Artificial Intelligence Review (2025) 58:315
https://doi.org/10.1007/s10462-025-11321-8

A review of named entity recognition: from learning methods to modelling paradigms and tasks

Wei Liang Seow[1] · Iti Chaturvedi[2] · Amber Hogarth[2] · Rui Mao[1] · Erik Cambria[1]

Accepted: 2 July 2025

Abstract

Named Entity Recognition (NER) is commonly used when summarising news articles and legal documents. It can extract the names of politicians or organisations and help determine the aspect of a positive or negative sentiment. Previous surveys have only provided a shallow review of NER with respect to a certain datatype. In contrast, here a much deeper coverage of different approaches is provided. First articles with respect to the learning method are discussed, such as supervised or unsupervised. Next, popular models that combine two or more learning methods are introduced in a bottom-up approach. The most popular NER algorithms are compared on a recently crawled 2024 election dataset from Australia. The effect of different parameters such as number of epochs and learning rate is explored. It is concluded that pre-trained NER models are limited in their ability to model new entities and disambiguate their context. Using the sentiment score together with a state space model over entities in a sentence might help overcome these challenges.

Keywords Natural language processing · Named entity recognition · Survey

✉ Iti Chaturvedi
iti.chaturvedi@jcu.edu.au

Wei Liang Seow
weiliang003@e.ntu.edu.sg

Amber Hogarth
amber.hogarth@my.jcu.edu.au

Rui Mao
rui.mao@ntu.edu.sg

Erik Cambria
cambria@ntu.edu.sg

[1] College of Computing and Data Science, Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore, Singapore

[2] College of Science and Engineering, James Cook University, 1 James Cook Dr, Townsville, QLD 4811, Australia

Published online: 16 July 2025

1 Introduction

Named Entity Recognition (NER) is a sub-task in Natural Language Processing (NLP) that focuses on identifying different types of entities within a text and categorising them into predefined classes (Mao et al. 2024c; Zhong and Cambria 2021). Common named entity categories include person names (PER), location names (LOC), and organisation names (ORG). The NER task typically involves analysing a sequence of words or tokens and assigning labels to those corresponding to named entities. NER is crucial for information extraction and supports various downstream tasks, including event extraction (Xiang and Wang 2019), text summarisation (Gambhir and Gupta 2017), relation extraction (Nasar et al. 2021), question answering (Khalid et al. 2008), and machine translation (Ugawa et al. 2018). Figure 1 illustrates a chronological summary of state-of-the-art NER techniques focused on in this survey.

Current deep learning approaches employ artificial neural networks to automatically learn complex features from the training data with non-linear activation functions in an end-to-end manner, improving accuracy on NER tasks. Despite advances in deep neural networks, NER still presents various challenges, including expensive data annotation, diverse languages, diverse domains, multiple modalities, nested/discontinuous/overlapping entities, low-resource languages/domains, and fine-grained entities. Moreover, deep learning-based NER models are predominantly black box models that suffer from interpretability and explainability issues (Cambria et al. 2023). Other challenges include continual learning of new entities in real-world scenarios such as virtual assistants and class-imbalance between “Others" entity tags and specific labelled entity tags. Despite these numerous challenges, NER is still a task worth studying as it is important in information retrieval, text understanding, automated data extraction and many downstream tasks.

The most recent NER surveys are domain-, task- or language-specific, e.g. NER from historical documents (Ehrmann et al. 2023), NER in Turkish legal texts (Küçük et al. 2017), and Clinical NER (Navarro et al. 2023), Chinese NER (Liu et al. 2022c), Joint NER and Relation extraction (Kambar et al. 2022), few-shot NER (Huang et al. 2021b; Moscato

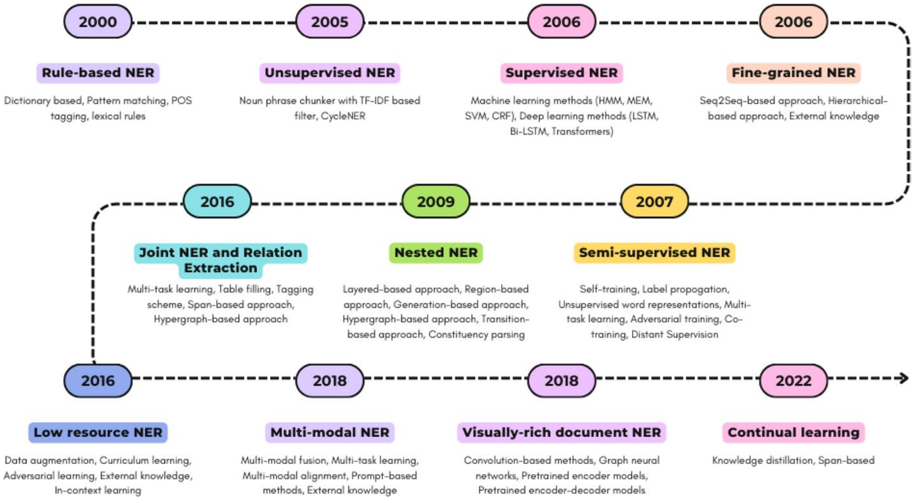
2000
Rule-based NER
Dictionary based, Pattern matching, POS tagging, lexical rules
2005
Unsupervised NER
Noun phrase chunker with TF-IDF based filter, CycleNER
2006
Supervised NER
Machine learning methods (HMM, MEM, SVM, CRF), Deep learning methods (LSTM, Bi-LSTM, Transformers)
2006
Fine-grained NER
Seq2Seq-based approach, Hierarchical-based approach, External knowledge
2007
Semi-supervised NER
Self-training, Label propogation, Unsupervised word representations, Multi-task learning, Adversarial training, Co-training, Distant Supervision
2009
Nested NER
Layered-based approach, Region-based approach, Generation-based approach, Hypergraph-based approach, Transition-based approach, Constituency parsing
2016
Joint NER and Relation Extraction
Multi-task learning, Table filling, Tagging scheme, Span-based approach, Hypergraph-based approach
2016
Low resource NER
Data augmentation, Curriculum learning, Adversarial learning, External knowledge, In-context learning
2018
Multi-modal NER
Multi-modal fusion, Multi-task learning, Multi-modal alignment, Prompt-based methods, External knowledge
2018
Visually-rich document NER
Convolution-based methods, Graph neural networks, Pretrained encoder models, Pretrained encoder-decoder models
2022
Continual learning
Knowledge distillation, Span-based

Fig. 1 Chronological order of the progress in NER from the year 1995 to the year 2023

et al. 2023b), nested NER (Wang et al. 2022f), multi-modal NER (Qian et al. 2023) and low resource NER (Tang et al. 2023a). To our knowledge, the latest comprehensive NER surveys are from Nasar et al. (2021) and Li et al. (2020c). Nasar et al. (2021) provided an in-depth study of state-of-the-art information extraction, including NER and Relation extraction with emphasis on deep learning approaches. They also survey prominent methods for the latest applications of deep learning in new NER problem settings and use cases. However, these works have a simpler taxonomy compared to ours and do not cover the different modelling paradigms and other recent NER trends, such as visually-rich document NER (Vrd-NER), continual learning NER and open-vocabulary NER.

Previous surveys have been predominantly written in a top-down approach, which is domain-, task- or language-specific. In contrast, our survey is written from a bottom-up approach and covers a wide taxonomy. Other areas that are covered which are not in previous surveys include theoretical research for NER, different NER modelling paradigms and a comprehensive review of different NER tasks, including the latest trends such as Vrd-NER, continual learning NER and open-vocabulary NER. Vrd-NER aims to extract entities from semi-structured documents such as forms, receipts, and invoices which involves different modalities such as text, image and layout. Continual learning NER is the ongoing process of acquiring knowledge about new entities using the NER model in real-world settings such as voice-enabled assistants. Open-vocabulary NER aims to train an NER model to be capable of recognising entities of any new type based on their textual names or descriptions.

The contribution of this survey is summarised below:

- A review of the latest NER techniques from multiple perspectives, e.g., learning methods, modelling paradigms, and data diversity (e.g., low-resource, Vrd-NER, multi-modal NER, cross-lingual NER, cross-domain NER). The taxonomy of this survey is more systematic than previous NER surveys in representing the different technical trends.
- A critical analysis of the advantages and disadvantages of different works amongst different technical trends, summarising the main challenges in this domain.

The survey is structured as follows: firstly, we review the taxonomy of NER techniques (see Sect. 2), theoretical research and different learning methods of NER (see Sect. 3). Next, different NER modelling paradigms (see Sect. 4) are presented from traditional models such as Sequence labelling to more recent models such as prompt based learning. Moving on, we analyse the different NER datasets and evaluation metrics in (see Sect. 5). Furthermore, different techniques employed in common NER tasks (see Sect. 6) and other NER approaches (see Sect. 7) were surveyed. In Sect. 8, we evaluate the accuracy and parameter settings of BERT on QLD 2024 Twitter Election dataset. Finally, different challenges faced in NER were discussed (see Sect. 9) and this survey is concluded in Sect. 10.

2 The taxonomy of NER techniques

Our study involves a comparative analysis with previous surveys within the domain of taxonomy. Establishing a suitable taxonomy within an NER survey is of paramount importance. This taxonomy offers a well-organised framework for classifying and structuring the dynamic landscape of relevant techniques, methodologies, and applications. Serving as a valuable guide for researchers, it facilitates comprehension and navigation through the intricate domain of NER. Moreover, practitioners benefit from this taxonomy as it enables the swift identification of the most pertinent technique for a particular task, enhancing efficiency in their application of NER methods.

Figure 2 shows our NER taxonomy, and Table 1 shows our NER taxonomy compared to other surveys. Our survey covers finer-grained NER taxonomy. We not only categorise NER techniques into different learning methods, e.g. rule-based, supervised, unsupervised, semi-supervised and weakly-supervised learning, but also ground the category based on different techniques, e.g. self-training and multitask learning. We also classify NER taxonomy based on different modelling paradigms and NER tasks.

Previous surveys provide a shallow, top-down review of NER algorithms. Here, different challenges on certain types of data are first introduced, and then training methods for each challenge are discussed. In contrast, our survey provides a much deeper coverage of different approaches without focusing on challenges on a specific dataset (Chaturvedi et al. 2018). Figure 3 compares our taxonomy in Fig. 2 with two recent surveys on NER. The first survey by Ehrmann et al. (2023) considers approaches specific to historical documents where optical character recognition is required to extract text from images. It discusses challenges in visually rich documents such as noisy images, spelling errors and domain bias in training models. Finally, it explores the use of deep learning to overcome these challenges. The second survey in Fig. 3 is by Nasar et al. (2021), and focuses on named entities in different news articles benchmark datasets. This survey suggested that news articles can be in English, low-resource foreign languages and different domains. For each of these three types, they surveyed training approaches that are neural (deep learning), non-neural and hybrid of both the former types.

3 Learning methods

3.1 Theoretical research

NER is an NLP sub-task involving the identification of specific mentions within a given text and classifying them into predefined entity types, including person, location, or organisation. Experts from diverse fields hold varying perspectives on how a named entity is defined. In this section on Theoretical Research, the definition of named entities based on two different perspectives: unique identification (Computer Science) and domain of application are reviewed. In the field of Computer Science, named entities are defined using the concept of unique identification. MUC conferences require that expressions annotated in text are the "unique identifier" of entities. However, Marrero et al. (2013) pointed out that this approach brings about ambiguity and subjectivity, as not all entities can be uniquely identified. Additionally, the same mention within a text may sometimes refer to different entity categories in various contexts, further complicating the process of accurate NER. From the perspective of Marrero et al. (2013), named entities are defined based on the specific domain of application, such as defence and biomedical domains. Under this definition, an entity labeled in a particular domain may be a non-entity in another domain. This could result in inconsistency in the labelling of entities across different domains, leading to confusion during model learning and continual learning of new entities.

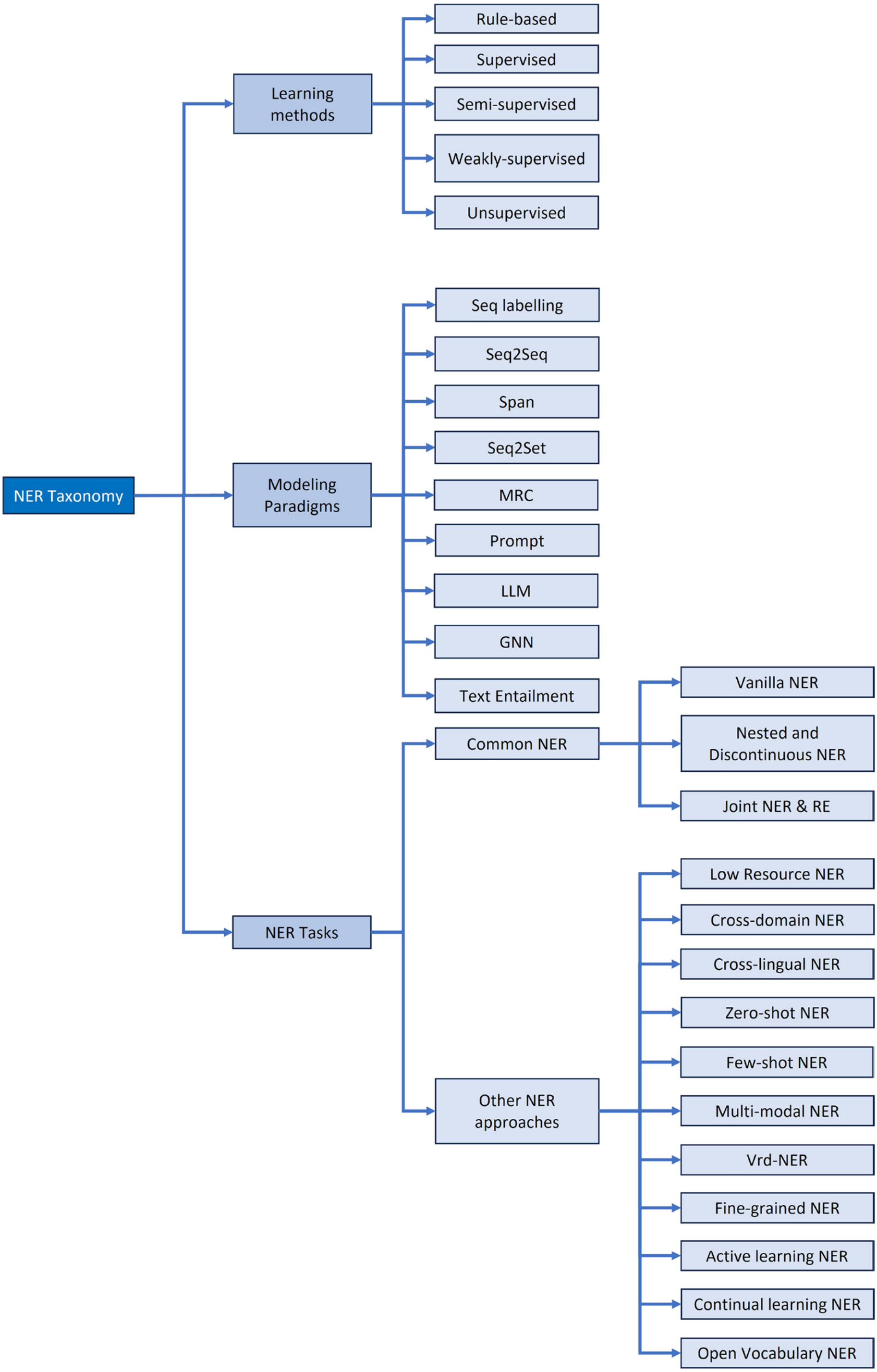

Fig. 2 NER taxonomy for this survey based on different learning methods, modelling paradigms and NER tasks

Table 1 Comparison of the NER taxonomy versus other surveys based on different learning methods, modelling paradigms and NER tasks

Taxonomy	Nasar et al. (2021)	Li et al. (2020c)	Ours
Rule-based	✓	✓	✓
Supervised learning	✓	✓	✓
Semi-supervised learning	✓		✓
Weakly-supervised learning			✓
Unsupervised learning	✓	✓	✓
Distant Supervision			✓
Self-training			✓
Multi-task learning		✓	✓
Seq labelling			✓
Seq2seq			✓
Span			✓
Seq2set			✓
MRC			✓
Prompt			✓
LLM			✓
GNN			✓
Text entailment			✓
Vanilla NER	✓		✓
Nested NER	✓		✓
Discontinuous NER			✓
Joint NER-RE			✓
Low resource			✓
Cross-lingual NER	✓		✓
Cross-domain NER			✓
Zero-shot NER			✓
Few-shot NER			✓
Continual learning			✓
Active learning		✓	✓
Multi-modal			✓
Vrd-NER			✓
Fine-grained NER			✓
Open Vocabulary			✓

Seq sequence, *Seq2Seq* sequence to sequence, *Seq2Set* sequence to set, *MRC* machine reading comprehension, *LLM* large language models, *GNN* graph neural networks, *Vrd-NER* visual-rich document NER, *Joint NER-RE* joint NER and RE

3.2 Rule based

Rule-driven methods in NER rely on establishing patterns, structures, or linguistic guidelines to extract entities from text. Quimbaya et al. (2016) employed a dictionary-based method to extract named entities from electronic health records (EHR). More recently, Popovski et al. (2019) used computational linguistics and semantic knowledge to extract food-related entities through rule-based techniques. In general, rule-based methods are durable and easy to understand and interpret, allowing quick implementation and modification of rules. However, approaches based on handcrafted rules are often rigid and do not generalise well.

