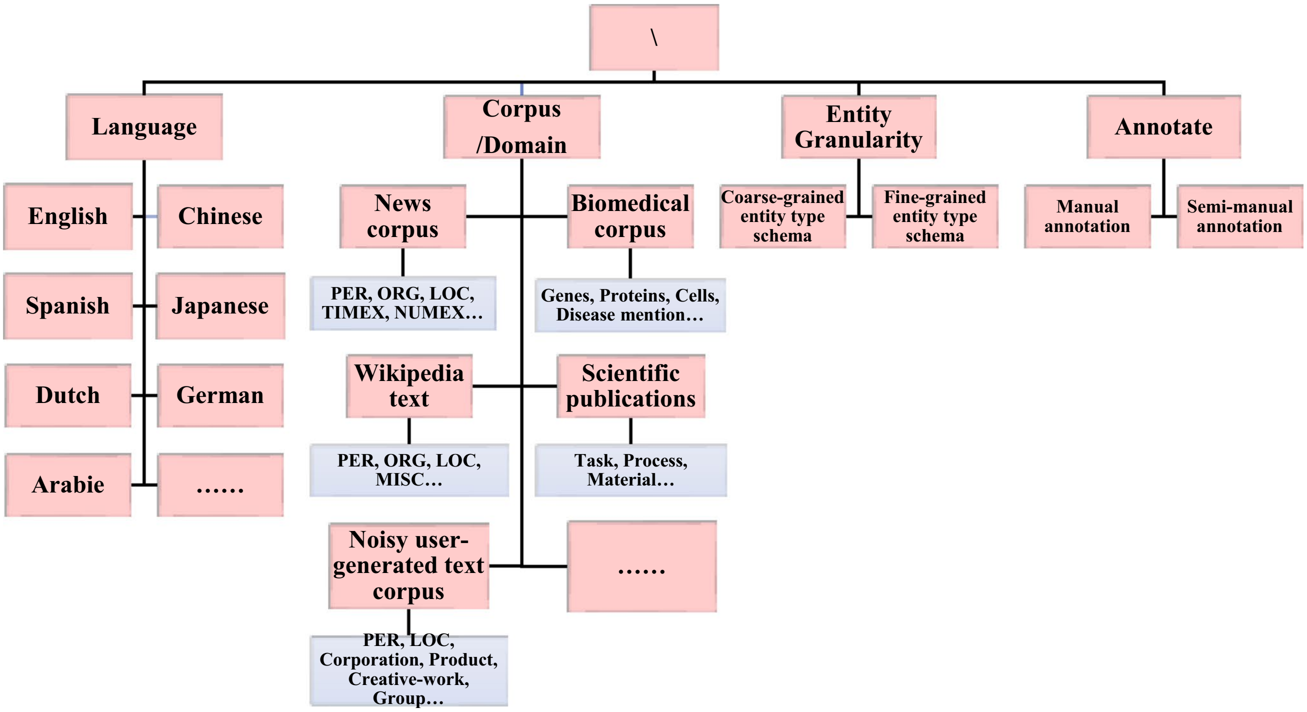

Fig. 1 Taxonomy of NER datasets

of multilingual NER, which used corpora from Chinese, Spanish, and Japanese to create NER datasets. With this attempt to introduce other languages, MET expects to examine whether the NER task will differ between languages. Furthermore, this initial trial provides research ideas for the development of the NER system in terms of portability between languages. Following this, the Japanese corpus was used in the Japanese-initiated IREX conference. CoNLL 2002 and CoNLL 2003 used Spanish, Dutch as well as English and German, respectively. According to [23], CoNLL's multilingual corpus aims to explore more general features for NER system training that are not restricted by language. ACE 2004 and ACE 2005 and OntoNotes 5.0, created in 2013, use English, Chinese, and Arabic. The N^3 corpus uses English and German. The above non-exhaustive examples of languages used, combined with the ranking of the most spoken languages in the world by https://www.berlitz.com/en-uy/blog/most-spoken-languages-world, show that the commonly used NER dataset already includes roughly the most frequently spoken languages. The aforementioned link is a blog by Berlitz which highlights the most spoken languages in the world as on September 23rd 2021. As per the blog, English is the most spoken language ranked at number 1 with 1,132 million speakers. For future research, more languages will be considered. NER datasets will not only be created in mainstream languages, but also in other languages for different research needs.

2.2 Corpus/Domain

2.2.1 Research Field

For NER datasets, commonly used generic domain datasets are NER datasets constructed from news-based corpora, which usually contain a substantial amount of familiar text and are more accessible than other domains, and therefore do not require a domain expert to guide the construction of the dataset. This is the reason why news texts have been used as a common applicable corpus to build NER datasets in the early stage of NER development. For illustration, the initial MUC, MET, IREX, CoNLL, ACE, and later the datasets BBN and NNE for the study of fine-grained entities and the large multilingual corpus OntoNotes were all NER datasets constructed using news-like texts. Nevertheless, as NER progressed, researchers embarked on other fields of research. Initially, the IREX conference introduced restricted domain (arrested) texts out of the necessity to study the portability of NER systems and the impact of domain texts on NER performance [22]. Since then, in addition to continued research in the general domain, research in the biomedical domain has also been ongoing, for example, the GENIA project, the GENETAG dataset, the StemNet project, the NCBI-Disease dataset, and the BioCreative V project have given impetus to the development of NLP technology through the continuous development of a substantial number of biomedical corpora. Evidently, these corpora provide effective data support for text mining tasks in the field of biomedicine. In addition to this, Wikipedia-type texts have since been introduced for research on NER tasks, and for example, WikiGold, WiNER. In more recent years, the Mention-level keyphrase identification sub-task for scientific publications was proposed at SemEval 2017 Task 10 for researchers searching for scientific documents [31]. Immediately afterwards, SCI-ERC further explored scientific documents by increasing the amount of data and adding more entity categories based on SemEval 2017 Task 10 and SemEval 2018 Task 7 [40]. In furthermore, apart from using professional, normalized texts to create NER datasets, there has also been a great interest in user-generated texts in re-cent years, and WNUT has been working on user-generated noisy texts for many years. The WNUT workshop emphasizes on natural language processing being applied to user-generated text which are noisy. These are usually found in social media, online reviews, web forums, clinical records, and language learner essays. As the online environment continues to open, users can generate many more texts on current events of the day, this phenomenon that provides a large textual resource for relevant NLP research on noisy texts. This type of user-generated text is more suitable for identifying emerging and rare named entities. In the later developmental stages, there is the emergence of datasets in the domains of speech and writing, food recipes, legal, and experimental protocols, such as DaNE [82], TASTEset [83], E-NER [84], the dataset proposed in the 2020 WNUT [85]. Meanwhile, some multi-domain datasets also exist, for example, CrossNER [79], MultiCoNER [86], Universal NER [87], and so on.

These trends show that at the beginning of the research process, the dataset was created from easy-to-understand news texts. The breadth and depth of the dataset have slowly expanded to some extent as different research needs have been explored in different fields. The most typical of these are biomedical, Wikipedia, and scientific texts, which are used as data sources for relevant research purposes. As a final point, the NER task is a fundamental core task, and in line with current trends and the future research needs of various industries in the field of NLP, the future creation of NER datasets will be more extensive in terms of corpus selection, and the texts used will contain, but not be limited to, all the types listed above.