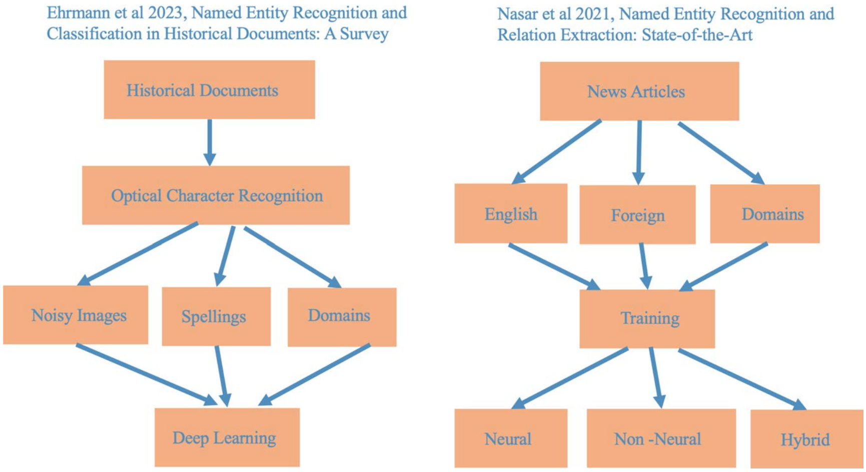
Ehrmann et al 2023, Named Entity Recognition and Classification in Historical Documents: A Survey
Historical Documents
Optical Character Recognition
Noisy Images
Spellings
Domains
Deep Learning
Nasar et al 2021, Named Entity Recognition and Relation Extraction: State-of-the-Art
News Articles
English
Foreign
Domains
Training
Neural
Non -Neural
Hybrid

Fig. 3 Comparison of taxonomy of two previous surveys on NER with ours provided in Fig. 2

3.3 Supervised learning

Deep learning methods utilise deep neural networks and non-linear activation functions to automatically learn intricate features from training data. Consequently, this reduces the need for manually-crafted, human-engineered features that require domain expertise. Deep learning employs forward propagation to compute predictions, backward propagation to calculate gradients for neural network weights, and updates weights based on the loss function gradient using gradient descent. Typically, a deep learning NER model consists of a word embedding layer, responsible for representing words as vectors, a context encoder to model contextual dependencies, and a tag decoder, e.g. softmax or Conditional Random Field (CRF), to predict individual tags for each word token in the input. Typical encoder models employed in deep learning include Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), Bidirectional Long Short-Term Memory (Bi-LSTM) (Ma and Hovy 2016; Lample et al. 2016) and Transformers (Vaswani et al. 2017), which are further elaborated later.

Early deep learning architectures for NER relied on a vanilla LSTM/Bi-LSTM CRF neural network to learn contextual embeddings. Numerous word embedding types such as Count Vector, TF-IDF Vector, Co-occurrence Vector and pre-trained word embeddings, e.g. GloVe (Pennington et al. 2014) and fastText (Bojanowski et al. 2017), act as input to these neural networks. Convolution Neural Network (CNN) can also be used to learn character embeddings. Such input embeddings are concatenated and input to the Bi-LSTM to learn both forward and backward hidden states. These states are then concatenated and input into a CRF layer for entity decoding. For example, Ma and Hovy (2016) presented a Bi-LSTM-CNN-CRF network that used word and character embeddings for end-to-end sequence labelling including NER. Lample et al. (2016) presented two neural architectures relating to Bi-LSTM CRF and Stack-LSTM, a Transition-Based Chunking Model. Stack-LSTM constructs and labels segments through a transition-based approach, drawing inspiration from shift-reduce parsers. Kuru et al. (2016) proposed a NER approach that treats a sentence as a character sequence, using it as input to stacked Bi-LSTMs to generate tag probabilities for each character as output. However, LSTM-based methods process data sequentially, which requires longer processing time and suffers from information loss, despite the reliance on memory/gates to retain information from old-to-new time steps. Although LSTM was invented to solve the issue of vanishing/exploding gradients in Recurrent Neural Networks, it is still prone to this.

Subsequently, Transformer, as explored by Vaswani et al. (2017), was introduced, which surpassed the performance of LSTM/Bi-LSTM techniques. The Transformer has an encoder-decoder architecture which utilises both stacked multi-head attention mechanisms and point-wise, fully connected layers as key components in each transformer layer. The transformer encoder architecture is commonly used for supervised learning in NER. Pre-trained examples of transformer encoders include Bidirectional Encoder Representations from Transformers (BERT) and its multilingual variant, multilingual BERT (M-BERT), as introduced by Devlin et al. (2019). Using masked language modelling and next sentence prediction tasks, BERT is pre-trained; following pre-training, its performance can then be fine-tuned to become state-of-the-art. As M-BERT is only pre-trained based on the top 104 languages in Wikipedia, Wang et al. (2020d) expanded M-BERT to include additional languages by increasing vocabulary and continually pre-training on target languages. Compared to LSTM-based models, transformer-based models offer advantages such as parallel processing, the ability to capture long-range dependencies using multi-head self-attention, and more interpretable attention mechanisms. However, it is also computationally expensive to process long sequences of text with quadratic computational complexity in self-attention.

3.4 Semi-supervised learning

3.4.1 Self-training

Self-training is widely used for semi-supervised NER to mitigate the requirement for large amounts of labelled data. Self-training involves training a teacher model to generate pseudo labels for a set of unlabelled examples, and then re-training the student model on the labelled and pseudo-labelled examples. This procedure is repeated and the current student model is used as a teacher in the next iteration to get pseudo-labels for training another (student) model. Liao and Veeramachaneni (2009) introduced a self-training method that utilises independent evidence separate from classifier features such as multi-mention property and high precision independent context to assign high-precision labels to unlabelled data. This approach automatically extracts data that is both highly accurate and non-redundant, thereby improving the classifier significantly in subsequent iterations. Due to insufficient labelled data for every language, Zafarian et al. (2015) proposed utilising unlabelled bilingual corpora to extract valuable features from transferring information from resource-rich language toward resource-poor language and using it to train a supervised classifier with a small set of labelled data. Finally, a CRF-based supervised classifier is trained using self-training. He and Sun (2017) presented a central model capable of leveraging both out-of-domain corpora and in-domain unannotated texts. First, the cross-domain learning model is utilised to acquire out-of-domain information by considering domain similarity. Then, the semi-supervised learning model focuses on learning from in-domain, unannotated data through self-training.

3.4.2 Multi-task learning

Multi-Task Learning improves a model's generalisation by training it on a main NER task alongside related auxiliary tasks, sharing layers to learn richer, more transferable representations. The goal is to utilise shared model parameters, enabling knowledge acquired from auxiliary tasks to support the main NER task's learning process. Two main types of parameter sharing are hard and soft parameter sharing. Hard parameter sharing involves sharing weights/parameters among multiple tasks to learn a unified representation space. This shared space is used to model different tasks, with task-specific layers added independently for each task. On the other hand, soft parameter sharing promotes similarities among related parameters instead of directly sharing identical parameter values. Each task has its own model, but a constraint is applied to penalise differences between parameters in these models. Soft parameter sharing allows for more flexibility among tasks as it loosely couples shared space representations.

Researchers mainly focus on hard parameter sharing for NER tasks. Clark et al. (2018) applied a semi-supervised learning algorithm centred around Cross-View Training (CVT) with multi-task learning that improves the representations of a Bi-LSTM sentence encoder using a mix of labelled and unlabelled data. It uses standard supervised learning for labelled examples, as shown in Eq. 1 where $\mathcal{D}_l$ represents the labelled dataset and $p_\theta(y|x_i)$ is the output distribution over classes produced by the model with parameters θ on input x_i. For unlabelled examples, auxiliary prediction modules share intermediate representations, minimising the distance function D between probability distributions using KL divergence. This ensures the auxiliary modules align with the primary prediction module on the unlabelled data, as shown in Eq. 2 where $\mathcal{D}_{ul}$ is the unlabelled dataset. The total loss is shown in Eq. 3, representing the summation of loss from both supervised training and cross-view training. Rei (2017) introduced a semi-supervised NER approach employing a Bi-LSTM CRF model for multi-task learning. Their method incorporates sequence labelling as the primary objective and integrates a secondary training objective focused on predicting each word's neighbouring words. In the sequence of words, predicting the next word and previous word stems from forward-moving LSTM and backward-moving LSTM respectively. textcolorblueThe hidden forward and backward representations are then concatenated to predict the final NER tag.

$$\mathcal{L}_{sup}(\theta) = \frac{1}{\mathcal{D}_l} \sum_{x_i, y_i \epsilon \mathcal{D}_l} CE(y_i, p_\theta(y|x_i)) \tag{1}$$

$$\mathcal{L}_{CVT}(\theta) = \frac{1}{\mathcal{D}_{ul}} \sum_{x_i \epsilon \mathcal{D}_{ul}} \sum_{j=1}^{k} D(p_\theta(y|x_i), p_\theta^j(y|x_i)) \tag{2}$$

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{CVT} \tag{3}$$

3.5 Weakly-supervised learning

A common technique used in weakly-supervised learning is distant supervision. This method leverages distant labels obtained from cross-referencing strings in text with a pre-defined entity dictionary such as gazetteers or an external knowledge base. Ritter et al. (2011) proposed a distantly supervised method using Labelled LDA (Ramage et al. 2009) for named entity classification, leveraging open-domain database FreeBase for distant supervision. Labelled LDA is used to capture the distribution of unlabelled entities and their potential types, constrained by each entity's set of possible types from Freebase. Moreover, automatically generated entity labels from distant supervision are often noisy. To address this problem, Shang et al. (2018) introduced two neural models tailored for noisy distant supervision from dictionaries. A modified fuzzy CRF layer is introduced to handle tokens that may have multiple possible labels. Subsequently, AutoNER is proposed, which is based on a Tie or Break scheme which emphasises the connections between neighbouring tokens, determining whether they form part of the same entity mention or are split into two segments. Li et al. (2020f) proposed negative sampling to avoid training NER models with unlabelled entities from distant supervision. To further reduce noise from distant supervision, Yang et al. (2018) developed a method that uses reinforcement learning to create an instance selector, which identifies positive sentences from automatically generated annotations based on distant supervision.

Variations of the Hidden Markov Model (HMM) can be used to enhance the accuracy of noisy annotations acquired through distant supervision. Li et al. (2021d) introduced a Conditional Hidden Markov Model (CHMM) to deduce actual labels from multiple noisy labels obtained from distant supervision using contextual representation abilities of pre-trained language models (PLM). Building on the CHMM, Li et al. (2022c) proposed Sparse Conditional Markov Model (Sparse-CHMM) which focuses on estimating the diagonal of the emission matrix instead of predicting the entire emission matrix, enhancing the accuracy and efficiency of the model.

Other studies integrated self-training for distant-supervised NER to mitigate noisy labels. BOND, introduced by Liang et al. (2020), employed a teacher-student framework to discard distant labels and use pseudo-labels to gradually improve the generalisation ability of the model. Similar to BOND, SCDL (Zhang et al. 2021b) co-trained two teacher-student networks to form inner and outer loops to alleviate label noise. Recently, Qu et al. (2023) proposed ATSEN, a self-training framework, to jointly train two teacher-student networks, promoting comprehensive student learning and implementing a fine-grained student ensemble that updates each segment of the teacher model.

Positive and Unlabelled (PU) learning (Liu et al. 2002) can be used to cope with the low recall score problem in distant supervised NER. Positive and Unlabelled learning trains a binary classifier based on labelled positive data (P) and unlabelled (U) data, where the unlabelled data contains positive or negative samples. The method was extended to multinomial classification for NER tasks. Zhou et al. (2022b) proposed a CONFidence-based Multi-class Positive and Unlabelled learning (Conf-MPU) technique by first performing token-level binary classification to predict the likelihood of each token being a named entity before another neural network model utilises these confidence scores for risk estimation.

Recently, Zhang et al. (2022a) introduced BINDER, a bi-encoder setup tailored for both supervised and distant supervised NER settings and nested and flat NER alike. It employs distinct encoders for entity types and text. BINDER adopts contrastive learning to align text spans and entity types within a shared vector space, treating NER as a representation learning task.