2.2.2 Types of Named Entity

The above classification of the NER dataset is based on the different domain corpora used to create it. The corpora used are generally from the news domain (generic domain, unrestricted domain), the biomedical domain, the Wikipedia domain, scientific documents, and noisy user-generated texts. The fact that NER datasets constructed on different domain texts are constructed from different data sources inevitably leads to differences in the types of entities that need to be identified when completing the NER task later. In brief, the type of entity to be recognized by the NER task can only be confirmed once the domain data source has been determined. For example, the NER dataset built on generic domain text, the main types of entities to be recognized are Entity (ENAMEX) (Person (PER), Organization (ORG), Location (LOC)), Time (TIMEX), and Number (NUMEX), which are the three main entity categories. These are the entity categories defined at the beginning of the development of NER (MUC, MET) and can basically encompass the named entity types that appeared in general scenarios. In addition, the recognition of other entity types than those listed above was also requested during the development of NER in accordance with research needs. For example, although IREX was also created based on a news-based text, the Japanese organization proposed the identification of the entity category ARTIFACT [22]. Subsequently, with the maturity of the two types of entity category recognition technologies, time (TIMEX) and number (NUMEX), these two categories have rarely been part of entity recognition since the CoNLL conference. In particular, the CoNLL conference suggested that in addition to the above-mentioned identification of PER, ORG, and LOC, there was also a demand for the recognition of miscellaneous items (MISC), i.e., the need to identify the name of any other entity that does not belong to the three previously mentioned types [23]. Then, when the three entity categories are shown above (ENAMEX) were no longer satisfactory for the research needs, new entity types were continuously added to the entity recognition task according to the needs of the research task. For example, ACE2004 and ACE2005, as early multi-category NER datasets, added the following entity categories: Facility, Weapon, Vehicle, and Geo-Political Entity [47]. The BBN dataset created in the same period not only introduced more new entity categories but also provided a more detailed delineation of entity categories. Specifically, BBN proposes 12 named entity types, 9 nominal entity types, and 7 numeric types [35]. OntoNotes 5.0, proposed in 2013, contains 18 named entity categories, which are broadly consistent with the BBN entity categories as it draws somewhat on the BBN dataset's entity category delineation.

When researchers are annotating NER datasets created based on the biomedical domain, the entities of interest are very different from when annotating newswire texts. For example, the GENIA project, set up to promote the development and evaluation of information extraction in the medical field, focused on gene and protein and cell identification [25, 26], followed by the GENETAG dataset and the PROGENE dataset, which focused only on gene and protein recognition. Since then, research in the biomedical field has been in full swing, and specific research in this area has become more practically oriented. For example, the

NCBI-Disease, disease name corpus, presented in 2013, is a valuable research resource in the field of biomedical natural language processing and has become a highly representative NER dataset in the identification of disease names [53, 54]. In addition, the BC5CDR dataset is annotated with relevant chemical entities as well as disease entities for the sub-task of Disease Named Entity Recognition (DNER), to facilitate research related to chemical-disease relationships [29]. The ultimate aim is to improve the chemical safety, reduce toxicity, and improve the survival of pharmaceutical compounds by identifying adverse drug reactions (ADRs) that may exist between chemicals and diseases, thereby facilitating research into new drugs and enhancing drug safety management [29]. It is evident that for the field of biomedicine, the datasets proposed later that can be used for NER tasks are becoming more and more targeted, and have more and more practical significance in terms of medical practicability.

The types of entities to be recognized in the NER datasets that are later created on scientific documents are also very diverse from those mentioned above. For NER tasks on scientific publications, the main objective is to use the key phrases of tasks, technologies and resources that appear in the scientific documents and the possible relationships between them to help researchers with a need for such articles to search for the target article precisely. Therefore, the key phrases that need to be identified for this type of NER dataset revolve around Task, Process and Material (i.e., the three types of keyphrases that need to be identified for the Mention-level keyphrase identification sub-task of SemEval 2017 Task 10). After this, researchers continued to explore scientific documents with the expectation of better training the NER system by expanding the dataset and extending it with more entity types. The SCI-ERC dataset proposed in 2018 is another relevant dataset created following SemEval 2017 Task 10 and SemEval 2018 Task 7. The SCI-ERC dataset aims to increase the coverage of the scientific information domain and is based on previous datasets created by extending entity types and relationship types [40]. As a result, the SCI-ERC required the identification of more keywords than the NER tasks of SemEval 2017 Task 10 and SemEval 2018 Task 7. In addition, the subsequently proposed SoMeSci dataset serves as a comprehensive corpus on software identification in the domain of scientific information, which can help to maximize the identification of software types and their associated mentions [88].

In addition, mention must be made of the NER datasets constructed on the basis of the user-generated noisy text. These datasets were originally created to detect emerging and rare named entities on user-generated noisy text. The current open online environment has led to an increasing number of online users willing to contribute their own opinions and insights on real-time hot topics. In this background, the increasing amount of user-generated text on current hot topics directly provides a considerable amount of textual resources for NLP research on noisy text. Since its inception, the WNUT project has been dedicated to the study of user-generated noisy texts. Due to the specificity of its research purpose and the complexity and diversity of its data sources, WNUT prefers to identify the categories of entities that online users are likely to talk about from the text. For example, in addition to the above entity types that are commonly identified on news-based NER datasets, in WNUT 2016 the entity types Company, movie, music artist, Product, Sports team, and Tv show also need to be recognized [30].

The types of entities that need to be identified have been described above according to different research areas, and based on this it is possible to define a general pattern of the types of entities that need to be identified for NER tasks over the years. As elaborated above, the generic domain-based named entity recognition dataset mainly recognizes PER, ORG, and LOC, but will be adjusted correspondingly with the conference and the creator's goals, for example, some of the datasets add the recognition of Facility, Vehicle, geo-political, nationality, product. Among them, the BBN and NNE datasets basically contain all the entity type tags in the previously proposed datasets. In addition, the subsequent CrossNER and FEW-NERD as multi-domain datasets involve more refined entity types. The later datasets created for the biomedical domain, scientific information domain, and Wikipedia, are more focused on the recognition of entity types within the domain, which are more specialized and domain-specific and can better contribute to the development of natural language processing tasks in the current domain. In general, the recognition of entity types in each dataset needs to be based more on the research covered by the domain, and the more specialized it is, the more it can provide some support for subsequent research. But it is inevitable that some entities will have type ambiguity in their identification. For example, an entity defined as Location in one dataset may be defined as Organization, company, etc. in other datasets. In particular, this is the case with names like some universities and companies. However, the actual problem behind this is much more than simply inconsistent entity type definitions. Further dissection of this shows that inconsistencies in the definition of entity types across different NER datasets may directly lead to the training of NER systems that are not well generalized. This means that a NER system that works well by being trained on one dataset may yield very different results if it is tested on another dataset. In other words, NER systems trained on different datasets are not comparable and can only be simply compared to systems trained on the same dataset for accuracy. Therefore, the comparison between NER systems trained on different datasets is somewhat one-sided. However, looking through the phenomenon, this current situation indirectly provides research ideas for the future development of NER datasets.

In terms of inconsistent definitions of named entity categories, an attempt can be made to integrate as many corpora and datasets as possible and to standardize and refine their definitions of named entity types. In this way, the trained-NER systems can be compared in a meaningful way.

2.3 Entity Granularity

Most datasets constructed based on news texts require the identification of entity types involving PER, LOC, and ORG, and some also include miscellaneous categories (MISC), time (TIMEX) and numeric (NUMEX) expressions, e.g., MUC-6, MUC-7, MET, CoNLL 2002, CoNLL 2003. MET, CoNLL 2002, CoNLL 2003, these early datasets were only concerned with the recognition of coarse-grained named entities as shown above. However, as related technologies continue to advance and research progresses, it is not sufficient for NER, which is the core foundation task, to simply identify coarse-grained entities. The NER task provides the underlying support for many practical applications such as relationship extraction, entity linking, question answering system and many more, so further processing and classification of coarse-grained entity classes into fine-grained entity classes is inevitable for future developments. Two gold standard datasets, ACE 2004 and ACE 2005, provide a more fine-grained delineation of named entity categories. The ACE 2004 dataset, for example, contains the following entity categories: PER—no subtypes, ORG—5 subtypes, LOC—10 subtypes, Facility (FAC—8 subtypes), Weapon (WEA—9 subtypes), Vehicle (VEH—5 subtypes), and Geo-Political Entity (GPEs-6 sub- types), and 5 to 10 subtypes under each entity type [47]. In addition to this, at basically the same time, the BBN dataset was proposed to more refine the categories of entities, with a total of 12 named entities, 9 nominal entity types and 7 numeric types, several of which can be further subdivided into subtypes, for a total of 64 entity categories [35]. The presentation of the BBN dataset implies a reference for a more fine-grained entity classification for NER in the generic domain. Further, it is not only the general domain NER that has a fine-grained entity classification but also in specific domains such as the biomedical domain, where the requirement for terminological precision is very high, such NER datasets usually have a finer classification of entity types. Specifically, the GENIA dataset for the biomedical domain contains a total of 36 different entity types in biology, with finer-grained differences between the different types.

Other than the above due to the continuous development of NER, the various applications based on NER and the domain-specific delineation of fine-grained entity types, the better control of data through the use of fine-grained entity delineation is another reason that cannot be ignored, which also promotes fine-grained entity delineation. For example, the WikiGold dataset and the WiNER dataset, two Wikipedia-based datasets, do not perform new entity typing, but instead obtain directly from the named entity tags (PER, ORG, LOC, MISC) from the previous CoNLL 2003 gold standard dataset for entity annotation. However, it is worth noting that the entity annotation is performed separately using coarse-grained entity labels and fine-grained entity labels, with the final mapping of fine-grained labels to coarse-grained labels to complete the annotation task. In this process, however, it was found that mapping fine-grained labels to coarse-grained labels resulted in more consistent entity annotation results [36]. This particular annotation approach not only provides a reference for subsequent annotation of other datasets but also reflects the fact that datasets annotated with fine-grained labels can be adapted to other entity classification schemes to some extent by mapping. This further illustrates that datasets annotated with fine-grained labels can be applied to different tasks to a greater extent than other datasets in general.

It is important to mention, the challenges that the delineation of fine-grained entities poses for performing NER tasks. Fine granularity directly implies a significant increase in the number of named entity types and the complexity introduced by a named entity having multiple subtypes at the same time [3]. Notwithstanding this, fine-grained entity delineation is now the dominant direction in the development of NER datasets, and more and more researchers will work on developing fine-grained entity NER datasets in the future.

2.4 Annotation

The creation of a named entity recognition dataset begins with the determination of the research area and the objectives of the project, followed by the selection of a suitable corpus based on the specific needs and the preparation of data annotation guidelines, and finally the arrangement of the relevant researchers to annotate the entity types. As the final step in the creation of a dataset, the quality of the annotation is crucial to a dataset. Over the years NER datasets have evolved to varying degrees in a variety of aspects. However, in the quest for consistency in the annotation of named entities, researchers have continued to introduce new annotation schemes in an attempt to achieve a high level of consistency in this task. In addition to the linguistic knowledge of syntax and semantics required for the annotation task, a certain degree of domain expertise is also required when it comes to the annotation of named entities in specific domains. In addition, different NER datasets have also designed different schemes to achieve the consistency of entity annotations. For example, WikiGold has adopted the scheme of mapping fine-grained tags to coarse-grained tags to pursue consistency in named entity annotation [36]. ACE performs consistency checking of data by crossing teams and languages [24]. OntoNotes 5.0 integrates all annotations into one database to aid in the consistency checking of data annotations [28, 55]. In addition, due to the increasing demands on the size of datasets today, some datasets are annotated using semi-manual methods in addition to fully manual annotation, with the help of experts to correct the automatic annotation results. For example, the annotation of the GENETAG dataset was first performed by AbGene tagger and then manually corrected by biochemistry, genetics and molecular biology experts through a web interface [33, 34]. Reuters-128 in the N^3 corpus was primarily annotation done by having domain experts manually modify named entity annotation errors caused by FOX annotation [56]. The annotation of the SemEval 2018 Task 7 dataset was first done using automatic annotation of named entities, followed by error correction by manual annotators, especially for the entity boundary misannotation problem [32].

High-quality annotation provides a good entity annotation dataset for the subsequent training of entity recognition models so that researchers can continue to produce high-performance NER systems. It is important to mention that the consistency of the annotation task plays an important role in subsequent entity recognition research, however, it is generally accepted that the annotation task is not a simple task to perform even for professional linguists or domain annotation experts with a linguistic background [57]. Therefore, finding a suitable scheme for the annotation of named entities is an urgent task and is essential for the creation of a high-quality NER dataset. Furthermore, the repeated increase in the accuracy required of NER systems has led to a pressing need for large corpora of high-quality annotations. The construction of such a corpus cannot rely solely on manual annotation by experts, and therefore it is inevitable that the annotation task for the NER dataset will evolve from manual to semi-automatic or even fully automatic annotation. Reducing the degree of manual intervention in the subsequent creation of NER datasets will be the principal goal of the annotation work.

3 Overviews of Commonly Used NER Datasets

In addition to the above classification of the NER datasets, it is essential to understand this work in terms of the creation of each dataset. Table 1 details the high-quality NER datasets mentioned in this paper in the chronological order of their creation. Starting with the introduction of the NER concept at the MUC-6 conference, these datasets are organized by year of creation, language, and research area, while the datasets are subsequently elaborated in terms of their creation goals and contributions to the NER mission, as well as their storage format. Meanwhile, the above research work on NER datasets is synthesized, describing the emphasis of the research work on NER datasets in terms of the early, mid and late development. The expectation is to provide as detailed a description as possible of the overall development of the dataset through the evolution of the mainstream NER dataset. Also, Table 2 systematically summarizes the tagged entity types in the dataset to help readers better understand the types of entities that need to be recognized in different domains, as well as the needs and goals of the named entity recognition task from another dimension.

Pre-term development of the NER dataset: MUC-6 as the starting point for the development of NER, providing a definition of named entity recognition and specification of tasks and the annotation format of the data, provided the basis and reference for subsequent work on the creation of datasets. In the year that followed MET made its first attempt at NER tasks in languages other than English. This experiment was not only the starting point for multilingual NER but also provided research ideas for the development of NER systems that could be transferred between different languages. In the same year as MET-2 (1998), IREX, a conference based on information retrieval and extraction in Japanese, introduced a new domain text to study the portability of NER systems and the effect of domains on NER performance. In this conference, models were trained and evaluated using texts from two different domains, restrained domain (Arrest) and unrestricted domain (News category) [22]. Furthermore, in addition to the recognition of PER, ORG, LOC and time (TIMEX) and number (NUMEX) expressions, IREX also added the recognition of ARTIFACT types. By this time, the NER dataset had already experimented with other linguistic and domain texts and introduced new entity types. In other words, in the first 3 years of the NER task, researchers have been investigating possible variations of the NER dataset in terms of language, domain and entity type.

Mid-term development of the NER dataset: CoNLL proposed in 2002. At this time, due to the continuous development of related technologies, CoNLL has made adjustments in the formulation of the NER task and the direction of its research in line with the technological development. First, TIMEX and NUMEX were no longer identified as entity types in CoNLL, as they could already be recognized very well. Second, rule-based NER systems were no longer advantageous in the context of multilingual corpora, and therefore at that time, CoNLL aimed to discover more general features that were not restricted by language to develop statistical-based NER systems. In addition, as many researchers at the time were dedicating a great deal of effort to machine learning-based research, the CoNLL dataset, as the largest dataset available for NER research at the time, provided reliable data support for the development of machine learning-based NER systems. It can be seen that at that time CoNLL 2002 and CoNLL

Table 1 List of commonly used NER dataset

Dataset/conference	Year	Language	Corpus/domain
MUC-6	1995	English	News
MUC-7	1998	English	News
MET-1	1996	Chinese, Spanish, Japanese	News
MET-2	1998	Chinese, Japanese	News
IREX	1998–1999	Japanese	News, restricted domain (arrest)
CoNLL 2002	2002	Spanish, Dutch	News
CoNLL 2003	2003	English, German	News
ACE 2004	2004	English, Chinese, Arabic	News
ACE 2005	2005	English, Chinese, Arabic	News
GENIA	2004	English	Biomedical
GENETAG	2005	English	Biomedical
BBN	2005	English	News
WikiGold	2009	English	Wikipedia
FSU–PRGE/PROGENE	2010	English	Protein
WiNER	2013	English	Wikipedia
OntoNotes 5.0	2013	English, Chinese, Arabic	News
NCBI-Disease	2013	English	Biomedical
N3	2014	German, English	News
BC5CDR	2015	English	Biomedical
WNUT 2016	2016	English	User-generated text
WNUT 2017	2017	English	User-generated text
SemEval 2017 Task 10	2016	English	Scientific publications
SemEval 2018 Task 7	2017	English	Scientific publications
SCI-ERC	2018	English	Scientific publications
NNE	2019	English	News
CoNLL + +	2019	English	CoNLL 2003
CrossNER	2020	English	Politics, natural science, music, literature, and AI
DaNE	2020	Danish	Speech and writing
WNUT-2020 Task 1	2020	English	Experimental protocols
FEW-NERD	2021	English	Wikipedia
RadGraph	2021	English	Chest X-ray radiology reports
SoMeSci	2021	English	Scientific articles
TASTEset	2022	English	Food recipes
MultiCoNER	2022	Multilingual	Wiki, questions, and search queries
E-NER	2022	English	Legal
Universal NER	2023	Multilingual	Mainly involves general domains, such as news, blogs, email, reviews, wiki, web, etc