3.6 Unsupervised learning

Unsupervised learning in NER involves training the NER model with unlabelled data. Unlike supervised learning that utilises input text paired with corresponding entity labels, the unsupervised method aims to discern data structures and patterns without explicit human guidance from entity labels. Unsupervised learning techniques are capable of discovering hidden patterns and eliminate manual labelling costs. However, this method suffers from less interpretability and a lack of direct supervision of ground truth labels, leading to lower accuracy. Luo et al. (2020b) presented an NER model relying solely on pretrained word embeddings. It applies a Gaussian Hidden Markov Model (GHHM) and Deep Autoencoding Gaussian Mixture Model (DAGMM) to word embeddings for detecting entity spans and predicting entity types, while a reinforcement learning-based instance selector is used to refine noisy annotations. Iovine et al. (2022) introduced an unsupervised training technique for NER centered around cycle consistency. Their approach employs two functions, sentence-to-entity and entity-to-sentence, without any labelled data for model training. Equation 4 refers to the reconstruction loss, i.e. average cross entropy loss, between input sentence S and generated sentence $S^{'}$ of S-cycle training where p(.) and g(.) represent the real and predicted token probabilities, $|s|$ represents the sentence length, s_i and $s_i^{'}$ are the i-th token in s and $s^{'}$, and $|S|$ is the number of input sentences. Equation 5 refers to the reconstruction loss i.e. average cross entropy loss between input entity sequence Q and generated entity sequence $Q^{'}$ of E-cycle training where p(.) and g(.) represent the real and predicted token probabilities, $|q|$ represents the fixed entity sequence length, q_i and $q_i^{'}$ are the i-th token in q and $q^{'}$, and $|Q|$ is the number of entity sequences. Veena et al. (2023) presented an unsupervised weighted distributional semantics method for entity labelling in the agricultural domain, leveraging an extended BERT model integrated with Latent Dirichlet Allocation (LDA). This integration combines the strengths of both LDA and BERT for enhanced performance.

$$\mathcal{L}_{\phi}(S, S^{'}) = -\frac{\sum_{s \epsilon S} \sum_{i<|s|} p(s_i) \log g(s_i^{'})}{|s| * |S|} \tag{4}$$

$$\mathcal{L}_{\theta}(Q, Q^{'}) = -\frac{\sum_{q \epsilon Q} \sum_{i<|q|} p(q_i) \log g(q_i^{'})}{|q| * |Q|} \tag{5}$$

3.7 Summary

Table 2 summarises the different NER learning methods based on different categories and the positives and negatives of each NER learning technique. The majority of the papers surveyed fall under supervised learning and semi-supervised learning. Rule-based techniques depend on manually created rules. Supervised learning learns NER based on direct supervision of labelled data. Supervised learning approaches can be classified into methods

Table 2 Summary for advantages and disadvantages of different NER learning methods (rule-based, supervised learning and unsupervised learning)

Category	Advantages	Disadvantages
Rule-based	Easy to understand, quick implementation, easy to modify rules and durable	Rigid pre-defined rules, poor generalizability to new data
Supervised learning	Direct access to supervision from labelled data	Large amount of labelled training data required
Unsupervised learning	Discovery of hidden patterns, reduced labelling costs	Lack of ground truth, interpretability and difficult to evaluate
Weakly-supervised learning	Reduce the need for hand-annotated data in supervised training	Noisy labels from weak supervision
Semi-supervised learning	Improved accuracy, reduction in labelling cost by utilising both labelled and unlabelled data	Lower reliability, reliance on selection of less noisy pseudo-labelled data

based on machine learning and those based on deep learning. Various deep learning based methods were covered under supervised learning including LSTM, Bi-LSTM models as well as transformer models such as BERT. Semi-supervised learning utilises little labelled data and abundant unlabelled data for training to reduce the requirement of labelled data. Semi-Supervised methods can be categorised into self-training and multi-task learning. Weakly-supervised learning uses labelling functions to obtain weak labels instead of precise, human-annotated labels. A subtype of weakly-supervised learning is distant supervision using a pre-defined entity dictionary. Lastly, unsupervised learning learns NER based solely on unlabelled data through identifying patterns and structures in the data without explicit guidance from entity labels.

4 Modelling Paradigms

4.1 Sequence labelling

Sequence labelling involves assigning specific labels to each word in a given text sequence as represented in Fig. 4. Equation 6 details that Sequence labelling first encodes the input text into contextualised features with an encoder Enc(.) followed by a decoder Dec(.) to predict the entity labels y_1,..., y_n for each token x_1,..., x_n in the input sequence X. Common tagging schemes for sequence labelling NER include IO (Inside, Outside) tagging, BIO (Begin, Inside, Outside) tagging and BIOES tagging (Beginning, Inside, Outside, Ending, Singleton). Traditionally, sequence labelling NER models are divided into two types: generative models, like HMM, and discriminative models, such as CRF and Maximum Entropy Markov Model (MEMM). An HMM is a Markov model in which observations are dependent on a latent (or “hidden") Markov process. Morwal et al. (2012) employed HMM along with the Viterbi algorithm to decode the most probable sequence of hidden states in an HMM for NER. Despite HMM being capable of learning the joint distribution of words and labels, it can only learn the local context. MEMM considers the relationships among neighbouring states and the entire sequence, exhibiting better expression ability compared to HMM. Alam and Islam (2020) applied an MEMM model coupled with POS tagging for

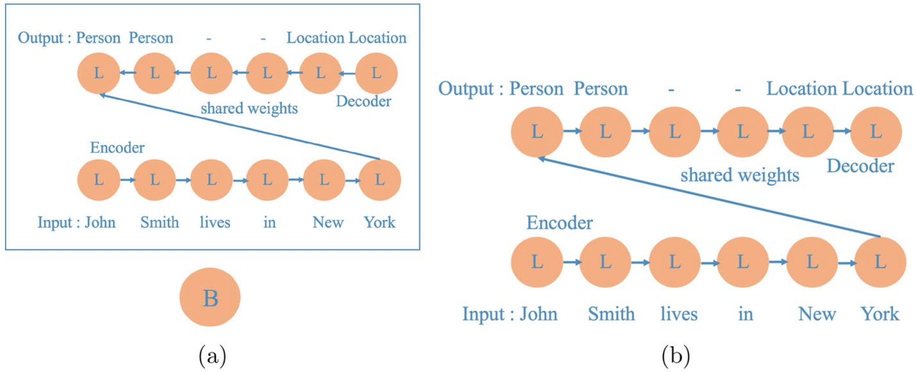
Output : Person Person - - Location Location
L L L L L L
shared weights
Decoder
Encoder
L L L L L L
Input : John Smith lives in New York
B
(a)
Output : Person Person - - Location Location
L L L L L L
shared weights
Decoder
Encoder
L L L L L L
Input : John Smith lives in New York
(b)

Fig. 4 **a** A bi-directional LSTM model (B) for NER. Here each hidden unit is an LSTM (L). The encoder takens the input and the decoder maps to the labels using a CRF (Ren et al. 2023). **b** The decoder in BART does not allow bi-directional edges, instead it is auto-regressive (Cui et al. 2021)

Bengali NER. However, the MEMM model suffers from a labelling bias issue, which is tackled using another discriminative model based on CRF. CRF learns the conditional probability of tagging words with a label and is capable of learning both the global and local context. Liu et al. (2017) leveraged CRF as a core approach to NER. Subsequently, CRF approaches are improved using LSTM and Bi-LSTM encoders, e.g. Bi-LSTM CRF models (Lample et al. 2016) to improve sequence labelling performance.

Recent strides in models such as BERT (Devlin et al. 2019) and Embeddings from Language Models (ELMo) (Peters et al. 2018) attained state-of-the-art performance for downstream NER sequence labelling tasks. Because self-attention in BERT can model dependencies between neighbouring (and even distant) tokens, BERT typically utilises a softmax classifier instead of a CRF classifier. Sequence labelling methods are better at dealing with entities that are long and have low label consistency, and are mainly utilised for flat NER tasks. For more complex tasks, such as nested and discontinuous NER, specialised tagging schemes are required to be designed. The sequence labelling approach still suffers from various limitations. It does not fully leverage label information, treating entity types as one-hot vectors without considering semantics during prediction. Moreover, CRF-based sequence labelling takes into account the context by considering the sequence of labels which can result in error propagation, where a misclassified label influences subsequent classifications.

$$y_1, ..., y_n = \mathrm{Dec}(\mathrm{Enc}(x_1, ..., x_n)) \tag{6}$$

4.2 Sequence-to-sequence based

Recently, the Sequence-to-Sequence based (Seq2Seq) model featuring an architecture built around an encoder-decoder has gained popularity in the field of NER. Unlike sequence labelling methods which are mainly used for flat entities, Seq2Seq-based models are capable of handling flat, nested and discontinuous entities. Moreover, Seq2Seq-based models can handle variable-length inputs and able to be train in an end-to-end manner. However, Seq2Seq models remain computationally expensive to train, have limited interpretability, and suffer from challenges related to decoding efficiency, error propagation and exposure bias. As shown in Eq. 7, the Seq2Seq model first encodes a variable-length source sequence X = x_1,..., x_n into a fixed-length vector using an encoder Enc(.) and decodes it back into a variable-length target sequence Y = y_1,..., y_m using the decoder Dec(.).

Several researchers experimented with Seq2Seq models for NER tasks. Yan et al. (2021) utilised a Bidirectional and Auto-Regressive Transformer (BART) (Lewis et al. 2020) model for the generation of entity pointer index sequences from input sentences using a pointer mechanism. Other researchers employed the Text-To-Text Transfer Transformer (T5) (Raffel et al. 2020) model for Sequence-to-Sequence NER. Zhang et al. (2023d) proposed 2INER which incorporates in-context examples and a new auxiliary task of type extraction to identify all the entity types presented in the original sentence.

$$y_1, ..., y_m = \mathrm{Dec}(\mathrm{Enc}(x_1, ..., x_n)) \tag{7}$$

4.3 Span-based

Span-based NER typically involves a two-step process as depicted in Fig. 5, studied by Sohrab and Miwa (2018). This approach first input sentence x into the encoder Enc(.) to obtain the semantic token representations followed by classifying both the start position and end positions (y_s and y_e) of each span as shown in Eq. 8. Figure 5 demonstrates that concatenating the start and end tokens' representations with inside representation result in span representation. Span representations are obtained by enumerating over all possible spans and then classifying using softmax classifier. Unlike Sequence labelling methods, Span-based approaches are more capable of handling nested entities, Out-of-Vocabulary (OOV) words and entities with medium length. However, longer computation time is required to enumerate all possible spans, which is quadratic to the length of the sentence. Furthermore, span-based NER faces several challenges, including the absence of explicit boundary supervision, overlapping spans, not as effective to capture sequence context and less efficient for long entities.

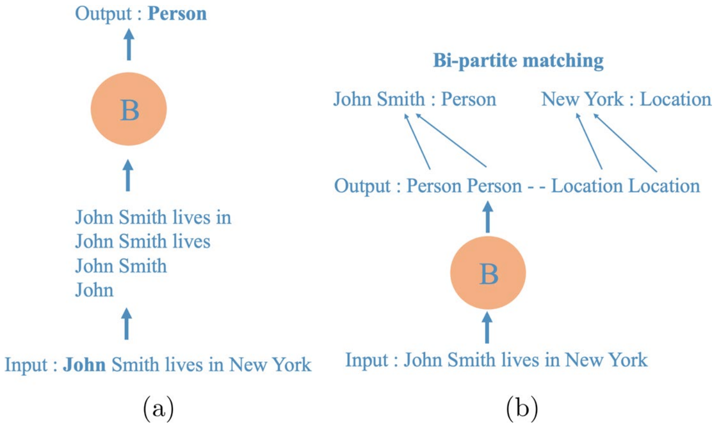

Fig. 5 **a** State diagram of a span-based NER model which concatenates all possible sub-spans for each prediction in the Bi-LSTM (B) (Sohrab and Miwa 2018). **b** This model uses bi-partite matching between the Gold entity set and the predicted entity set to train the decoder (Tan et al. 2021)

Yu et al. (2020a) proposed a span-based model by reformulating NER as dependency-parsing. It used BERT embeddings and character embeddings from a CNN as input to a Bi-LSTM, and a biaffine classifier assigns scores to all possible entity spans. To improve the span representation and tackle the absence of boundary information in span-based NER, researchers applied multi-task learning to incorporate boundary supervision. Tan et al. (2020a) presented an enhanced neural span classification model that integrates boundary detection as an auxiliary task, employing multi-task learning for joint training. Zhang et al. (2023f) proposed SMARTSPANNER using multi-task learning for Named Entity Head (NEH) prediction and span classification.

Span-based methods suffer from the problem of overlapping spans between positive and negative instances. To mitigate this problem, Yu et al. (2020a), Li et al. (2020f) explored greedy decoding algorithms to acquire a set of non-overlapping entities. However, greedy decoding tends to suffer from myopic bias, choosing spans without regard to future decisions. Hence, other researchers (Sarawagi and Cohen 2004; Kong et al. 2016; Ye and Ling 2018) formulated Span NER as joint segmentation and labelling using Semi-Markov CRF. It leverages a globally-normalised model to compute the probability of each labelled segmentation which ensures no overlap between output entities. However, Semi-CRF suffers from quadratic complexity over sequence length and inferior performance compared to CRF. To address this problem in Semi-CRF, Zaratiana et al. (2023) proposed Filtered Semi-CRF that utilises a filtering step to remove irrelevant segments using a lightweight local segment classifier. Besides Semi-CRF based methods, Zaratiana et al. (2022) introduced GN-Ner, utilising Graph Neural Network (GNN) to refine span representations, reducing the occurrence of overlapping spans during prediction.

Recently, researchers experimented with other innovative methods for Span-based NER. Zhu and Li (2022) introduced boundary smoothing as a regularisation method for neural models. Nguyen et al. (2023b) proposed using information bottleneck (IB) models comprising of two Variational Autoencoder (VAE) components for span reconstruction and synonym generation, and one Variational Information Bottleneck (VIB) component to compress span representations. Shen et al. (2023a) introduced DiffusionNER, which conceptualises NER as a process of refining boundaries to extract named entities from spans that are initially noisy. Zhu et al. (2023) discussed DSpERT, comprising a conventional Transformer and a span Transformer to generate span representations that encapsulate deep semantic information. The span Transformer utilises span representations from lower layers as queries while gradually collecting token representations as keys and values from the lower to the upper layers. Existing models often overlook semantic dependencies between spans. To address this, Geng et al. (2023) proposed a planarised sentence representation for nested named entities and implemented a bi-directional, two-dimensional, recurrent operation to capture these dependencies effectively.

$$y_s, y_e = \mathrm{CLS}(\mathrm{Enc}(x)) \tag{8}$$

4.4 Sequence-to-set

Sequence to Set reformulates NER as an entity set prediction task that can better handle complex NER scenarios such as nested NER. As shown in Eq. 9, Sequence-to-Set typically first encodes the input sentence x using an encoder Enc(.). Then, the entity set decoder

Dec(.) utilises the context vector from the encoder together with entity queries x_q as input to predict the start position y_s, the end position y_e of each entity span and the entity class y_c respectively. Tan et al. (2021) first introduced NER as a task based upon the prediction of entity sets. As shown in Fig. 5 (b), their model utilised a sequence encoder, an entity set decoder and a loss function based on bipartite matching. The sequence encoder captures contextual details from the input sentence, while the entity set decoder, aided by entity queries, predicts the boundaries and categories of the entity set. Bipartite matching ensures a unique prediction for each target entity. However, the approach of Tan et al. (2021) assumed that each entity is a span and cannot handle the recognition of discontinuous mentions. To address this problem, He and Tang (2022) introduced a new framework for entity set generation in general NER contexts, treating each entity as a sequence rather than a span, allowing it to identify discontinuous mentions. To incorporate relationships between entity spans in Span-based methods, Wu et al. (2022) presented Propose-and-Refine Network (PnRNet), a two-stage network designed for set prediction in nested NER. During the propose stage, a span-based predictor generates some coarse entity predictions as entity proposals. In the refine stage, proposals interact with each other to incorporate richer contextual information into the proposal representations.

$$y_s, y_e, y_c = \mathrm{Dec}(\mathrm{Enc}(x), x_q) \tag{9}$$

4.5 Machine reading comprehension

Machine Reading Comprehension (MRC) reformulates NER into a question-answering task with the goal of locating entities by predicting their start and end positions within a given context as shown in Fig. 6. Li et al. (2020d) developed an MRC framework for NER based on BERT. Their approach is explained in Eq. 10 where MRC model first encodes the input sentence x together with specific queries x_q into contextualised features using an encoder Enc(.) followed by two linear classifiers CLS(.) for predicting the initial starting position y_s along with the final ending position y_e of each entity span respectively. Then, an additional classifier is used to verify if the start and end positions correspond to the same entity. This approach is able to handle both flat and nested entities and benefits from the incorporation of prior information from entity types through MRC queries. Subsequently, MRC-based NER