2003 already integrated the feature that the dataset could be multilingual and new entity types could be introduced. Furthermore, it has to be mentioned that the creation of the CoNLL dataset can reflect some extent the changes that occurred in NER technology at that time. The ACE project, which has been conducted since 2002 as a successor to the MUC named entity recognition task [62], shifted the emphasis of the research from the initial entity recognition to entity resolution. Compared to MUC, ACE has not only changed by adding more entity types and performing subtyping but also by considering the annotation of nested entities. ACE 2004 and ACE 2005, the most commonly used NER datasets in the ACE project, provided a reference for the creation of subsequent datasets in terms of entity types, sub-categories division and annotation of nested entities. In the same period, research in the biomedical field was also conducted, with the creation of GENIA in 2003 and GENETAG in 2005 as commonly used datasets for extracting biological entities, providing resources for the use of NLP techniques for text mining in the biomedical domain. As can be seen, the creation of datasets in this period has been more varied than before,

Table 2 Entity type tags for commonly used NER datasets

Dataset/conference	Entity type tags
MUC-6, MUC-7 MET-1, MET-2	ENAMEX (PERSON, ORGANIZATION, LOCATION), TIMEX (DATE, TIME), NUMEX (MONEY, PERCENT)
IREX	ENAMEX, TIMEX, NUMEX, ARTIFACTS
CoNLL 2002, CoNLL 2003	PER (persons), ORG (organizations), LOC (locations), and MISC (miscellaneous items)
ACE 2004, ACE 2005	PER (persons), ORG (organizations), GPE (Geo-political Entity), LOC (locations), FAC(Facility), VEH (Vehicle), and WEA(Weapon)
GENIA	Covers biological entities such as proteins, genes, and cells, with a total of 36 species
GENETAG	The acceptable alternatives for gene and protein names are tagged
BBN	In addition to the common PERSON, ORGANIZATION, LOCATION, FACILITY, GPE, DATE, TIME, PERCENT, MONEY, there are also NATIONALITY, PRODUCT, EVENT, WORK OF ART, LAW, LANGUAGE, CONTACT-INFO, PLANT, ANIMAL, SUBSTANCE, DISEASE, GAME, ORDINAL and CARDINAL. INFO, PLANT, ANIMAL, SUBSTANCE, DISEASE, GAME, ORDINAL, and CARDINAL, for a total of 64 named entity types
WikiGold	PER (persons), ORG (organizations), LOC (locations), and MISC (miscellaneous items)
FSU–PRGE/PROGENE	Protein, protein family or group, protein complex, protein variant, protein enum
WiNER	PER (persons), ORG (organizations), LOC (locations) and MISC (miscellaneous items)
OntoNotes 5.0	PERSON, ORGANIZATION, LOCATION, FACILITY, GPE, NORP (NATIONALITY or RELIGIOUS, POLITICAL or OTHER), PRODUCT, EVENT, WORK OF ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL
NCBI-Disease	There are four types of disease tagged: Composite mentions, Modifiers, Disease Class mentions, and Specific Diseases
N3	PERSON, ORGANIZATION, LOCATION
BC5CDR	Relevant chemical substance entities as well as disease entities are tagged
WNUT 2016	PERSON, LOCATION, CORPORATION, FACILITIES, FILMS, MUSIC ARTISTS, PRODUCT, SPORTS TEAMS, TV PROGRAMES and OTHERS
WNUT 2017	PERSON, LOCATION, CORPORATION, PRODUCT, CREATIVE-WORK and GROUP
SemEval 2017 Task 10, SemEval 2018 Task 7	Labels the three types of entities Task, Material, and Process that appear in scientific publications
SCI-ERC	Task, Method, Metric, Material, Other-Scientific Term and Generic
NNE	The entity types were extended based on the BBN dataset and a total of 114 entity types were tagged
CoNLL + +	PER (persons), ORG (organizations), LOC (locations) and MISC (miscellaneous items)
CrossNER	Different domains are tagged with different types of entities. For example, the Politics domain is tagged with politician, political party, event, election, etc. The AI domain is tagged with field, task, product, algorithm, researcher, metrics, etc
DaNE	PER (persons), ORG (organizations), LOC (locations) and MISC (miscellaneous items)
WNUT-2020 Task 1	CONSTITUENTS, QUANTIFIERS, SPECIFIERS, ACTION, and MODIFIERS
FEW-NERD	The eight entity types of Person, Location, Organization, Art, Building, Product, Event, and Miscellaneous are tagged, where different entity types are tagged with different more fine-grained types. For example, Organization contains the following specific types: company, Education, Government, Media, Political/party, Religion, Sports League, Sports Team, Show ORG, and others
RadGraph	Anatomy, Observation (Definitely present, uncertain, definitely absent)
SoMeSci	Type of software, Type of Mention, and additional information
TASTEset	FOOD, QUANTITY, UNIT, PROCESS, PHYSICAL QUALITY, COLOR, TASTE, PURPOSE, PART
MultiCoNER	PER (persons), CORP (corporation), LOC (locations), CW (creative-work), GRP (groups), PROD (product)
E-NER	Location, Person, Business, Government, Court, Legislation/Act, Miscellaneous
Universal NER	PER (persons), ORG (organizations), LOC (locations), OTH (other)

both in terms of depth and breadth of work. The datasets created during this period are still widely used today.

Post-term development of the NER dataset: The WikiGold dataset, created in 2009 based on Wikipedia, was annotated using the named entity annotation scheme in CoNLL 2003. The WiNER, created in 2012, also uses the same coarse-grained named entity annotation scheme as WikiGold to perform the annotation task. In addition, SemEval 2017 Task 10 and SemEval 2018 Task 7, designed for keyword identification based on scientific publications, provided the basis for the creation of SCI-ERC. The SCI-ERC is based on the datasets published in SemEval 2017 Task

10 and SemEval 2018 Task 7 and was created by adding entity types and relationship types, with the aim of increasing the coverage of the scientific information domain as comprehensively as possible. In addition to this, the NNE dataset created in 2019 references the fine-grained entity schema of the BBN dataset at the entity granularity aspect, expanding from the 64 entity types of BBN to the current 114. Furthermore, CoNLL + +, proposed in the same year, was created based on the modification of entity annotation errors in CoNLL 2003 and resulted in a more accurate NER test set than before. The CrossNER dataset proposed in 2020 covers multiple languages and domains, which to some extent provides valuable reference for the creation of subsequent datasets. In the same year, the other two datasets DaNE and WNUT-2020 Task 1 were proposed to explore the new domains of speech and writing and experimental protocols, respectively. After that, FEW-NERD, RadGraph, and SoMeSci, though all of them are in the researched domains, gradually become more specialized based on the original ones. The datasets created in 2022 broaden the scope of research even further, with TASTEset covering the domain of recipes, MultiCoNER covering questions, and search queries, and E-NER covering the domain of legal. The datasets created are substantially improved in terms of both size and quality compared to similar datasets introduced previously. Therefore, from the point of view of creating a new dataset, it is possible to refer to the work done on previously created datasets, for example, by correcting errors in the previous dataset, adding or deleting the types of entities to be recognized, etc. to create a new NER dataset that meets the needs of the research.

In addition, more researchers have been focusing more on named entity recognition in small language specialization domains in recent years, collectively working on the overall development of the natural language processing domain. For example, the following datasets were proposed in 2021 the African language dataset MasakhaNER [89], the Modern Hebrew language dataset NEMO^2 [90], the Korean language dataset KLUE [91], and the Czech language dataset SumeCzech [92], and the LegalNERo dataset focus on the legal domain [93]. In 2022, KazNERD is a Kazakh dataset for recognizing the news domain [94], HiNER is a Hindi dataset for recognizing the news domain and tourism domain [95], MobIE is a German dataset for recognizing entities in social media texts and traffic reports corpus [96], KIND is an Italian Multi-Domain Dataset for recognizing entities in news, literature, and political discourses [97]. Recently, the newly proposed Naamapadam integrates a large corpus of Indian languages for named entity recognition, and BanglaCoNER focuses more on complex named entity recognition in Bangla [98, 99].