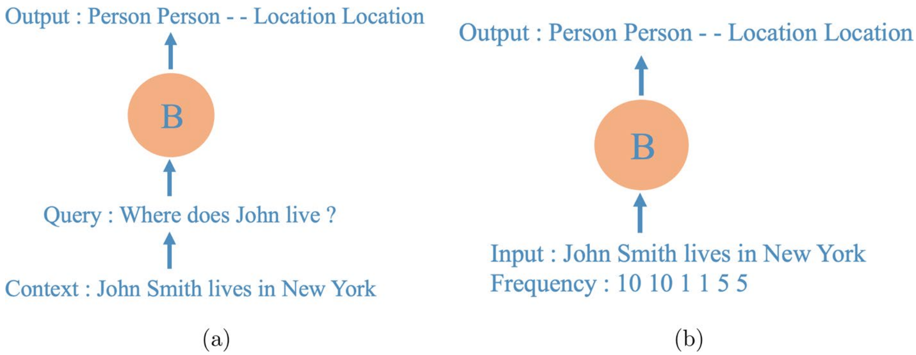
Output : Person Person - - Location Location
B
Query : Where does John live ?
Context : John Smith lives in New York
(a)
Output : Person Person - - Location Location
B
Input : John Smith lives in New York
Frequency : 10 10 1 1 5 5
(b)

Fig. 6 **a** The MRC based model transforms each sentence to a query and then predicts the label using Bi-LSTM (Sun et al. 2021a). **b** This is a template-free model that does not require a span and instead uses words with highest frequency for labels (Ma et al. 2022d) was applied to other domains such as Biomedical (Sun et al. 2021a) and Financial (Zhang and Zhang 2023) domains. Existing Masked Language Models (MLM) suffer from differences among model pre-training tasks and downstream fine-tuning tasks such as NER. To address this, Xu et al. (2023a) introduced the Pre-trained Machine Reader (PMR), which adapts pre-trained MLM for use in pre-trained MRC models without the need for labelled data. This is done by incorporating an MRC head onto existing MLM and conducting continual pre-training with large-scale MRC-styled data. The MRC models can easily adapt to NER tasks, particularly in scenarios that are low-resource. Despite the progress in MRC models for NER, it ignores potential relationships between different entity types during entity extraction (Liu et al. 2023) and extracts one entity type at a time inefficiently. To address this problem, Shen et al. (2022) introduced the Parallel Instance Query Network (PIQN), which employs comprehensive and adaptable instance queries that enable simultaneous querying of all entities with each instance predicting a specific entity type.

$$y_s, y_e = \mathrm{CLS}(\mathrm{Enc}(x, x_q)) \tag{10}$$

4.6 Prompt-based training

Unlike conventional NER supervised learning, prompt-based learning relies on language models that directly predict text probabilities $P(x)$ instead of predicting output label y based on input x as $P(y \mid x)$. In this approach, as shown in Eq. 11, the original input x is altered using a prompting function f_{prompt} to create a text prompt x' with unfilled slots. The language model then probabilistically fills these gaps to generate a final string $\hat{x}$, from which the output y is derived through a verbaliser. This framework's strength lies in the model's pre-training on extensive raw text and its adaptability through new prompting functions. Hence, it can help connect pre-training with diverse downstream tasks, facilitating few-shot or zero-shot learning scenarios (Mao et al. 2023, 2024b). Broad categories for prompt-based training include Template-based, Template-free, continuous/soft prompts, and Question Answering methods.

Cui et al. (2021) introduced a template-driven strategy using BART (Lewis et al. 2020) as shown in Fig. 4, treating NER as a challenge of ranking by a language model. This technique involves crafting templates manually for each class, filled with potential entity spans from the input sentence. Labels are then assigned to these entity span candidates based on the template score. However, this template-based approach is time-consuming due to the need to iterate through all potential entity spans. To resolve this problem of span enumeration, Shen et al. (2023b) introduced PromptNER, which merges entity detection and classification into a single round of prompt learning. This approach eliminates the need to iterate over entity spans or types through employing position slots [P] and type slots [T] within the prompt template.

$$y = verbalizer(\mathrm{CLS}(f_{prompt}(x))) \tag{11}$$

Addressing the span enumeration problem in template-based prompt NER models, template-free approaches were proposed by researchers for prompt-based NER. Ma et al. (2022d) introduced a template-free prompt method without using a pre-defined prompt template. As shown in Fig. 6 (b), their approach involves creating label words and predicting label words for actual entity token positions while the non-entity token positions predict themselves. Specifically, given the original input X, the LM is trained to find the maximum probability P($X^{Ent}|X$) of the target sentence X^{Ent} as shown in Eq. 12. Using contrastive learning, He et al. (2023) proposed another Template-free prompt-based method. First, external knowledge from label descriptions is utilised to initialise semantic anchors for each entity type. These anchors are simply appended with input sentence embeddings as template-free prompts (TFPs). Then, prompts and sentence embeddings are in-context optimised with their proposed semantic-enhanced contrastive loss.

$$\mathcal{L}_{EntLM} = -\sum_{i=1} \log(P(x_i = x_i^{Ent}|X)) \tag{12}$$

The above-mentioned methods utilised discrete (hard) prompts with template words linked to natural language phrases. Hard prompts utilise knowledge contained in PLM. However, hard prompts are not flexible enough and require prior expertise to construct. To resolve this limitation, continuous (soft) prompts are introduced, which operate directly within the model's embedding space. Soft prompts alleviate two limitations: (1) they loosen the criteria for template words' embeddings to be natural language embeddings, and (2) they remove the constraint that the template is parameterised by a pre-trained LM's parameters. Despite these advantages, soft prompts are hard to interpret and typically cannot be easily transferred. Chen et al. (2022b) proposed LightNER based on BART (Lewis et al. 2020) which integrates continuous prompts within the self-attention layer, directing attention for prompt tuning. Liu et al. (2024) integrated a deep prompt tuning framework with threefold knowledge (TKDP), encompassing internal context knowledge, external label knowledge and sememe knowledge. TKDP encodes these three sources of information into soft prompt embeddings, which are then embedded into a language model pre-trained for enhancing prediction accuracy.

Lastly, another type of prompt-based NER, Question-answering NER, combines the tasks of NER with the ability to answer questions related to identified entities. Previous prompt-based NER methods suffer from several limitations of high computational complexity, manual prompt engineering, the lack of prompt robustness and low transferability. Arora and Park (2023) decomposed NER into two distinct sub-tasks based on question-answering: Span Detection, which focuses solely on identifying entity mention spans without considering their types, and Span classification, which categorises these spans into specific entity types.

4.7 Large language models

LLMs are prevalent today and are used for many tasks (Mao et al. 2024a). The most successful LLM is built around the architecture of transformers. Transformers, the backbone of modern NLP models, are built upon self-attention mechanism using multi-head self-attention. This technique enables models to assess the significance of various parts of the input sequence when capturing word dependencies or making predictions. The equation of self-attention can be shown in Eq. 13, where the query vector is represented by Q, the key vector is represented by K, d_k is the dimension of the key vector and V is the value vector. Through the use of self-supervised pre-training tasks, including masked language modelling and next token prediction, the LLMs are pre-trained. Depending on their model architectures, they can be categorised into different types. Earlier LLMs were based on encoders such as BERT (Devlin et al. 2019) and decoders such as GPT (Radford et al. 2018). Subsequently, LLMs were created based on encoder-decoder such as BART (Lewis et al. 2020) and T5 (Raffel et al. 2020). These LLMs are able to adapt to downstream NER tasks through fine-tuning or prompting. A unified Seq2Seq approach (Yan et al. 2021) based on BART (Lewis et al. 2020), integrates a pointer mechanism to generate entity pointer index sequences from input sentences. Zhang et al. (2024) introduced LinkNER, which integrates smaller fine-tuned NER models with LLMs through an uncertainty-based linking strategy, Recognition-Detection-Classification (RDC). LLMs are scalable, versatile, and adapt to downstream NER tasks by fine-tuning or prompting. However, with a large number of parameters, LLMs have several disadvantages, including the requirement for a large training dataset for pretraining, high computational costs and problems of bias and hallucination.

$$H = \mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{13}$$

Recently, in Song et al. (2024) authors manipulated words in a sentence to generate additional synthetic data in languages or domains with scarce labelled data. They considered the CoNLL2003 and MIT Restaurant and Movie corpus from 2013. They showed that the accuracy of the model increases with the number of times prompts are changed for entity and context in a sentence on all three datasets. They proposed a self-consistency method in which prompts with inconsistent predictions are discarded. However, they considered pseudo annotation of unlabelled data that can reduce the reliability of the results.

In Merdjanovska et al. (2024), the authors developed a new benchmark dataset for evaluated NER called Noisebench. This dataset covers different types of noise in labels such as expert errors and automatic annotation errors. The experiments showed a higher accuracy with noisy samples, confirming that models can identify general patterns in the presence of noise. They also showed that in real noise, the model has similar prediction on both seen and unseen data; however, when noise is simulated, the seen data significantly outperform unseen data.

4.8 Graph neural networks

GNN is a unique type of neural network that uses the graph data structure. A graph contains nodes representing each word and edges connecting two nodes. A general approach for GNN involves message passing between nodes and the aggregation of information from neighbouring nodes $N(v)$. The message passing is shown in Eq. 14 where h_u^k is the embedding corresponding to each node, h_u^{k+1} is the updated node, σ represents the element-wise non-linearity, and W_k and B_k represent trainable parameter matrices. Chen et al. (2021a) proposed Entity Relation Graphs (EnRel-G) which explicitly connect entity mentions based on both global co-reference and local dependency relations, intended for the creation of improved entity mention representations. Wang et al. (2023) introduced Graph Neural Network Sequence labelling (GNN-SL), which enhances the standard sequence labelling (SL) model by incorporating similar tagging examples from the entire training set, improving its ability to address long-tail instances in the SL task. Sui et al. (2022) proposed a trigger-based graph neural network (Trigger-GNN) for nested NER. It obtains the complementary annotation embeddings through entity trigger encoding and semantic matching, and tackles nested NER using a GNN. The advantages of GNNs include the ability to learn from neighbouring nodes and built-in inductive capabilities. However, GNN's architecture is typically limited to shallow networks because of the over-smoothing problem from the overlapping of the receptive field between two nodes with increasing GNN layers. In addition, GNNs architecture is also constantly changing and suffers from scalability issues.

$$h_v^{k+1} = \sigma\left(W_k \sum_{u \in N(v)} \frac{h_u^k}{|N(v)|} + B_k h_v^k\right), \quad \forall k \in \{0, \ldots, K-1\} \tag{14}$$

4.9 Text entailment

NER can be reformulated as a text entailment task (Bowman et al. 2015) that predicts the relation of two sentences, Premise (P) and Hypothesis (H): whether H is true given P. For example, for the given input sentence "Paracetamol is effective in alleviating fever and pain", "Paracetamol" is a MEDICINE entity. The input sentence acts as a premise, while the assertion "Paracetamol is a MEDICINE" acts as a hypothesis. Subsequently, an NER task is formulated as a text entailment problem to determine the truth of the hypothesis considering the premise. As shown in Eq. 15, the permise x_p and hypothesis x_h are encoded by the encoder Enc(.) followed by a classification layer CLS(.), which performs classification to obtain entailment scores y_c. This method is suitable for low-resource NER scenarios, as it requires specifying labels for only certain entities during training, rather than needing complete annotations for the entire sequence as is the case with sequence labelling, which is often prone to noisy annotations. However, one disadvantage is the enumeration over all possible text spans or words in the input sentence to obtain named entity candidates similar to span-based methods and template-based prompt methods. Li et al. (2022a) was the first to propose Prompt-based Text Entailment (PTE) for low-resource NER by treating the original sentence as premise and the entity type-specific prompt as hypothesis. Given an entity type, the Premise and Hypothesis are fed into PLMs to get entailment scores for each candidate. The entity type with the top entailment score is chosen as the final label. Liang et al. (2023) improved on the approach of Li et al. (2022a) for BioNER tasks by proposing Textual Entailment with Dynamic Contrastive learning (TEDC), which reduces noisy labelling from gazeteers and improves the discrimination ability between entities and non-entities via contrastive learning.

$$y_c = \mathrm{CLS}(\mathrm{Enc}(x_p, x_h)) \tag{15}$$

4.10 Summary

Table 3 summarises the advantages and disadvantages of different NER modelling paradigms. NER modelling paradigms are categorised into 9 categories; the majority of the papers fall under sequence labelling, span-based, MRC and prompt-based. We also provide

Table 3 Summary for advantages and disadvantages of different NER modelling paradigms

Modelling paradigms	Advantages	Disadvantages
Seq labelling	Suitable for flat NER task and better at dealing with those entities that are long and with low label consistency	Not fully utilising label information, Not suitable for nested and discontinuous NER, design of specialised tagging schemes required for more complex NER tasks
Seq2Seq	Able to handle flat, nested and discontinuous entities, capable of capturing sentence context, handling variable length inputs and benefits from end-to-end training	Computationally expensive to train, limited interpretability, challenges in decoding efficiency and exposure bias
Span-based	Suitable for nested entities, OOV words and medium length entities	Computationally expensive, lack explicit boundary supervision, overlapping spans, not as effective to capture sequence context, less efficient for long entities
Text entailment	Suitable for low resource NER, benefits from prior label information	Computationally expensive due to span enumeration
Seq2Set	Ability to handle nested entities, benefits from prior knowledge of all entity queries	Problem with discontinuous entities
MRC	Suitable for both flat and nested entities, prior information about entity types can be included through queries	Extract one entity type at a time, disregard relationship between entity types, manually constructing MRC queries and span enumeration
Prompt	Ability to adapt pre-trained model using prompting functions, enabling few-shot and zero-shot learning and able to handle nested entities	Reliant on prompt engineering, lack of consistency and reliability, limited control over output format, suffer from scalability and latency concerns, soft prompts are hard to interpret
LLMs	Scalable and versatile, able to fine tune and adapt to downstream tasks	Large training dataset required, high computational cost, bias and hallucination
GNN	Able to learn from neighbouring nodes, built-in inductive capabilities	Shallow networks, constantly changing and scalability issues

Seq sequence, *Seq2Seq* sequence to sequence, *Seq2Set* sequence to set, *MRC* machine reading comprehension, *LLMs* large language models, *GNN* graph neural network a brief explanation of each NER modelling paradigm and its corresponding equations in Table 4.

In Sect. 3 we introduce different learning methods to improve the generalisation of the model to unseen data. However, in Sect. 4 we introduce popular models for NER that use a combination of two or more learning methods. BERT in Fig. 4a for example is a pre-trained model for NER that only uses labels for fine-tuning. Word vectors are generated in a completely unsupervised manner using co-occurrence information. However, the fine-tuning of BERT on a specific task is a supervised learning. For example, in Fig. 6a we show MRC where each sentence is paired with a corresponding query prompt prior to training. This additional information based on common sense can be considered as an unsupervised part of the model. The template-free model in Fig. 6b uses the frequency of words instead of the span of words to determine the boundaries of a named entity. This is another example of unsupervised learning used in conjunction with supervised learning. Another interesting approach is bipartite matching in Fig. 5b useful for nested entities where the cardinality of named entitles is used to match graphs. This is another example of using an unsupervised method inside a supervised framework. Lastly, in this article, we perform automatic labelling of election tweets using ChatGPT. Since this can introduce errors, we can consider this a type of weakly supervised model.

5 NER datasets

NER has wide application in many real-world scenarios. In finance, for example, it can be used to extract critical terms from legal contracts such as parties involved, payment dates, and penalties. Another application is to forecast election results from news articles. It can extract names of politicians, political parties and social issues allowing for real time tracking of majority votes. During product recommendation, NER can help determine the target aspect for a positive or negative sentiment. Similarly, a major hotel chain can use NER to extract aspects such as “room cleanliness", “staff behaviour" and identify recurring problems and improve customer experience. Table 6 lists the common NER benchmark datasets for tasks such as biomedical articles, social media, multi-lingual text, and scanned documents.