Different dataset/conference has different goals. MUC-6/MUC-7: Facilitating and evaluating information extraction studies. MET-1/MET-2: Investigating whether the NER task varies between languages [58]. IREX: Research on Japanese-based information retrieval and extraction. CONLL 2002/CONLL 2003: Use of multilingual corpora to explore more general features that are not restricted by language [52]. Development of more statistically based NER technology. Construction of the largest dataset at the time to facilitate the study of ma-chine learning-based NER systems [52]. ACE 2004/ACE 2005: The research focuses on key technologies that promote relevant automatic entity recognition, relationship recognition and event recognition. GENIA: Supporting the natural language processing in the field of molecular biology [25, 26]. GENETAG: The creation of a large available corpus containing gene/protein tags to evaluate AbGene previously developed by researchers [59]. A different annotation format from the GENIA corpus leads to a more meaningful assessment of the performance of the NER system [34]. BBN: Provides a fine-grained entity annotation reference. WikiGold: Using Wikipedia's large, semi-structured features to create NER datasets. FSU–PRGE/PROGENE: The goal of the PROGENE is to create a large, comprehensive and reliably annotated protein/gene corpus that can be used for supervised training and quality assessment based on machine learning in the domain of biology [27]. WiNER: Training the NER system by continuously creating NER datasets based on encyclopedic texts to improve the performance of the system. OntoNotes 5.0: The goal of the OntoNotes project is to create a research resource that is applicable in many aspects of the field of natural language processing by annotating a large corpus. NCBI-Disease: Promoting automated disease name recognition technology. N^3: A collection of datasets that can be used for named entity recognition and disambiguation. BC5CDR: The aim is to improve the chemical safety, reduce toxicity, and improve the survival of pharmaceutical compounds by identifying ADRs that may exist between chemicals and diseases, thereby facilitating research into new drugs and enhancing drug safety management [29]. WNUT 2016/WNUT 2017: The aim is to identify emerging named entities in the user-generated text [61]. SemEval 2017 Task 10/SemEval 2018 Task 7: Targeting keywords such as tasks, technologies, resources, and discovering relationships between them in scientific documents helps researchers to conduct the next research through keyword extraction. SCI-ERC: Better training of scientific document based NER systems by increasing the size of the dataset and adding more types of entities [40]. NNE: This dataset was created primarily for the research of nested named entities. CoNLL + +: Accurate re-evaluation of the NER system by modifying annotation errors in the test set. CrossNER: addresses the problems of named entity recognition in terms of domain adaptation. DaNE: provides the largest gold annotated dataset available for research. FEW-NERD: fine-grained large-scale dataset created around rare entities. RadGraph: dataset used in the medical domain for recognizing entities in chest X-ray radiology reports. SoMeSci: a dataset for identifying software entities and mentions in the scientific domain. TASTEset: a dataset designed to facilitate the extraction of information from recipes. MultiCoNER: provides a cross-language diverse text corpus to address the current challenges of named entity recognition. E-NER: remedies the difficulty of accurately extracting entities from legal texts by the current common models. UniversalNER: covers entities in multiple language contexts to meet the diverse needs of information extraction.

Different dataset/conference has different explainations. MUC-6/MUC-7: The starting point for NER. MET-1/MET-2: The starting point for multilingual NER. IREX: The introduction of new entity type. The introduction of new domain texts. CONLL 2002/CONLL 2003: The creation of NER datasets using other languages. The introduction of new entity types. Responding to trends in technology. ACE 2004/ACE 2005: The creation of NER datasets using other languages. The introduction of new entity type. All mentions of each entity are annotated, and nested mentions are also annotated. Further annotation according to the category of the entity (NEG, ATR, SPC, GEN, USP). GENIA: The creation of NER datasets using other languages. A suit- able nested entity annotation structure has been developed based on the constitutive form of the biological terms. the GENIA corpus was semantically annotated by experts using descriptors from the GE- NIA ontology [26]. GENETAG: Creation of datasets in the biomedical field that can be used for NLP research. WikiGold: The introduction of new do-main texts. The gold standard NE tag was used to annotate 145 articles selected from Wikipedia. FSU–PRGE/PROGENE: The PROGENE corpus is in-tended to cover as many do-mains of biology as possible, and the entire corpus consists of 11 sub-corpora, any two of which are independent of each other [27]. WiNER: Same as the WikiGold dataset, with text from Wikipedia annotated using the gold standard NE tag. OntoNotes 5.0: The BBN dataset was referenced for the determination of the entity types. The multiple annotation layers of the corpus consider structural information and shallow semantics. NCBI-Disease: The NCBI-Disease was developed based on the AZDC corpus, which is more informative and complete than the AZDC [60]. NCBI-Disease is a valuable research resource in the domain of biomedical natural language processing and is a highly representative dataset for identifying disease names. N^3: The entire dataset consists of 3 sub-datasets. The NLP Interchange Format (NIF) was used to facilitate interoperability, considering the storage of the dataset [39]. BC5CDR: The datasets that have been proposed for use in NER tasks are becoming more and more targeted, and have more and more practical significance in terms of medical practicability. WNUT 2016/WNUT 2017: The large amount of noisy text currently available provides a substantial data resource for NLP research. SemEval 2017 Task 10/SemEval 2018 Task 7: The introduction of new do-main texts. NNE: Created based on the BBN dataset. CoNLL++: A more accurate NER test set was obtained by modifying data annotation errors that appeared in the CoNLL 2003 test set. CrossNER: improves the generalization of the model under multi-domain and multi-language by integrating and annotating multiple resources. DaNE: follows the entity types in the CoNLL2003 dataset. FEW-NERD: the first dataset for rare entity recognition. RadGraph: further advances natural language processing in the healthcare domain. SoMeSci: the most comprehensive dataset for recognizing software mentions in the current domain. TASTEset: allows for the extraction of more complex entities in recipes. MultiCoNER: further enhances the performance of the model by performing entity recognition in challenging scenarios. E-NER: a new dataset for the legal domain that enhances the performance of models for recognizing legally relevant entities. UniversalNER: Promotes model generalization and cross-language and cross-domain recognition capabilities.

Dataset Storage Preferences. At present, the storage of named entity recognition datasets is usually in the form of TXT files, and some datasets are also stored in CSV and JSON formats. Although the formats can be converted to each other, the choice of storage format depends more on the actual usage requirements, for example, whether the data storage is easy-to-understand, whether the data is easy to analyze, and how easy it is to train the subsequent model.

4 Interaction of NER Dataset with NER Technology

Given that NER datasets are created to test NER research techniques, it is undoubtedly necessary to analyze NER datasets from the perspective of NER techniques. Therefore, in addition to the above classification of NER datasets, the mainstream NER methods and NER datasets in the order of development are discussed next. The progress of this research domain is provided comprehensively through the changes in the technical routes and the development of NER datasets over the years. In the following, the past research work is reviewed in terms of three named entity recognition methods: rule-based methods, machine-learning-based algorithms, and multi-technology fusion methods, respectively.

4.1 Rule-Based Methods

The rule-based method means that entity recognition relies on the rules manually formulated in advance by domain experts, and when rules are complete, good entity recognition results can be obtained. However, also owing to its specific entity recognition approach, rule-based methods are difficult to be transferred to other domain datasets. Rule-based methods were first used for the MUC dataset as the first technique of its kind for entity recognition. Already in 1995, the FASTUS system [63] and LaSIE system were used for the NER dataset presented at the MUC-6 conference [64, 65]. The FASTUS is a system which is used for extracting information from free text in English to be entered into a database or any other applications. The LaSIE system is also known as large-scale information extraction system which was developed at the University of Sheffield as part of their research on natural language engineering. It is a single integrated system that develops a unified model of a text which helps in generating outputs for all the tasks in MUC-6. It is implemented as a cascaded non-deterministic finite state automation. Rule-based methods have been proposed continuously since then, such as the LaSIE-II system [66] and the FACILE system [67], and the SRA system for the MUC-7 dataset in 1998 [68]. The rule-based approach does not require much from the dataset itself, with the exception that the rules prepared by domain experts correspond as comprehensively as possible to the named entities to be extracted from the dataset.