Table 5 illustrates examples of NER prediction by BERT and GLINER for the QLD Election dataset described in Sect. 8. We can see that GLINER is able to predict all the entities correctly. However, BERT has two incorrect predictions both of which are of type “Person". This could be because there are a large number of politicians. This can also be observed in Fig. 7. Here we show the number of entity samples for three entity types “Person", “Organisation" and ‘Location’ in six different datasets. The “Person" entity type has the highest frequency in most datasets. This is followed by “Location". The type “Organisation" has the fewest samples in most datasets or, like the “Person" class.

5.1 Vanilla NER

CoNLL-2002 dataset (Tjong Kim Sang 2002) is available in Spanish and Dutch for language-independent NER. The dataset includes entities for person, location, organisation and miscellaneous entities. CoNLL-2003 dataset (Tjong Kim Sang and De Meulder 2003)

Table 4 Summary for the explanation of each modelling paradigm and its corresponding equations

Modelling paradigms	Explanation	Equations
Seq labelling	Sequence labelling assigns specific labels to each word in a given text sequence	$y_1, ..., y_n = \mathrm{Dec}(\mathrm{Enc}(x_1, ..., x_n))$
Seq2Seq	Seq2Seq encodes a source sequence and decoding it into target sequence	$y_1, ..., y_m = \mathrm{Dec}(\mathrm{Enc}(x_1, ..., x_n))$
Span	Span-based model obtain semantic representations at the span level, followed by enumerating over all possible spans for classification	$y_s, y_e = \mathrm{CLS}(\mathrm{Enc}(x))$
Seq2set	Sequence to Set encode the input sentence and entity queries to predict start, end position and entity class of each entity span	$y_s, y_e, y_c = \mathrm{Dec}(\mathrm{Enc}(x), x_q)$
MRC	MRC reformulates NER into a question-answering format for predicting their start and end positions of entity spans	$y_s, y_e = \mathrm{CLS}(\mathrm{Enc}(x, x_q))$
Prompt	Prompt-based learning relies on language models that directly predict text probabilities using prompting functions	
		$y = verbaliser(\mathrm{CLS}(f_{prompt}(x)))$
		$\mathcal{L}_{EntLM} = -\sum_{i=1} \log(P(x_i = x_i^{Ent} \mid X))$
LLMs	LLMs utilise self-attention and adapt to downstream task through fine-tuning or prompting	$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$
GNN	GNNs learn word embeddings through message passing between nodes and aggregating information from neighbouring nodes	$h_v^{k+1} = \sigma\left(W_k \sum_{u \in N(v)} \frac{h_u^k}{\lvert N(v)\rvert} + B_k h_v^k\right)$
Text entailment	Reformulates NER as text entailment task to predict the relation of two sentences (Premise and Hypothesis)	$y_c = \mathrm{CLS}(\mathrm{Enc}(x_p, x_h))$

Seq sequence, *Seq2Seq* sequence to sequence, *Seq2Set* sequence to set, *MRC* machine reading comprehension, *LLMs* large language models, *GNN* graph neural network is available in English and German. The English data originated from the Reuters Corpus, and the German data from the ECI Multilingual Text Corpus. The dataset features entities such as person, location, organisation, and miscellaneous entities.

OntoNotes−5.0 dataset (Weischedel et al. 2013) features a variety of text genres, such as news articles, telephone conversational dialogues, weblogs, Usenet newsgroups, broadcasts and talk shows. It is available in English, Chinese, and Arabic languages. It encompasses 18 different types of entity, including cardinals, dates, events, laws, and languages.

NCBI Disease corpus (Doğan et al. 2014) consists of 793 PubMed abstracts divided into 593 abstracts for training set, 100 abstracts for validation set, and 100 abstracts for test set. Each abstract is annotated with disease mentions.

BC5CDR dataset (Li et al. 2016) includes 1,500 PubMed articles with annotations for 4,409 chemicals, 5,818 diseases, and 3,116 chemical-disease relations. The dataset is divided into 5228 examples in the training set, 5330 examples in the validation set, and 5865 examples in test set.

Broad Twitter Corpus (BTC) (Derczynski et al. 2016) is based on various regions, time periods, and types of Twitter users. The dataset includes entities for persons, organisations, and locations. The dataset comprises 9551 documents, 165739 tokens, 5271 person entities, 3114 location entities and 3732 organisation entities.

5.2 Nested NER

GENIA dataset (Kim et al. 2003), consists of 2000 annotated MEDLINE abstracts from the biomedical literature. These abstracts were collected through a PubMed search using the MeSH terms “human", “blood cells”, and “transcription factors”. The dataset provides annotations for five types of biological entities based on DNA, RNA, Protein, Cell line, and Cell category and includes 36 detailed entity categories across the 2000 abstracts.

5.3 Discontinuous NER

CADEC dataset (Karimi et al. 2015) is obtained from AskaPatient, a platform for patients to discuss their medication experiences. It comprises 1253 posts with 7597 sentences. CADEC includes annotations for entities such as drugs, Adverse Drug Events (ADEs), diseases, symptoms and findings. It contains 679 discontinuous mentions out of 6318 mentions.

ShARe13 (Pradhan et al. 2013) and ShARe14 (Mowery et al. 2014) focus on annotating mentions of disorders in various clinical documents, such as discharge summaries, electrocardiograms, echocardiograms, and radiology reports. ShARe13 dataset contains 1,088 discontinuous mentions out of 11,148 mentions. ShARe14 dataset contains 1,650 discontinuous mentions out of 19,047 mentions.

5.4 Joint NER and relation extraction

SciERC dataset (Luan et al. 2018a) comprises 500 scientific abstracts from AI papers annotated with 8089 scientific entities, 4716 relationships, and 1023 coreference clusters. It features 6 entity types (Task, Method, Metric, Material, Other-Scientific-Term, Generic) and 7 relationship types (Compare, Conjunction, Evaluate-For, Used-For, Feature-Of, Part-Of,

Table 5 Comparison of NER prediction by GLINER and BERT on QLD election dataset

Sentence	Ground truth		Prediction	
	Entity	Type	GLINER	BERT
Like all but 3 members of his party, QLD opposition leader Crisafulli, voted against legalising abortion in 2018. #qldvotes	QLD opposition leader Crisafulli	Person	Correct	Correct
	Abortion	Social Issue	Correct	Incorrect
Here is a scare campaign. The Liberals will attack women's rights and thereby attack human rights. #qldvotes #auspol	Liberals	Political Party	Correct	Correct
	Women's rights	Social Issue	Correct	Correct
	Human rights	Social Issue	Correct	Incorrect
Message for Qlders, it's not too long for Labor in Govt., Premier Steven Miles getting so much done. Vote Labor, don't go back to the LNP Campbell Newman plan. #qldvotes	Labor	Political Party	Correct	Correct
	Premier Steven Miles	Person	Correct	Correct
	LNP	Political Party	Correct	Correct
	Campbell Newman	Person	Correct	Incorrect

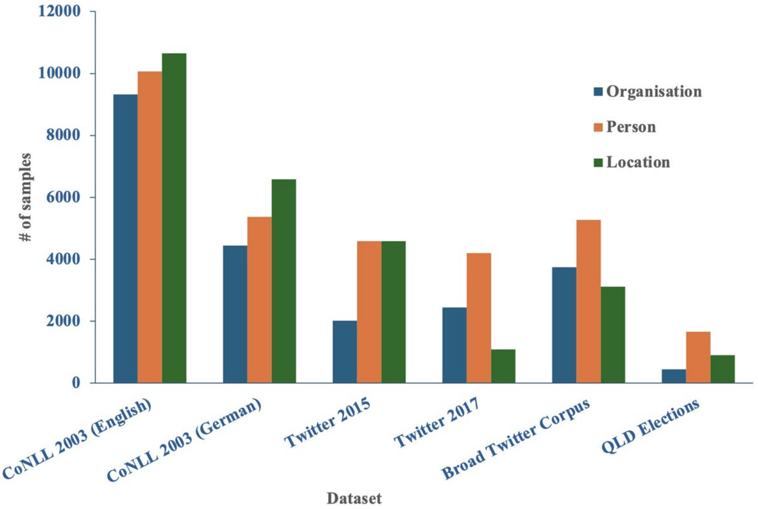

Fig. 7 Frequency of entity types in different datasets. Person class has the highest frequency in most datasets

Hyponym-Of) across 2,687 sentences. The dataset is divided into 1861 sentences for the training set, 275 sentences for the validation set, and 551 sentences for test set.

ACE-2004 Multilingual Training Corpus (Doddington et al. 2004) includes texts from various genres in English (158,000 words), Chinese (307,000 characters or 154,000 words), and Arabic (151,000 words), all annotated for entities and relations. Walker et al. (2006) created the ACE 2005 Multilingual Training Corpus, a mixed-genre dataset with 1800 files in English, Arabic, and Chinese languages, annotated for entities, relations, and events. Both the ACE-2004 and ACE-2005 datasets feature seven entity types based on Person, Organisation, Facility, Location, Geo-Political Entity, Weapon, and Vehicle.

CoNNL-2004 corpus (Roth and Yih 2004) consists of 1437 sentences, each containing at least one relation. Among the sentences, there are 5336 entities, and 19048 pairs of entities (binary relations). The dataset is annotated with four named entity types including 1685 persons, 1968 locations, 978 organisations and 705 others and five relation labels include 406 located_in, 394 work_for, 451 orgBased_in, 521 live_in, 268 kill, and 17007 none.

5.5 Cross-domain NER

CrossNER dataset (Liu et al. 2021) is designed for cross-domain NER. It spans five distinct domains based on politics, natural science, music, literature, and artificial intelligence with specific entity categories for each domain. Additionally, CrossNER provides unlabelled corpora related to each domain.

5.6 Cross-lingual NER

Wikiann NER corpus (Rahimi et al. 2019) comprises 41 languages selected for their alignment with multilingual word embeddings. The corpus uses IOB2 format tags for location, person, and organisation. To handle label imbalance, the dataset was balanced and divided into training, development, and test sets.

5.7 Few-shot NER

Few-NERD dataset (Ding et al. 2021) for Few-Shot NER includes 188,238 sentences annotated with 491,711 entities based on 8 coarse-grained entity types and 66 fine-grained entity types. There are three dataset variants of FEW-NERD (SUP), FEW-NERD (INTRA) and FEW-NERD (INTER) based on the benchmark tasks. FEW-NERD (SUP) dataset is used for supervised settings with training, validation, and test sets containing 131,767, 18,824, and 37,648 samples, respectively. FEW-NERD (INTRA) dataset is randomly divided by coarse type with training, validation, and test sets containing 99,519, 19,358, and 44,059 samples, respectively. FEW-NERD (INTER) dataset is randomly divided within coarse type, meaning each file includes all 8 coarse types but features different fine-grained types with training, validation, and test sets containing 130,112, 18,817, and 14,007 samples, respectively.

5.8 Multi-modal NER

Two multi-modal Twitter NER datasets, namely Twitter2015 (Zhang et al. 2018) and Twitter2017 (Lu et al. 2018) cover entity types such as person, location, organisation, and other/

miscellaneous. The Twitter2015 dataset is divided into 4000 tweets for the training set, 1000 tweets for the development set, and 3257 tweets for test set. The Twitter2017 dataset is divided into 3373 tweets for the training set, 723 tweets for the development set, and 723 tweets for the test set.

SnapCaptions dataset (Moon et al. 2018) consists of image-caption pairs generated by 10,000 users. Expert annotators manually identified the named entities in the captions and categorised them into person, location, organisation, and miscellaneous entity types. The dataset is divided into 70% for the training set, 15% for the validation set and 15% for test set. The caption dataset has an average length of 30.7 characters (5.81 words) and a vocabulary size of 15,733, out of which 6612 tokens are not found in the Stanford GloVe embeddings (Pennington et al. 2014) and are treated as unknown.

Twitter-GMNER dataset (Yu et al. 2023) is designed for Grounded Multi-modal NER and builds on the Twitter2015 and Twitter2017 datasets. It includes annotations for entities and their types in each multi-modal tweet such as person, location, organisation, and other/miscellaneous. The dataset is divided into 70% for the training set, 15% for the validation set, and 15% for test set. It contains 16,778 entities, with approximately 60% lacking grounded bounding boxes. For the rest of 6716 groundable entities, 8,090 bounding boxes were manually annotated, with some entities corresponding to multiple bounding boxes.

5.9 Visually-rich document NER

FUNSD dataset (Jaume et al. 2019) contains 199 real, fully-annotated forms, 31485 words, 9707 semantic entities, 5304 relations. The noisy scanned forms divided into 149 for training set and 50 for test set, suitable for a range of tasks such as text detection, optical character recognition, spatial layout analysis, and entity labelling/linking. It features four semantic entity categories of question, answer, header, and other.

SROIE dataset (Huang et al. 2019b) consists of 1000 whole scanned receipt images and annotations for semantic entity recognition. The dataset contains company, date, address, and total labels. The dataset is divided into a training/validation set of 600 images and a test set of 400 images.

5.10 NER evaluation metrics

NER evaluation assesses the performance of NER models by relying on three primary metrics of Precision, Recall and F1 score. Evaluation metrics can be further categorised into two types: exact match and relaxed (partial) match. The exact match score measures entities where the predicted entity precisely matches the ground truth entity. In contrast, the relaxed (partial) match score considers partial matches between the predicted entity and the groundtruth entity, allowing for some level of variation. These evaluation approaches enable a comprehensive assessment of the effectiveness of the NER system in accurately capturing named entities. In the context of NER, precision refers to the percentage of the NER system results that are correctly predicted, while recall refers to the percentage of total entities correctly predicted by the NER system. F1 score is then defined as the harmonic mean between the precision and recall scores. Mathematically, precision, recall and the F1 score are calculated based on the number of false positives, number of false negatives and number of true positives which are further defined below.

- False Positive (FP): a named entity predicted by the NER system but not present in the ground truth.
- False Negative (FN): a named entity present in the ground truth but missed by the NER system.
- True Positive (TP): a named entity correctly predicted by the NER system and present in the ground truth.

$$Precision = \frac{\#TP}{\#TP + \#FP} \tag{16}$$

$$Recall = \frac{\#TP}{\#TP + \#FN} \tag{17}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

5.11 Summary

Table 6 covers the different datasets used for different NER tasks based on corpus, category, #Tags and tasks. The NER datasets are categorised into Vanilla NER, Nested NER, Discontinuous NER, Joint NER and RE, Multi-modal NER, VrD-NER, Few-shot NER task, Cross-domain NER, and Cross-lingual NER task. The majority of the NER datasets belong to the Vanilla NER task, followed by Cross-lingual NER. Vanilla NER datasets are derived from various sources such as Wall Street, news domains, biomedical and tweets. Nested NER datasets come from biomedical, news domains and multi-lingual datasets. Discontinuous NER datasets are from the medical domain. Cross-lingual datasets are based on news domains and Wikipedia. Cross-domain datasets are derived from diverse domains such as Politics, Natural Science, Music, Literature, and Artificial Intelligence. Joint NER and RE datasets are from biomedical domains, mixed genres and multi-lingual datasets. Few shot NER datasets come from mixed genres and comprise coarse- and fine-grain entities. VrD-NER datasets consist of documents from forms, invoices and receipts. Multi-modal NER datasets are from social media domains such as Twitter and Snapchat.

6 Common NER tasks

6.1 Vanilla NER

Vanilla NER task refers to the extraction of flat named entities from a given input text under single language, single domain and high resource data scenarios. There are various techniques used by researchers for NER which includes data augmentation, document-level context, external knowledge, multi-task learning, adversarial learning, transfer learning, self-attention, ensemble learning and knowledge distillation.

Table 6 NER datasets for common and more complex NER tasks

Corpus	Category	#Tags	NER	Nested	Dis	JNER	CD	CL	Few	Multi-modal	VrD
CONLL 2002	Reuters news	4	✓					✓			
CoNLL-2003	Reuters news	4	✓					✓			
GENIA	Biomedical	36	✓	✓							
OntoNotes—5.0	Mixed genres	18	✓								
NCBI-disease	Disease	1	✓								
BC5CDR	Chemical disease	2	✓			✓					
BTC	Twitter data	3	✓								
Wikiann	Multilingual	3						✓			
SciERC	Scientific abstract	6				✓					
ACE-2004	Multilingual	7		✓		✓					
ACE-2005	Multilingual	7		✓		✓					
CoNLL-2004	General	3				✓					
CADEC	Medical	5			✓						
ShARe 13	Disorder	1			✓						
ShARe 14	Disorder	1			✓						
FUNSD	Noisy Scanned Docs	4									✓
CrossNER	Cross-domain	-					✓				
Few-NERD	Mixed genres	8							✓		
SROIE	Receipts	4									✓
Twitter2015	Social media	4									✓
Twitter2017	Social media	4									✓
SnapCaptions	Social media	4								✓	
Twitter-GMNER	Social media	4								✓	

Dis discontinuous, *JNER* joint NER and RE, *CD* cross-domain, *CL* crosslingual, *Few* few-shot, *VrD* visually-rich document

6.1.1 Data augmentation

Data augmentation is a common technique used to generate artificial training samples to augment an original dataset. Dai and Adel (2020) experimented with some simple NER data augmentation such as label-wise token replacement, synonym replacement, mention replacement, and segment shuffling. Despite the promising results of data augmentation, these manual data manipulation methods suffer from noisy and mislabelled samples and augmented data may be syntactically and/or semantically incorrect. To address this issue, Ke et al. (2023) proposed SAINT which includes syntactic features in a pre-trained language model to generate samples with pre-defined entities. To reduce reliance on labelled data, Chen et al. (2020a) proposed the LADA method to create artificial samples by interpolating sequences that are similar. Intra-LADA performs interpolation within a single sentence, while Inter-LADA involves interpolation between tokens from different sentences.

6.1.2 Document-level context

While NER is traditionally modeled at the sentence-level, transformer-based models offer a natural option to capture document-level features by passing a sentence with its surrounding context which can be useful to classify named entities within the sentence. Luoma and Pyysalo (2020) utilised BERT models to obtain cross-sentence information for NER and propose Contextual Majority Voting (CMV) to combine these different predictions. Other researchers propose multiple-level contexts by incorporating document-level context with sentence-level context or word-level context. Hu et al. (2019) proposed to encode sentences within a document using a Bi-LSTM layer at the sentence level. Subsequently, a document-level module is used to encode the relationships between occurrences of specific tokens. Luo et al. (2020a) proposed a NER model that integrates hierarchical contextualised representations. At the sentence level, the model improves word representations by using a label embedding attention mechanism. At the document level, a key-value memory network is utilised to obtain document-specific information for each word. Chen et al. (2020b) proposed to incorporate multi-level contexts for NER through document-level context and word-level context. Document-level context is obtained from interactions between sentences via multi-head self-attention while word-level context is obtained from an auxiliary task to predict the type of each word. Yang et al. (2022b) proposed AMFF which aims to capture multi-level global and local features based on character-level and word-level comprehensively within the current context. Additionally, document-level features are incorporated through Context-Aware Attentive Multilevel Feature Fusion (CAMFF).

6.1.3 External knowledge

External knowledge can introduce valuable additional information to augment the original input into the NER model. Different types of external knowledge explored by researchers include retrieval augmented methods, gazetteers and various syntactic knowledge sources from POS tags and dependency trees. In social media and E-commerce domains, user search queries, tweets and short comments may lack context, affecting the accuracy of NER. To solve this problem, various studies augment the original text with external context retrieved from search engines to improve NER performance. Wang et al. (2021b) combined external context from search engines with original input as retrieval-based input view. Subsequently, both the original input view and the retrieval-based input view are trained using cooperative learning. Zhang et al. (2022b) used the Elasticsearch engine to find related samples for a given text and used a transformer-based, multi-instance cross-encoder to model correlated samples.

Gazetteers are collected dictionaries or lexicons consisting of long lists of entity names constructed from external knowledge bases such as Wikidata. Liu et al. (2019) utilised a Hybrid Semi-Markov CRF which scores spans derived from token label scores and another module to score candidate entity spans based on how closely it softly matches the gazetteer.

However, the above approach fails to consider contexts when applying entity dictionaries to NER. To address this, Wu et al. (2020a) leveraged dictionary knowledge with contextual information and context-dictionary attention to learn relationships between contexts and entities within the dictionaries.

Syntactical knowledge obtained from POS tags and dependency trees can introduce external knowledge to NER systems. POS tags categorise words based on their function in a sentence, such as nouns, verbs, adjectives, etc. Dependency tree structures capture intricate syntactic relationships and long-distance dependencies among words within a sentence. Aguilar et al. (2018) utilised grammatical details like POS and dependency features for NER. However, previous models do not consider the noise from POS and need to re-extract features from token representations. To alleviate POS noise, Bai et al. (2020) proposed incorporating POS features through an attention mechanism and adversarial training. Jie and Lu (2019) introduced a LSTM-CRF model guided by dependencies which encodes entire dependency trees to capture essential dependency information for NER tasks. Stacking LSTM and GCN architectures for NER, as proposed by Jie and Lu (2019), yields only modest improvements. Xu et al. (2021a) propose Syn-LSTM which incorporates a graph-encoded representation to enhance memory and hidden state updates, allowing for more effective integration of structured information from dependency trees. Nie et al. (2020) proposed incorporating various types of syntactic information such as POS labels, syntactic constituents, and dependency relations through an attentive ensemble using Key-Value Memory Networks (KVMN), as proposed by Miller et al. (2016).

6.1.4 Multi-task learning

Multi-task learning trains the main NER task together with other auxiliary tasks by sharing the network's layers and parameters across different tasks. To alleviate the lack of labelled training data, Liu et al. (2018) jointly trained the sequence labelling task with a neural language modelling task for character-level understanding using multi-task learning. As single-target learning can limit the performance and model efficiency for NER, Hu et al. (2021) proposed multi-task learning of boundary labelling and type labelling subtasks and aggregated the predictions of sub-tasks together. In a recent study, Zhong et al. (2022) explored the joint modelling of NER and Named Entity Classification (NEC) to determine whether semantics play a role in aiding syntax. Results showed that NER remains primarily a syntactic task and the simultaneous modelling of NER and NEC does not improve NER outcomes. NER could benefit from linguistic dependency knowledge; however, existing NER models can currently utilise this information only if the datasets include dependency annotations.

6.1.5 Adversarial learning

Two main types of adversarial learning used for NER include adversarial samples and adversarial adaptation. Adversarial samples are training samples that are modified as adversarial attacks used in model training to build a more robust model. Reich et al. (2022) proposed expert-guided heuristics to create adversarial examples and a mix-up strategy between the original examples and their adversarial samples to enhance generalisation and reduce overfitting. On the other hand, adversarial adaptation aims to create a shared embedding space between source and target datasets using a domain discriminator. To better incorporate Part of Speech knowledge into the model, Bai et al. (2020) proposed Adversarial NER with POS label embedding (ANP) that utilises adversarial training and task-attention mechanism to map shared information between POS and NER tasks into a shared feature space. A discriminator is used to determine the task from which the training sentence comes from, and the shared encoder is trained to produce sentence representations that prevent the discriminator from identifying the task. This is done by optimising the min-max objective function as shown in Eq. 19, where $\hat{r}^{ner}$ and $\hat{r}^{pos}$ are the probabilities that the training sentence comes from the NER task and POS task respectively, with r=1 for NER tasks and r=0 for POS tasks. To better integrate shared knowledge from multi-task sequence labelling models, Wang et al. (2020c) proposed MTAA, a symmetric, multi-task sequence labelling model which extracts shared knowledge among POS, NER and Chunking tasks by adversarial learning and proposes an attention mechanism for merging feature representations.

$$L_{adv} = \underset{\theta_E}{min}(\underset{\theta_J}{max})(r.\log(\hat{r}^{ner})) + (1-r).\log(\hat{r}^{pos}) \tag{19}$$

6.1.6 Transfer learning

Deep learning methods for NER have recently garnered significant attention because they enable end-to-end learning of model parameters without requiring manually engineered features. However, deep learning is highly dependent on high-quality labelled data, which is expensive to obtain. Transfer learning can help address this issue by transferring knowledge gained from a self-supervised pre-training task to support the downstream NER task. Wang et al. (2018b) applied a label-aware double transfer learning framework (La-DTL), which includes the label-aware Maximum Mean Discrepancy (MMD) for transferring feature representations and a label-aware L2 constraint for parameter transfer, with a theoretical upper bound. Gligic et al. (2020) proposed to bootstrap neural networks through transfer learning, utilising pre-trained word embeddings derived from a secondary task on unannotated electronic health records.

6.1.7 Self-attention

The self-attention mechanism enhances NER by creating a comprehensive representation of each sequence. It assigns a query, key, and value vector to each token. The model calculates attention scores by comparing the query vector of one token with the key vectors of all others using a scaled dot product. These scores are then normalised with a softmax function to determine the importance of each token. The final output for each token is a weighted sum of value vectors, where the weights reflect the computed attention scores. This approach allows the model to integrate information from different positions within the sequence more effectively. Devlin et al. (2019) introduced BERT, which uses multi-headed self-attention layers that can be adapted for various downstream NER tasks. Yamada et al. (2020) presented LUKE, which integrates an entity-aware self-attention mechanism that extends BERT's approach by considering whether tokens represent words or entities when computing attention scores. Wu et al. (2023a) proposed Adversarial Self-Attention (ASA), which introduces adversarial biases to attention mechanisms to reduce reliance on specific features like keywords and promote a broader semantic exploration.

6.1.8 Ensemble learning

Ensemble learning aims to combine multiple models to create an ensemble model by integrating the model output to improve the accuracy and robustness of the NER predictions. Several methods can be used for this purpose, including averaging, majority voting, weighted averaging, and stacking. Averaging involves calculating the mean prediction of all models. Majority voting aggregates the predictions of each model, with the class receiving the most votes being selected. Unlike simple averaging, weighted averaging assigns different weights to models based on their performance. Stacking, on the other hand, uses the outputs of multiple models as inputs for a new model to make the final predictions. Florian et al. (2003) combined four different classifiers based on a linear classifier, a maximum entropy model, transformation-based learning, and an HMM. They explore various combination strategies such as weighted voting and equal voting to optimise performance under different conditions. Akkasi and Varoğlu (2017) introduced a two-step approach by generating diverse baseline classifiers utilising CRF with distinct feature sets and employing Particle Swarm Optimisation (PSO) and Bayesian combination techniques to efficiently select and merge these classifiers.

6.1.9 Knowledge distillation

Knowledge distillation (KD) is a model compression method in which a smaller model is trained to mimic the behaviour of a larger pre-trained model or a group of models. Initially proposed by Buciluǎ et al. (2006) and later expanded by Hinton et al. (2015), this approach is often referred to as “teacher-student" training. In this process, knowledge is transferred from the larger teacher model to the smaller student model by minimising the KL divergence between their predictions. Zhou et al. (2021b) proposed a multi-grained knowledge distillation by utilising k-best predictions from the Viterbi algorithm to distil knowledge from the teacher model to the student. Additionally, CRF adjustments, fuzzy objective and data augmentation were incorporated to improve distillation performance. In the biomedical domain, publicly available datasets often differ in entity types, leading to inadequate ground truth for training multi-task models. To address this issue, Moscato et al. (2023a) introduced TaughtNet, a method that facilitates fine-tuning a single multitask student model utilising ground truth data and knowledge from single task teachers.

6.1.10 Summary

Table 7 shows a summary of the vanilla NER task categorised by author and different techniques. Techniques used for vanilla NER tasks are categorised into nine categories, which include data augmentation, document-level context, external knowledge, multi-task learning, adversarial learning, transfer learning, self-attention, ensemble learning and knowledge distillation. Most of the papers use external knowledge as a technique to improve NER performance, followed by multi-task learning and self-attention. The dataset used for each paper and the F1 score achieved are also included.

Table 7 Summary for vanilla NER tasks categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	DA	DLC	EK	MTL	Adv	TFL	Att	EL	KD
Dai and Adel (2020)	I2B2	87.2	✓								
Chen et al. (2020a)	CoNLL-2003	91.83	✓								
Ke et al. (2023)	OntoNotes	72.75	✓								
Luoma and Pyysalo (2020)	CoNLL-2003	87.95		✓						✓	
Hu et al. (2019)	CoNLL-2003	92.96		✓							
Luo et al. (2020a)	CoNLL-2003	93.37		✓							
Chen et al. (2020b)	CoNLL-2003	92.68		✓		✓					
Yang et al. (2022b)	CoNLL-2003	94.53		✓							
Wang et al. (2021b)	WNUT-2017	60.45			✓						
Zhang et al. (2022b)	Ecommerce	83.61			✓						
Liu et al. (2019)	CoNLL-2003	92.75			✓						
Wu et al. (2020a)	CoNLL-2003	92.20			✓						
Aguilar et al. (2018)	WNUT-2017	45.55			✓						
Bai et al. (2020)	CoNLL-2003	92.86			✓		✓				
Jie and Lu (2019)	OntoNotes−5.0	89.88			✓						
Xu et al. (2021a)	OntoNotes−5.0	90.85			✓						
Nie et al. (2020)	OntoNotes−5.0	90.32			✓						
Liu et al. (2018)	CoNLL-2003	91.85				✓					
Hu et al. (2021)	CoNLL-2003	92.8				✓					
Zhong et al. (2022)	CoNLL-2003	94.12				✓					
Reich et al. (2022)	CoNLL-2003	90.53	✓				✓				
Wang et al. (2020c)	CoNLL-2003	93.45				✓	✓				
Wang et al. (2018b)	Chinese Medical	71.15						✓			
Gligic et al. (2020)	I2B2	94.6						✓			
Devlin et al. (2019)	CoNLL-2003	92.8							✓		
Yamada et al. (2020)	CoNLL-2003	94.3			✓				✓		
Wu et al. (2023a)	WNUT-2017	57.3					✓		✓		
Florian et al. (2003)	IBM dataset	91.63								✓	
Akkasi and Varoğlu (2017)	ChemDNER	87.02								✓	
Zhou et al. (2021b)	CoNLL-2003	91.17	✓								✓
Moscato et al. (2023a)	NCBI-disease	89.20				✓					✓

DA data augmentation, *DLC* document-level context, *EK* external knowledge, *MTL* multi-task learning, *Adv* adversarial learning, *RL* reinforcement learning, *TFL* transfer learning, *Att* self-attention, *EL* ensemble learning, *KD* knowledge distillation, *Meta* meta learning

6.2 Nested and discontinuous NER

Unlike traditional flat NER that extracts entities as separate and non-overlapping spans, nested and discontinuous NER identifies hierarchical named entities within text which are nested, overlapping, or discontinuous entities. "British officials" includes both a person entity, "British officials", and a geopolitical entity, "British", and is an example of a nested entity. The sentence "productive cough with white or bloody sputum" contains two discontinuous, overlapping entities of "productive cough white sputum" and "productive cough bloody sputum". Approaches related to nested and discontinuous NER can be divided into layered-based approach, region-based approach, generation-based approach, hypergraph-based approach, transition-based approach, constituency parsing and other approaches.