4.2 Machine-Learning-Based Algorithms

The supervised learning-based approach is gradually being applied to NER tasks along with the rule-based approach. Its use of high-quality large-scale labeled datasets to train models that can recognize named entities. In 1997, Bikel et al. pioneered the use of the Hidden Markov model (HMM) for the NER task [69]. Bikel used HMM not only on the English dataset (MUC-6) but also on the Spanish dataset (MET). Meanwhile, the Maximum Entropy model (MEM) was applied to the MUC-7 NER dataset by Borthwick et al. and the portability of this model was validated using capitalized English text [70]. In addition, there are other methods based on supervised learning. The Conditional Random Fields (CRF) has been commonly applied to problems such as natural language processing by the end of the twentieth century [71]. The CRF are class of statistical modeling which is also applied in pattern recognition and machine learning for achieving structured prediction. It's a commonly used approach in NER wherein a linear chain CRF connects to a labeler in which the tag assignment depends only on the tag of the previous word. Reference [71] formally used the CRF model for the entity recognition problem. In 2002 McNamee et al. used support vectors machines (SVMs) to identify entities on the Spanish and Dutch datasets of CONLL 2002 [72]. In 2006, reference [73] illustrated how to use decision tree for entity recognition on English (CoNLL 2003) and Hungarian (Szeged corpus). From the above research work, the non-portable nature of rule-based methods is largely overcome by such algorithms. According to the nature of supervised learning methods, the demand to improve the accuracy of NER models from a dataset perspective tends to require the acquisition of a larger amount of data while also improving the accuracy of entity annotation. However, since manual entity labeling is a time-consuming and laborious task, NER systems based on Semi-supervised learning methods and Unsupervised learning methods soon evolved for this reason. Semi-supervised learning uses only a small amount of labeled data for entity recognition through iterative and continuous learning, and Unsupervised learning uses a dataset without any entity annotation for entity recognition. Semi-supervised learning-based and Unsupervised learning-based NER systems therefore largely solve the problem of expensive entity annotation and avoid the difficulty of annotating entities across languages and research domains. In general, machine learning-based algorithms can generally perform NER tasks on different languages and research domains without any major modifications, thus maintaining good portability [73]. On the other hand, considering from the perspective of NER datasets, machine learning-based NER research methods provide approaches to improve recognition accuracy. The most important problem of expensive entity annotation can be solved by developing a suitable entity annotation scheme and studying NER datasets that can be automatically annotated. For example, the SemEval 2017 Task 10 dataset is automatically annotated, and the problem of automatic entity boundary annotation errors is corrected by manual annotation, which in turn improves the accuracy of annotation [32].

4.3 Multi-Technology Fusion Methods

Multi-technology fusion approaches often have the advantages of each of the methods being used, for which reason researchers are constantly working to combine related technologies to improve the accuracy of entity recognition. In other words, the strengths of one method compensate for the possible deficiencies of another, or several methods are used to jointly perform the named entity recognition task by facilitating each other. For example, reference [74] NER system incorporates both CRF and rule-based approaches, and the combination of these two methods improves the efficiency as well as the accuracy of entity recognition. Reference [75] trained the domain-independent NER model to perform the entity recognition task by combining two machine learning methods, SVM and HMM, together. Reference [76] proposes a NER system dedicated to tweets by combining CRF and clustering-based methods through a two-stage approach to cope with the characteristics of tweets text. The multi-technology fusion approach is mainly designed to explore technology combi-nation methods to construct models with higher accuracy compared to rule-based and machine learning-based methods alone through suitable combination methods.

To summarize, at the beginning of the NER concept, since there was no mature entity recognition technology at this time, therefore, the commonly used method at this time relied heavily on the rules hand-coded by experts. In addition, the total number of NER datasets created in this period was relatively limited, the dataset size was quite small, the relevant corpus was generally news texts, and various methods based on machine learning were not yet widely used for such tasks. Therefore, taking these factors together, it is possible to state that the early stages of NER development were dominated by rule-based methods and that the size of the relevant dataset was suitable for rule-based method studies. In other words, rules elaborated by experts are likely to cover the dataset exhaustively. With the development of technology, attempts based on machine learning methods soon emerged and some datasets such as MET-1, MET-2, provided data support for relevant researchers to verify the portability of their machine learning-based systems. If suitable training datasets are available, it is reasonable to assert that systems trained on them can be ported to datasets in different languages and even in other research domains [70]. In addition, the emergence of machine learning-based methods has greatly saved the cost of manual annotation and has flexible portability, but since the accuracy of their supervised learning-based models is significantly limited by the large-scale high-quality annotated NER dataset, researchers have explored machine learning algorithm-based models while still maintaining their enthusiasm for rule-based methods, and the subsequent NER systems have also partially incorporated artificial rules. Moreover, since both rule-based and machine-learning-based approaches have their advantages, researchers have immediately introduced multi-technology fusion approaches. The datasets used in these methods are partly constructed by researchers themselves for their different research purposes. In conclusion, the continuous development of NER datasets has contributed to the diversification of NER techniques, and the demand for accuracy of NER techniques has in turn contributed to the continuous development of NER datasets.

5 Conclusion and Future Direction

This paper surveys the literature on the creation of NER datasets, profiling common NER datasets created at different times, in different conferences, and in different tasks. Since MUC-6 proposed the NER task, the development of the NER dataset over the years has been sorted out from the dimensions of the language used, the research domain, the type of entity, the granularity of the entity, and the entity annotation. A review of the evolution of datasets over the years and the different aspects mentioned above can provide ideas for the future development of datasets. In future NER research, there will be many large fine-grained datasets with high-quality entity annotations created to perform named entity recognition tasks.

Based on the above research work and development trends of NER datasets, we suggest following three future directions for NER datasets. (1) In terms of the research area to which the dataset belongs, in the future NER datasets will not only be created based on the research area of the researchers, but also more likely to explore new fields that have never been researched or to experiment with fields that may have industrial and commercial value, thus filling the gaps in the development of NER datasets. In addition, for a dataset in determined domains, its entity type can be defined against the terminology of the domain. (2) In terms of entity granularity, the delineation of fine- grained entities will inevitably emerge in future research. This is mainly due to the nature of the named entity recognition task. As named entity recognition is a fundamental part of many applications, the coarseness of its granularity will have a direct impact on the accuracy of the application. Therefore, fine-grained named entity recognition datasets will be the first to be considered by researchers, but coarse-grained named entity recognition datasets will also exist. (3) For entity annotation, reducing manual involvement and improving the accuracy and consistency of annotations will be the goal of the dataset creation. If researchers want to train a more accurate NER model, then the dataset must have both large-scale and high-quality annotation. However, if the size of the dataset is large, then full manual annotation is unlikely to be possible, and therefore research into suitable entity annotation schemes is also an important direction for the future development of NER datasets.

In general, as NER research continues to evolve, new languages and domains will be covered, and the range of entity types to be recognized by the NER task will become more diverse and the granularity of entity will become more refined.

Acknowledgements The authors would like to thank the Institute of Systems Engineering related to this work for their support.

Author Contributions Ying Zhang designed and performed the experiments and analyzed the data. Ying Zhang wrote the manuscript in consultation with Gang Xiao.

Funding This work was supported by the National Key Laboratory for Complex Systems Simulation Foundation (NO.6142006190301).