6.2.1 Layered-based approach

Layered-based approach employs models typically composed of multiple layers reflecting the hierarchical structure inherent in nested named entities. Each layer focuses on identifying a cluster of named entities, which can refer to entities at specific levels or of certain lengths. Ju et al. (2018) introduced a dynamic stacking approach to identify nested entities using flat NER layers. Their model combines Bi-LSTM and cascaded CRF to capture sequential context, updating the representation of detected entities at each layer and passing them to the next layer until no more outer entities are found. However, conventional layered schemes do not handle the broader overlapping scenario and suffer from layer disorientation. Wang et al. (2020a) introduced the Pyramid architecture where token or text region embeddings are processed through a series of flat NER layers arranged in a pyramid structure. To facilitate bi-directional interaction between layers, an inverse pyramid design is also implemented. Fisher and Vlachos (2019) deconstructed nested NER into two stages: token merging into entities at Level 1, followed by further merging with tokens or entities at higher levels. Luo and Zhao (2020) presented BiFlaG by combining a flat NER module to identify outermost entities with a graph module to identify entities in inner layers. Shibuya and Hovy (2020) proposed to model the tag sequence for nested entities as the second-best path within the span of their parent entity and a decoding method that identifies entities iteratively, starting from the outermost to the innermost entities. Yang et al. (2021) proposed HiTRANS which deconstructs sentences into multi-grained spans and enhances representation learning hierarchically. Specifically, a two-phase module aggregates context information using bottom-up and top-down transformer networks to generate span representations for each layer. Subsequently, a label prediction layer is designed to hierarchically recognise nested entities. Kim and Kim (2024) introduced a recursive label attention network designed to explicitly reflect nested levels and efficiently utilise lower-level label information through level-specific label embeddings.

6.2.2 Region-based approach

The region-based approach for nested NER treats the task as a multiclass classification problem. This approach is divided into boundary-based and enumeration-based approaches, to represent potential regions (subsequences or spans) before classifying them. The typical region-based approach employs a boundary-based strategy to establish representations of candidate regions (potentially entities) utilising boundary information, followed by entity classification. Zheng et al. (2019) introduced a boundary-aware model based on Bi-LSTM for joint training of entity boundary detection and categorical label prediction using multi-task learning. To resolve the lack of explicit boundary supervision in span-based methods, Tan et al. (2020b) introduced a multi-task neural span classification model enriched with boundary detection task augmenting span representation through additional boundary supervision. Li et al. (2020a) proposed the Recursively Binary Modification model, which utilises modification relationships between sub-entity types to infer the head component within a Bayesian framework. The recursive approach allows lower-level entities to improve the modelling of higher-level entities. Wang et al. (2020b) presented a model designed to proficiently detect nested named entities by modelling the boundary tokens or "head-tail pair" and the relationships between tokens within those boundaries as “token interaction". Li et al. (2021b) introduced SESNER that frames the nested NER task as a segment covering problem. This approach models entities as segments, detects segment endpoints, and identifies positional relationships for span classification. To better correlate semantics between words under different entity types, Xu et al. (2021b) used a supervised multi-head self-attention mechanism where each head is dedicated to each entity type and adaptively predicts the span type by evaluating the intensity of correlations between the head and tail under the corresponding entity type.

Other region-based approaches employ an enumeration-based strategy for nested NER. Sohrab and Miwa (2018) explicitly enumerated all potential spans derived from input sentences, which are subsequently input into a classifier for category prediction through multitask learning. Xia et al. (2019) proposed MGNER which firstly identifies entity positions across different granularities using a Detector through span enumeration, followed by the classifying of these entities. Li et al. (2021a) proposed a span-based model by enumerating over all possible text spans to obtain entity fragments followed by relation classification to predict if a given pair of entity fragments are overlapping or succession.

6.2.3 Generation-based approach

Generation-based approach reformulates nested NER as a sequence generation problem. Straková et al. (2019) viewed nested NER as a sequence-to-sequence challenge and employ hard attention on the word whose label is being predicted. Fei et al. (2021) proposed using pointer networks where the memory-augmented pointer decides at the same time if a token at each decoding step represents an entity mention and where the next token is. As earlier methods overlook the order of recognition and the boundary position relationships of nested entities, Yang et al. (2023) proposed GPRL which employs reinforcement learning to create entity triplets independent of the entity order in the gold labels, with the aim of determining an effective recognition sequence for entities through a process of trial and error.

6.2.4 Hypergraph-based approach

A hypergraph is an extended form of a traditional graph, distinguished by edges that can link an arbitrary number of vertices. Lu and Roth (2015) utilised mention hypergraphs to recognise overlapping mentions. Muis and Lu (2016) introduce a hypergraph model that utilises mention separators to mark gaps between words, enabling the recognition of overlapping mentions. Katiyar and Cardie (2018) modified the top-hidden layer of the decoder for a standard Bi-LSTM model to learn the nested entity hypergraph structure for an input sentence followed by entity classification. Huang et al. (2021a) proposed HGN that employs encoders to learn a hypergraph representation followed by tagging each hyperedge based on the entity type. As complex hypergraphs can pose challenges during training, Yan et al. (2023a) introduced the LHBN, which constructs several smaller, local hypergraphs to capture named entities instead of relying on a single large and complex hypergraph. Hypergraph-based methods offer flexibility and are adept at modelling various types of nested structure. However, their computational demands are influenced by dataset-specific characteristics such as sequence length, maximum length, depth of entity mentions, and the number of potential entity labels. Consequently, they may become computationally inefficient, particularly for larger datasets featuring complex entity structures.

6.2.5 Transition-based approach

Transition-based parsing constructs syntactic or dependency parse trees by applying predefined actions to a configuration of Stack, Buffer, and Dependencies. Initially, all sentence tokens reside in the buffer, with an empty stack and no dependencies. Transitions modify this state: Shift moves tokens from buffer to stack, Reduce combines stack items into units or completes dependencies, Left/Right Arc links stack items, and No-op proceeds without change. These actions iteratively build a parse structure until completion. Wang et al. (2018a) presented a neural transition-based method by representing sentences with nested entities as a forest structure. The system builds this structure from the bottom up using a sequence of three transition actions: SHIFT, REDUCE, and UNARY, and relies on a stack to temporarily hold processed elements. Dai et al. (2020) proposed a transition-based model using a set of transition actions of six actions based on (SHIFT, OUT, COMPLETE, REDUCE, LEFT-REDUCE, RIGHT-REDUCE). Ji et al. (2021) proposed NeuJoRN for disease NER and normalisation by defining the task as predicting an action sequence. For recognition, they introduce four actions (OUT, SHIFT, REDUCE, SEGMENT). Additionally, a normalisation action (LINKING) is proposed to link recognised entities to standard concepts.

6.2.6 Constituency parsing

Constituency parsing is a technique within natural language processing to analyse the grammatical structure of sentences. It is a type of syntactic parsing, focusing on identifying the constituents or sub-parts within a sentence and determining their relationships. Typically, a constituency parser generates a parse tree as its output, depicting the hierarchical relationship of the sentence's components. Finkel and Manning (2009) proposed a discriminative constituency parser which transforms each sentence into a tree structure, where constituents represent each named entity. Fu et al. (2021) approached nested NER as a form of constituency parsing using partially-observed trees, treating labelled entity spans as observed nodes and other spans as latent nodes within a constituency tree. However, the method of Fu et al. (2021) did not utilise entity heads, which can assist in entity mention detection and typing. In contrast, Lou et al. (2022) introduced a more sophisticated approach using lexicalised constituency trees where constituents are annotated with headwords to model nested entities. Yang and Tu (2022) introduced a pointing mechanism within a bottom-up parsing framework. By leveraging the insight that consecutive spans share boundaries in a post-order traversal of a constituency tree, their model utilises a pointer network to track and predict these shared boundaries iteratively.

6.2.7 Other approaches

Besides the above approaches, other researchers have explored other techniques for nested and discontinuous NER. Straková et al. (2019) unify multiple labels of nested entities into a single multi-label, which is subsequently predicted using the LSTM-CRF model. Rojas et al. (2022) train several flat NER models, each dedicated to a specific entity type followed by combining the outputs to predict entity labels. Li et al. (2022b) introduce W2NER, which treats NER as word-word relation classification within a 2D grid. They propose multi-granularity 2D convolutions to enhance grid representations before inferring word-word relations.

6.2.8 Summary

Table 8 summarises nested and discontinuous NER tasks based on the author and different techniques. The techniques used for nested and discontinuous NER tasks are categorised into five major categories, including layer-based approach, region-based approach, generation-based approach, hypergraph-based approach, transition-based approach, constituency parsing and other approaches. Most of the papers belong to the nested NER task and few papers to discontinuous NER. Based on the techniques surveyed, the majority falls under the region-based approach, followed by the generation-based approach and the hypergraph-based approach. The dataset used for each paper and the F1 score achieved are also included.

6.3 Joint NER and relation extraction

Joint NER and RE task combines the NER and RE tasks where NER identifies and categorises specific entities in a given text while RE focuses on discerning semantic relationships between these identified entities. Generally, a pair of entities and their relation are defined as a relational triplet (Li and Ji 2014), for example $\langle$ Paris, France, Located_in$\rangle$. The traditional method for extracting relational triplets involves first identifying named entities and then classifying their relationships. This sequential process, termed the pipeline method Chen and Guo (2022), is simple but does not allow two sub-tasks to interact, which can result in the propagation of errors (Li and Ji 2014). However, the relationship between two entities is typically closely related to the entities themselves. Research indicates that jointly extracting entities and their relations yields a more promising performance compared to the pipeline approach.

6.3.1 Multi-task learning

As Joint NER and RE consists of two subtasks of NER and RE, it is a natural approach to adopt multi-task learning for this joint task. Miwa and Bansal (2016) jointly trained a LSTM model for NER with a tree-based dependency LSTM layer for RE using multi-task learning. Zheng et al. (2017a) introduced a hybrid neural network comprising a NER module and a RE module, sharing a Bi-LSTM encoding layer. Then an LSTM layer explicitly model tag interactions. In contrast to Miwa and Bansal (2016), the Bi-LSTM encoding layer in the Zheng et al. (2017a) model can capture contextual information about entities, which aids in identifying relationships between them. However, the parameter-sharing approach through multi-task learning leads to a significant amount of redundant information, and the potential association features between entities and relations may not be fully utilised.

Table 8 Summary for nested and discontinuous NER tasks categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	Nested	Disc	Layer	Region	Gen	HG	Trans	CP	Others
Ju et al. (2018)	GENIA	74.7	✓		✓						
Wang et al. (2020a)	ACE-2005	79.42	✓		✓						
Fisher and Vlachos (2019)	ACE-2005	74.6	✓		✓						
Luo and Zhao (2020)	ACE-2005	75.1	✓		✓						
Shibuya and Hovy (2020)	ACE-2005	84.34	✓		✓						
Kim and Kim (2024)	ACE-2005	87.19	✓		✓						
Yang et al. (2021)	ACE-2005	87.04	✓		✓						
Zheng et al. (2019)	GENIA	74.7	✓			✓					
Tan et al. (2020b)	GENIA	78.3	✓			✓					
Li et al. (2020a)	GENIA	79.8	✓			✓					
Wang et al. (2020b)	GENIA	76.2	✓			✓					
Li et al. (2021b)	ACE-2004	77.0	✓			✓					
Xu et al. (2021b)	GENIA	79.6	✓			✓					
Sohrab and Miwa (2018)	GENIA	77.1	✓			✓					
Li et al. (2021a)	CADEC	69.5	✓	✓		✓					
Xia et al. (2019)	ACE-2005	78.2	✓			✓					
Fei et al. (2021)	CADEC	72.4	✓	✓			✓				

Table 8 (continued)

Paper	Dataset	F1 score	Nested	Disc	Layer	Region	Gen	HG	Trans	CP	Others
Yang et al. (2023)	GENIA	81.45	✓				✓				
Lu and Roth (2015)	ACE-2005	62.5	✓					✓			
Muis and Lu (2016)	SHEL-2013+SemEval-2014	59.2	✓	✓				✓			
Katiyar and Cardie (2018)	ACE-2005	70.5	✓					✓			
Huang et al. (2021a)	ACE-2005	78.9	✓					✓			
Yan et al. (2023a)	ACE-2005	87.83	✓					✓			
Wang et al. (2018a)	ACE-2005	73.0	✓						✓		
Ji et al. (2021)	NCBI	88.57	✓						✓		
Dai et al. (2020)	CADEC	69.0		✓					✓		
Finkel and Manning (2009)	GENIA	70.33	✓							✓	
Fu et al. (2021)	ACE-2005	85.4	✓							✓	
Lou et al. (2022)	ACE-2005	86.91	✓							✓	
Yang and Tu (2022)	ACE-2005	85.53	✓							✓	
Straková et al. (2019)	ACE-2005	84.33	✓				✓				✓
Rojas et al. (2022)	GENIA	77.6	✓								✓
Li et al. (2022b)	ACE-2005	86.79	✓								✓

Disc discontinuous, *Layer* layered-based, *Region* region-based, *Gen* generation-based, *HG* hypergraph-based, *Trans* transition-based, *CP* constituency parsing

6.3.2 Table filling

Multi-task learning enables sharing model parameters between NER and RE sub-task. However, it cannot completely eliminate the propagation of errors as the RE process still pairs entities obtained from the NER task for RE. Miwa and Sasaki (2014) proposed using table-filling method where a table is constructed first with each cell representing a pair of words from a sentence. The main diagonal of the table contains BILOU tags for individual tokens, while off-diagonal cells capture relations between token pairs. Relations are obtained by examining the last words of the entities involved. Gupta et al. (2016) followed a similar table-filling approach to Miwa and Sasaki (2014) but they use a bidirectional recurrent neural network to assign labels to each word pair.

6.3.3 Tagging scheme

Table-filling approach helps to reduce the effects of error propagation in Joint NER and RE. However, it still requires pairing up all sentence components within a table leading to substantial redundancy. To address this limitation, Zheng et al. (2017b) designed a tagging scheme to jointly label the entities and their relations and propose a Bi-LSTM encoder and LSTM decoder with biased loss. Qiao et al. (2022) adopted the same tagging scheme proposed by Zheng et al. (2017b) and introduce BERT on top of the Zheng et al. (2017b) model to better model contextual dependencies in the agricultural domain.

6.3.4 Span-based approach

Span-based approaches conduct a thorough search across all spans, allowing the handling of overlapping (often nested) entities for Joint NER and RE task. Dixit and Al-Onaizan (2019) utilised concatenated ELMo, word, and character embeddings as input to Bi-LSTM, followed by span representations generation. Luan et al. (2018b) presented a multi-task, span-based model for classifying entities, relations, and coreference clusters in scientific texts using a beam search approach. Luan et al. (2018b) further extend their work with DyGIE, introduced in a subsequent study (Luan et al. (2019)), which incorporated a graph propagation step to capture span interactions using a dynamic span graph. More recently, Wadden et al. (2019) proposed DyGIE++ which uses the BERT encoder in place of the Bi-LSTM encoder in DyGIE.

6.3.5 Hypergraph-based approach

Previous approaches, which are based on span-based methodologies, often focus on intricately modelling potential entity regions, neglecting crucial contextual cues. In response to this limitation, Wan et al. (2021) proposed RHGN which uses GCN and Bi-LSTM to generate hypernodes for each region to construct a region-based relation hypergraph. A sequence-enhanced graph (SEG) unit is designed to initialise and improve the features of the edges and hypernodes in the hypergraph. To mitigate error propagation issues, Yan et al. (2023b) proposed HGERE that constructs a hypergrah where nodes represent entities and their corresponding relations, while hyperedges model interactions between different relations or between a relation and entities. A high-recall pruner mechanism is also proposed to remove unlikely span candidates.

6.3.6 Summary

Table 9 shows a summary of the Joint NER and RE categorised by author and different techniques. We categorise the techniques used into five main categories and also other approaches. 5 main categories include multi-task learning, table filling, tagging scheme, span-based and hypergraph based approach. Most of the papers surveyed in Joint NER and RE task are based on span-based approaches. We also included the dataset used for each paper and the F1 score achieved.