Availability of Data and Materials The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification[J]. Lingvisticae Investigationes **30**(1), 3–26 (2007)
2. Grishman, R., Sundheim, B.M.: Message understanding conference-6: A brief history[C]. Coling: The 16th International Conference on Computational Linguistics **1** (1996)
3. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models[J]. arXiv preprint arXiv:1910.11470 (2019.
4. Goyal, A., Gupta, V., Kumar, M.: Recent named entity recognition and classification techniques: a systematic review[J]. Comput. Sci. Rev. **29**, 21–43 (2018)
5. Li, J., Sun, A., Han, J., et al.: A survey on deep learning for named entity recognition[J]. IEEE Trans. Knowl. Data Eng.Knowl. Data Eng. **34**(1), 50–70 (2020)
6. Mandl, T., Womser-Hacker, C.: The effect of named entities on effectiveness in cross-language information retrieval evaluation[C]. Proceedings of the 2005 ACM symposium on applied computing. 1059–1064 (2005)
7. Guo, J., Xu, G., Cheng, X., et al.: Named entity recognition in query[C]//Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 267–274 (2009)
8. Petkova, D., Croft, W. B.: Proximity-based document representation for named entity retrieval[C]. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 731–740 (2007)
9. Mollá, D., Van Zaanen, M., Smith, D.: Named entity recognition for question answering[C]. Proc. Australas. Lang. Technol. Workshop **2006**, 51–58 (2006)
10. Pizzato, L.A., Mollá, D., Paris, C.: Pseudo relevance feedback using named entities for question answering[C]. Proc. Australas. Lang. Technol. Workshop **2006**, 83–90 (2006)
11. Babych, B., Hartley, A.: Improving machine translation quality with automatic named entity recognition[C]. Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003. (2003)
12. Zhang, Z., Han, X., Liu, Z., et al.: ERNIE: Enhanced language representation with informative entities[J]. arXiv preprint arXiv:1905.07129 (2019)
13. Cheng, P., Erk, K.: Attending to entities for better text understanding[C]. Proc. AAAI Confer. Artific. Intellig. **34**(05), 7554–7561 (2020)
14. Nobata, C., Sekine, S., Isahara, H., et al.: Summarization System Integrated with Named Entity Tagging and IE pattern Discovery[C]. LREC (2002)
15. Aone C.: A trainable summarizer with knowledge acquired from robust nlp techniques[J]. Adv. Autom. Text Summariz. 71–80 (1999)
16. Bach, N., Badaskar, S.: A review of relation extraction[J]. Literat. Rev. Lang. Statist. **II**(2), 1–15 (2007)
17. Gundluru, N., Rajput, D. S., Lakshmanna, K., Kaluri, R., Shorfuzzaman, M., Uddin, M., & Rahman Khan, M. A. (2022). Enhancement of Detection of Diabetic Retinopathy Using Harris Hawks Optimization with Deep Learning Model. *Computational Intelligence and Neuroscience*, *2022*.
18. Kumar S.: A survey of deep learning methods for relation extraction[J]. arXiv preprint arXiv:1705.03645 (2017)
19. Getoor, L., Machanavajjhala, A.: Entity resolution: theory, practice & open challenges[J]. Proc. VLDB Endowment **5**(12), 2018–2019 (2012)
20. Zhao, J.: A survey on named entity recognition, disambiguation and cross-lingual co-reference resolution[J]. J. Chinese Inform. Process. **23**(2), 3–17 (2009)
21. Merchant, R., Okurowski, M.E., Chinchor, N. :The multilingual entity task (MET) overview[R]. Department of Defense Fort George G Meade MD (1996)
22. Sekine, S., Isahara, H.: IREX: IR & IE evaluation project in Japanese[C]. LREC. 1977–1980 (2000)
23. Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition[J]. arXiv preprint cs/0306050 (2003)
24. Doddington, G.R., Mitchell, A., Przybocki, M.A., et al.: The automatic content extraction (ace) program-tasks, data, and evaluation[C]. Lrec. **2**(1), 837–840 (2004)
25. Kim J.D., Ohta, T., Tateisi, Y., et al.: GENIA corpus—a semantically annotated corpus for bio-textmining[J]. Bioinformatics, **19**(suppl_1): i180-i182 (2003)
26. Kim, J.D., Ohta, T., Tateisi, Y., et al.: GENIA corpus manual-encoding schemes for the corpus and annotation[J]. Date of Release **15** (2006)
27. Faessler, E., Modersohn, L., Lohr, C., et al.: ProGene-A large-scale, high-quality protein-gene annotated benchmark corpus[C]. Proceedings of the 12th Language Resources and Evaluation Conference. 4585–4596 (2020)
28. Marcus, R., Palmer, M., Ramshaw, R.B.S.P.L., et al.: Ontonotes: a large training corpus for enhanced processing[J]. Joseph Olive, Caitlin Christianson, and John McCary, editors, Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation (2011)
29. Wei, C.H., Peng, Y., Leaman, R., et al.: Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task[J]. Database (2016)
30. Strauss, B., Toma, B., Ritter, A., et al.: Results of the wnut16 named entity recognition shared task[C]. Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 138–144 (2016)
31. Augenstein, I., Das, M., Riedel, S., et al.: Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications[J]. arXiv preprint arXiv:1704.02853 (2017)
32. Buscaldi, D., Schumann, A.K., Qasemizadeh, B., et al.: Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers[C]. Proceedings of the 12th international workshop on semantic evaluation. 679–688 (2018)
33. Tanabe, L., Xie, N., Thom, L.H., et al.: GENETAG: a tagged corpus for gene/protein named entity recognition[J]. BMC Bioinform. **6**(1), 1–7 (2005)

34. Ohta, T., Kim, J.D., Pyysalo, S., et al.: Incorporating GENETAG-style annotation to GENIA corpus[C]. Proceedings of the BioNLP 2009 Workshop. 106–107 (2009)
35. Weischedel, R., Brunstein, A.: BBN pronoun coreference and entity type corpus[J], p. 112. Linguistic Data Consortium, Philadelphia (2005)
36. Balasuriya, D., Ringland, N., Nothman, J., et al.: Named entity recognition in wikipedia[C]. Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (People's Web). 10–18 (2009)
37. Ghaddar, A., Langlais, P.: Winer: A wikipedia annotated corpus for named entity recognition[C]. Proceedings of the Eighth International Joint Conference on Natural Language Processing **1**: 413–422 (2017)
38. Lakshmanna, K., Khare, N.: Mining dna sequence patterns with constraints using hybridization of firefly and group search optimization. J. Intell. Syst.Intell. Syst. **27**(3), 349–362 (2018)
39. Röder, M., Usbeck, R., Hellmann, S., et al.: N^3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format[C]//Proceedings of the ninth international conference on language resources and evaluation (LREC'14). 3529–3533 (2014)
40. Luan, Y., He, L., Ostendorf, M., et al.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction[J]. arXiv preprint arXiv:1808.09602 (2018)
41. Ringland, N., Dai, X., Hachey, B., et al.: NNE: A dataset for nested named entity recognition in english newswire[J]. arXiv preprint arXiv:1906.01359 (2019)
42. Wang, Z., Shang, J., Liu, L., et al.: Crossweigh: Training named entity tagger from imperfect annotations[J]. arXiv preprint arXiv:1909.01441 (2019)
43. Jain, N., Sierra, A., Ehmueller, J., et al.: Generation of Training Data for Named Entity Recognition of Artworks[J]
44. Sahin, H.B, Tirkaz, C., Yildiz, E., et al.: Automatically annotated turkish corpus for named entity recognition and text categorization using large-scale gazetteers[J]. arXiv preprint arXiv:1702.02363 (2017)
45. Fu, R., Qin, B., Liu, T.: Generating Chinese named entity data from parallel corpora[J]. Front. Comp. Sci. **8**(4), 629–641 (2014)
46. Wang, X., Jiang, Y., Bach, N., et al.: Automated concatenation of embeddings for structured prediction[J]. arXiv preprint arXiv:2010.05006 (2020)
47. Linguistic Data Consortium. Annotation guidelines for entity detection and tracking(edt), version 4.2. 6 200400401[J]. http://www.ldc.upenn.edu/Projects/ACE/docs/EnglishEDTV4–2–6. PDF–Zugriff am, **4** (2004)
48. Lakshmanna, K., Khare, N.: FDSMO: frequent DNA sequence mining using FBSB and optimization. Int. J. Intellig. Eng. Syst. **9**(4), 157–166 (2016)
49. Zhong, Z., Chen, D.: A frustratingly easy approach for entity and relation extraction[J]. arXiv preprint arXiv:2010.12812 (2020)
50. Yu, J., Bohnet, B., Poesio, M.: Named entity recognition as dependency parsing[J]. arXiv preprint arXiv:2005.07150 (2020)
51. Lee, J., Yoon, W., Kim, S., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics **36**(4), 1234–1240 (2020)
52. Ringland, N.: Structured Named Entities[J]. (2015)
53. Leaman, R,, Miller, C., Gonzalez, G.: Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark[C]. Proceedings of the 2009 Symposium on Languages in Biology and Medicine **82**(9): 82–89 (2009)
54. Chowdhury, M.F.M., Lavelli, A.: Disease mention recognition with specific features[C]. Proceedings of the 2010 workshop on biomedical natural language processing. 83–90 (2010)
55. Pradhan, S.S., Hovy, E., Marcus, M., et al.: Ontonotes: A unified relational semantic representation[C]. International Conference on Semantic Computing (ICSC 2007). IEEE, 517–526 (2007)
56. Ngonga Ngomo, A.C., Heino, N., Lyko, K., et al.: Scms–semantifying content management systems[C]. International Semantic Web Conference. Springer, Berlin, Heidelberg 189–204 (2011)
57. Hellmann, S., Lehmann, J., Auer, S., et al.: Integrating NLP using linked data[C]. International semantic web conference. Springer, Berlin, Heidelberg 98–113 (2013)
58. Palmer, D.D., Day, D.: A statistical profile of the named entity task[C]. Fifth Conference on Applied Natural Language Processing. 190–193 (1997)
59. Tanabe, L., Wilbur, W.J.: Tagging gene and protein names in biomedical text[J]. Bioinformatics **18**(8), 1124–1132 (2002)
60. Dogan, R.I., Lu, Z.: An improved corpus of disease mentions in PubMed citations[C]. BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. 91–99 (2012)
61. Derczynski, L., Nichols, E., van Erp, M., et al.: Results of the WNUT2017 shared task on novel and emerging entity recognition[C]. Proceedings of the 3rd Workshop on Noisy User-generated Text. 140–147 (2017)
62. Sekine, S.: Named entity: History and future[J]. Project notes, New York University **4** (2004)
63. Appelt, D.E., Hobbs, J.R., Bear, J., et al.: FASTUS: A finite-state processor for information extraction from real-world text[C]//IJCAI. **93**: 1172–1178 (1993)
64. Appelt, D., Hobbs, J.R., Bear, J., et al.: SRI International FASTUS systemMUC-6 test results and analysis[C]. Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6–8, 1995. (1995)
65. Gaizauskas, R., Wakao, T., Humphreys, K., et al.: University of Sheffield: Description of the LaSIE system as used for MUC-6[C]//Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6–8, 1995. (1995)
66. Humphreys, K., Gaizauskas, R., Azzam, S., et al.: University of Sheffield: Description of the LaSIE-II system as used for MUC-7[C]. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. (1998)
67. Black, W.J., Rinaldi, F., Mowatt, D.: FACILE: Description of the NE system used for MUC-7[C]. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. (1998)
68. Aone, C., Halverson, L., Hampton, T., et al.: SRA: Description of the IE2 system used for MUC-7[C]. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. (1998)
69. Bikel, D.M., Miller, S., Schwartz, R., et al.: Nymble: a high-performance learning name-finder[J]. arXiv preprint cmp-lg/9803003 (1998)
70. Borthwick, A., Sterling, J., Agichtein., E, et al.: NYU: Description of the MENE named entity system as used in MUC-7[C]. Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. (1998)
71. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. (2001)
72. McNamee, P., Mayfield, J.: Entity extraction without language-specific resources[C]. COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). (2002)
73. Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms[C]. International Conference on Discovery Science. Springer, Berlin, Heidelberg 267–278 (2006)

74. Guanming, Z., Chuang, Z., Bo, X., et al.: CRFs-based Chinese named entity recognition with improved tag set[C]. 2009 WRI World Congress on Computer Science and Information Engineering. IEEE **5**, 519–522 (2009)
75. Atkinson, J., Bull, V.: A multi-strategy approach to biological named entity recognition[J]. Expert Syst. Appl. **39**(17), 12968–12974 (2012)
76. Liu, X., Zhou, M.: Two-stage NER for tweets with clustering[J]. Inf. Process. Manage. **49**(1), 264–273 (2013)
77. Liu, H., Qiu, Q., Wu, L., et al.: Few-shot learning for name entity recognition in geological text based on GeoBERT[J]. Earth Science Informatics 1–13 (2022)
78. VeeraSekharReddy, B., Rao, K.S., Koppula, N.: Enhanced Conditional Random Field-Long Short-Term Memory for Name Entity Recognition in English Texts[J]. (2022)
79 Liu, Z., Xu, Y., Yu, T., et al.: Crossner: Evaluating cross-domain named entity recognition. Proc. AAAI Confer. Artific. Intellig. **35**(15), 13452–13460 (2021)
80. Ding, N., Xu, G., Chen, Y., et al.: Few-nerd: A few-shot named entity recognition dataset[J]. arXiv preprint arXiv:2105.07464 (2021)
81. Jain, S., Agrawal, A., Saporta, A., et al.: Radgraph: Extracting clinical entities and relations from radiology reports[J]. arXiv preprint arXiv:2106.14463 (2021)
82. Hvingelby, R., Pauli, A.B, Barrett, M., et al.: DaNE: A named entity resource for danish[C]//Proceedings of the 12th language resources and evaluation conference. 4597–4604 (2020)
83. Wróblewska, A., Kaliska, A., Pawłowski, M., et al.: TASTEset--Recipe Dataset and Food Entities Recognition Benchmark[J]. arXiv preprint arXiv:2204.07775 (2022)
84. Au, T.W.T., Cox, I.J., Lampos, V.: E-NER--an annotated named entity recognition corpus of legal text[J]. arXiv preprint arXiv:2212.09306 (2022)
85. Tabassum, J., Lee, S., Xu, W., et al.: WNUT-2020 task 1 overview: Extracting entities and relations from wet lab protocols[J]. arXiv preprint arXiv:2010.14576 (2020)
86. Malmasi, S., Fang, A., Fetahu, B., et al.: Multiconer: a large-scale multilingual dataset for complex named entity recognition[J]. arXiv preprint arXiv:2208.14536 (2022)
87. Mayhew, S., Blevins, T., Liu, S., et al.: Universal NER: A gold-standard multilingual named entity recognition benchmark[J]. arXiv preprint arXiv:2311.09122 (2023)
88. Schindler, D., Bensmann, F., Dietze, S., et al.: Somesci-A 5 star open data gold standard knowledge graph of software mentions in scientific articles[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 4574–4583 (2021)
89. Adelani, D.I., Abbott, J., Neubig, G., et al.: MasakhaNER: Named entity recognition for African languages[J]. Trans. Assoc. Comput. Linguist. **9**, 1116–1131 (2021)
90. Bareket, D., Tsarfaty, R.: Neural modeling for named entities and morphology (nemoˆ2)[J]. Trans. Assoc. Comput. Linguist. **9**, 909–928 (2021)
91. Park, S., Moon, J., Kim, S., et al.: Klue: Korean language understanding evaluation[J]. arXiv preprint arXiv:2105.09680 (2021)
92. Marek, P., Müller, Š., Konrád, J., et al.: Text summarization of czech news articles using named entities[J]. arXiv preprint arXiv:2104.10454 (2021)
93. Păiş, V., Mitrofan, M., Gasan, C.L, et al.: Named entity recognition in the Romanian legal domain[C]//Proceedings of the Natural Legal Language Processing Workshop 2021. 9–18 (2021)
94. Yeshpanov, R., Khassanov, Y., Varol, H.A.: KazNERD: Kazakh named entity recognition dataset[J]. arXiv preprint arXiv:2111.13419 (2021)
95. Murthy, R., Bhattacharjee, P., Sharnagat, R., et al.: HiNER: a large hindi named entity recognition dataset[J]. arXiv preprint arXiv:2204.13743 (2022)
96. Hennig, L., Truong, P.T., Gabryszak, A.: Mobie: A german dataset for named entity recognition, entity linking and relation extraction in the mobility domain[J]. arXiv preprint arXiv:2108.06955 (2021)
97. Paccosi, T., Aprosio, A.P.: KIND: an Italian Multi-Domain Dataset for Named Entity Recognition[J]. arXiv preprint arXiv:2112.15099 (2021)
98. Mhaske, A., Kedia, H., Doddapaneni, S., et al.: Naamapadam: a large-scale named entity annotated data for indic languages[J]. arXiv preprint arXiv:2212.10168 (2022)
99. Sameen Shahgir, H.A.Z., Alam, R., Alam, M.Z.U.: BanglaCoNER: Towards Robust Bangla Complex Named Entity Recognition[J]. arXiv e-prints, arXiv: 2303.09306 (2023)

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