7 Other NER approaches

7.1 Low resource NER

Low-resource NER addresses the challenge of performing NER with limited labelled data in specific domains or languages. To address this challenge, researchers employ techniques that include data augmentation, curriculum learning, adversarial learning, incorporating external knowledge, and in-context learning.

7.1.1 Data augmentation

Data augmentation generates synthetic data to address data scarcity in low-resource scenarios. However, these methods often result in token-label misalignment for token-level tasks. To address this problem, synthetic data is generated from a model trained on sequences where each word is mapped to the correct label. Ding et al. (2020) proposed a generation method leveraging labelled sequence linearisation by pairing words with their corresponding tags, inserting tags before (or after) the words while excluding O tags. Subsequently, a language model is trained using linearised data to generate synthetic labelled data. However,

Table 9 Summary for joint NER and RE tasks categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	MTL	TF	TS	Span	HG
Miwa and Bansal (2016)	ACE-2005	55.6	✓				
Zheng et al. (2017a)	ACE-2005	56.5	✓				
Miwa and Sasaki (2014)	CoNLL-2004	61.0		✓			
Gupta et al. (2016)	CoNLL-2004	72.1		✓			
Zheng et al. (2017b)	NYT	52.0			✓		
Qiao et al. (2022)	NYT	55.7			✓		
Luan et al. (2018b)	SCIERC	39.3	✓			✓	
Luan et al. (2019)	ACE-2005	63.2				✓	
Dixit and Al-Onaizan (2019)	ACE-2005	62.83				✓	
Wadden et al. (2019)	ACE-2005	63.4				✓	
Wan et al. (2021)	CoNLL-2004	70.53					✓
Yan et al. (2023b)	ACE-2005	70.8					✓

MTL multi-task learning, *TF* table filling, *TS* tagging scheme, *Span* span-based, *HG* hypergraph

DAGA (Ding et al. 2020) generated synthetic data with both context and entities generated simultaneously which may suffer from generating ungrammatical and unfluent sentences using an untrained language model. Unlike Ding et al. (2020), Zhou et al. (2022d) introduced MELM which focuses on modifying entity tokens without changing context and exhibits better performance in low-resource scenarios. MELM (Zhou et al. 2022d) aims to mitigate token label misalignment problems during data augmentation by explicitly integrating NER labels into sentence context via labelled sequence linearisation followed by fine-tuned on these linearised sequences. Specifically, the sentence is corrupted as $\tilde{X}$ as input, MELM is trained to maximise the sum of probabilities for each word x_i in the sequence in the labelled training data. In this way, we can reconstruct the linearised sequence X using Eq. 20 where θ represents the parameters of MELM, n is the number of tokens in $\tilde{X}$, x_i is the original token in X, m_i = 1 if x_i is masked and otherwise m_i = 0. Liu et al. (2022a) leveraged BERT's knowledge for data augmentation using two prompting strategies based on label-conditioned word replacement through masked token prediction and prompting with question answering. Zeng et al. (2020) proposed the Counterfactual Generator, which enhances the original dataset by generating counterfactual examples through interventions on existing observational examples. Cai et al. (2023b) introduced GPDA which utilises graph propagation to propagate information from labelled data to unlabelled texts. A basic search engine is utilised to retrieve relevant texts related to the labelled data and propagates entity labels using anchor links.

$$\underset{\theta}{max}\ \log p_\theta(X|\tilde{X}) \approx \sum_{i=1}^{n} m_i \log p_\theta(x_i|\tilde{X}) \tag{20}$$

7.1.2 Curriculum learning

Synthetic data generated from data augmentation methods may suffer from noise that affects NER training. To alleviate noise from synthetic data, Curriculum learning (Bengio et al. 2009) can be used to train on these synthetic data using a progressive training strategy from simpler to more complex instances. Zhu et al. (2021) presented a framework that incorporates data augmentation and denoising techniques. Data augmentation leverages BERT (Devlin et al. 2019) to predict masked words based on the context, generating new sentences by replacing the original masked positions with these predictions to expand the training set. Furthermore, augmented data are de-noised using curriculum learning to enhance its quality. In another example, Lobov et al. (2022) used natural annotations to create synthetic training datasets, selectively choosing the most appropriate examples to improve NER performance through curriculum learning. They introduced a technique called “Natural" annotation, wherein desired linguistic properties are extracted from annotations generated incidentally during a natural activity, unrelated to the model's primary task.

7.1.3 Adversarial learning

Previous studies neglected the difference in representation between resources and enforced a shared feature representation across languages/domains. Fewer training sentences are available in low-resource languages such as Spanish compared to high-resource languages such as English. Addressing these issues, Zhou et al. (2019) proposed a new neural transfer method termed Dual Adversarial Transfer Network (DATNet) which unifies two types of adversarial learning, i.e., Generalised Resource-Adversarial Discriminator (GRAD) and Adversarial Training (AT), into one transfer learning model. GRAD impose resource weight to focus on hard examples, and AT introduce adversarial samples to improve generalisation and reduce overfitting.

7.1.4 External knowledge

External knowledge in the form of gazetteers can be incorporated to improve performance for low resource NER. Fetahu et al. (2022) proposed a method to encode sentences using the pre-trained XLMR model and enhance it with multilingual gazetteers from sources such as Wikidata and domain-specific resources. These gazetteers help in transferring NER knowledge and provide explicit signals about named entities in target languages or domains. The model combines information from both modules using a mixture of experts (MoE) to dynamically determine the relevant information for NER.

Entity triggers can be incorporated as external knowledge for low-resource NER. Entity triggers are defined as groups of words within a sentence that help explain why humans recognise entities. Lin et al. (2020) introduced entity triggers as external knowledge to enhance the NER task. They propose a Trigger Matching Network that integrates trigger representations with a soft matching module using self-attention. However, the Lin et al. (2020) approach generated the trigger representation that lacks detailed entity-specific information, limiting its effectiveness in entity identification. Furthermore, its attention-based fusion mechanism, a basic non-linear transformation, struggles to adequately integrate trigger information. Zhang et al. (2022d) improved the Lin et al. (2020) model and introduced LELNER, consisting of an information interaction module and an information fusion network. The information interaction module facilitates the interaction between triggers and sentences, enriching trigger representations with entity information. Meanwhile, the information fusion network effectively integrates these trigger representations into sentence sequences.

7.1.5 In-context learning

In-context learning (ICL) involves concatenating a query with a small set of few-shot demonstrations to prompt LLMs for prediction. Lee et al. (2022) introduced a demonstration-based learning method for low resource NER, which involves prefacing the input with task demonstrations to facilitate ICL. They conducted a systematic study of demonstration strategies, examining what to include (entity examples with or without a surrounding context), how to select the examples, and what templates to use. Wu et al. (2024) introduced ConsistNER comprising three stages that integrate ontological and contextual information for NER in low-resource settings. Initially, ConsistNER utilises LLMs to pre-identify potential entities in a zero-shot approach. Subsequently, it retrieves sentence-specific exemplars for each target sentence, focusing on ontological and contextual coherence. Finally, ConsistNER leverages these retrieved exemplars across all target sentences to prompt LLMs to make predictions.

7.1.6 Summary

Table 10 shows a summary of low resource NER tasks categorised by author and different techniques. We categorise the techniques used for low resource NER tasks into five main techniques including data augmentation, curriculum learning, adversarial learning, external knowledge, and ICL. We conclude that most previous authors used data augmentation for NER tasks in low resource domains. We also see that NER in English has a highest F1 score of 88%, while in Spanish it only 79%.

7.2 Cross-domain NER

Cross-domain NER aims to leverage entity information from one domain to assist in recognising entities in target domain where labelled data is scarce. A “domain" pertains to a specific subject area, such as news articles, medical texts, legal documents, social media posts, and scientific papers, each with its distinctive set of named entities and domain characteristics. In this section, we discuss three different approaches for cross-domain NER.

7.2.1 Data augmentation

Data augmentation is a effective technique to improve Cross-domain NER. Most existing techniques focus on augmenting in-domain data in low-resource scenarios where annotated data is quite limited. Chen et al. (2021b) focused on cross-domain data augmentation and proposed to map data from a high-resource domain to a low-resource domain using a cross-domain autoencoder. First, sentences in the source and target domain are linearised by inserting entity label before corresponding word. It utilise two methods of denoising reconstruction, which reconstructs each input sentence from its noisy version, and detransforming reconstruction, which reconstructs each input sentence from its transformed version in the opposite domain. Yang et al. (2022a) introduced FactMix, a method designed to enhance in-domain and out-of-domain (OOD) performance in cross-domain NER tasks.

Table 10 Comparison of F1 score and different methods for low-resource NER tasks. The most popular method is data augmentation and accuracy in English is higher than in Spanish or Chinese

Paper	Dataset	F1 score	DA	CurrL	Adv	EK	ICL
Ding et al. (2020)	CoNLL-2002/2003	81.02	✓				
Zhou et al. (2022d)	CoNLL-2002/2003	87.59	✓				
Liu et al. (2022a)	CoNLL-2003	70.1	✓				
Zeng et al. (2020)	CNER	78.8	✓				
Cai et al. (2023b)	CrossNER	74.81	✓				
Zhu et al. (2021)	CoNLL-2003	61.48	✓	✓			
Lobov et al. (2022)	WikiNER	78	✓	✓			
Zhou et al. (2019)	CoNLL-2002	79.46			✓		
Fetahu et al. (2022)	mLOWNER	77.2 (Avg)				✓	
Lin et al. (2020)	CoNLL-2003	86.5				✓	
Zhang et al. (2022d)	BC5CDR	75.52				✓	
Lee et al. (2022)	CoNLL-2003	65.11					✓
Wu et al. (2024)	CoNLL-2003	78.87					✓

DA data augmentation, *CurrL* curriculum learning, *Adv* adversarial learning, *EK* external knowledge, *ICL* in-context learning

FactMix employs a two-step augmentation process of entity-level semi-fact generation utilising prepared entity knowledge bases and context-level semi-fact generation based on masked token prediction. Zhang et al. (2023e) introduced SLC-DA, a data augmentation method using label-constrained pre-training task and a structure-constrained optimisation objectives. These strategies aim to generate domain-specific augmented data, facilitating the seamless adaptation of NER models from source to target domains.

7.2.2 Domain adaptation

Domain adaptation is a subcategory of transfer learning which aims to fill the gap between source data and target data. It is the ability to apply an algorithm that is trained on one or more source domains to a different target domain. In domain adaptation, the source and target data have the same feature space but from different distributions, while transfer learning includes cases where target feature space is different from source feature space. Lin and Lu (2018) introduced efficient techniques for conducting domain adaptation with neural NER models using lightweight methods, including sentence and output adaptation layers integrated into existing neural architectures. Jia et al. (2019) suggested utilising cross-domain language modelling as a bridge between domains for adapting NER tasks. This entails enabling knowledge transfer across domains and tasks by devising a novel parameter generation network. Liu et al. (2021) performed Domain-Adaptive Pre-training (DAPT) by continue pre-training the language model BERT (Devlin et al. 2019) on the unlabelled corpus for the domain adaptation. They also investigated the influence of different levels of the corpus on pre-training and the effectiveness between token-level and span-level masking in the DAPT. Jia and Zhang (2020) examined a multi-cell compositional LSTM architecture for multi-task learning, where each entity type is represented by its own cell state. Through the incorporation of entity-typed units, cross-domain knowledge transfer can occur at the level of entity types. Li et al. (2020b) proposed MetaNER, a technique for domain adaptation using meta learning for sequence labelling in NER. It leverages meta-learning and adversarial training methods to generate robust, general and transferable representations capable of adapting to unseen domains with small amount of annotated data. As data is typically fully-unlabelled in a completely new domain, Peng et al. (2021) introduced an unsupervised cross-domain model that utilises labelled data from a source domain to predict entities in an unlabelled target domain through adversarial training and an entity-aware attention module to guide the adversarial training process. Hu et al. (2022) introduced an autoregressive cross-domain NER framework for domain adaptation which enhances the connection between the source text and its named entity labels while improving the transfer of label information. However, cross-domain NER methods overlook the direct alignment of input word distributions between domains, a crucial aspect in word-level classification tasks like cross-domain NER. Ma et al. (2022c), introduced X-Piece, a subword-level domain adaptation method to address the shift in input word-level distribution in NER. Specifically, the input words from the source domain are re-tokenised to approximate the target subword distribution, treating it as an optimal transport problem. Instead of aligning subword distribution, Hu et al. (2023) suggested incorporating subsequence-level features to enhance feature adaptation for cross-domain NER to help the model distinguish different meanings of the same word in different domains. Chen et al. (2023b) proposed another approach of Collaborative Domain-Prefix Tuning for cross-domain NER (CP-NER) based on text-to-text generative PLMs with frozen

PLMs. They utilise text-to-text generation grounding domain-related instructors to transfer knowledge to new domain NER tasks without structural modifications. Several researchers use dependency information which is more consistent across domains for domain adaptation. Dou et al. (2023) introduced unsupervised domain-adapted method to transfer word-dependency knowledge from high-resource domains to low-resource ones for cross-domain NER. A multi-task learning framework is introduced which utilised Cross-domain Dependency Parsing (DP) as auxiliary learning task. To make better use of the cross-task knowledge between NER and DP, both tasks is unified in a shared network architecture for joint learning, using Maximum Mean Discrepancy (MMD).

7.2.3 Task decomposition

The majority of current cross-domain NER approaches are constructed within the sequence labelling framework, treating entity detection and type prediction as a unified process. However, the differing transferability of these subtasks are often overlooked: entity detection is generally robust across domains, while entity types vary significantly. Integrating them into a single learning task might increase the complexity of domain transfer. Hence, researchers decompose NER into its respective subtasks to facilitate cross-domain knowledge transfer. Zhang et al. (2022c) explored task decomposition in cross-domain NER into two subtasks (entity span detection and type classification) that are learned by separate functional modules to perform respective cross-domain transfer. Then the two subtasks are combined to achieve the final result with a modular interaction mechanism, and adversarial regularisation for generalised and robust learning in low-resource target domains. Hu et al. (2024) explored a similar approach to task decomposition in which the potential named entities obtained by the source domain models are first copied after the target domain sentence. Second, the embeddings predicted by the source domain models are transferred to the target domain model through the Knowledge Progressive Networks.

7.2.4 Summary

Table 11 shows a summary of cross-domain NER tasks categorised by author and different techniques. We categorise the techniques used for cross-domain NER tasks into 3 main techniques including data augmentation, domain adaptation and task decomposition. Most of the papers surveyed for cross-domain NER task are based on domain adaptation, followed by data augmentation and task decomposition. The dataset used for both source domain and target domain for each paper are included with the F1 score achieved.

7.3 Cross-lingual NER

Cross-lingual NER aims to transfer knowledge from a source language with rich labelled data to a target language with little or even no labelled data. In this section, we will discuss different techniques for cross-lingual NER.

Table 11 Summary for cross-domain NER tasks categorised by source domain, target domain, F1 score, and different techniques

Paper	Source domain	Target domain	F1 score	DA	DAdapt	TD
Chen et al. (2021b)	OntoNotes−5.0	Temporal Twitter	44.82	✓		
Yang et al. (2022a)	CoNLL-2003	CrossNER	74.62	✓		
Zhang et al. (2023e)	OntoNotes 5.0	CoNLL-2003	81.7	✓		
Lin and Lu (2018)	OntoNotes−5.0	Ritter11	66.40		✓	
Jia et al. (2019)	CoNLL-2003	BioNLP13PC	85.54		✓	
Liu et al. (2021)	CoNLL-2003	CrossNER	69.63		✓	
Jia and Zhang (2020)	CoNLL-2003	Broad Twitter	78.43		✓	
Peng et al. (2021)	CoNLL-2003	Twitter	64.1		✓	
Hu et al. (2022)	CoNLL-2003	CrossNER	74.06		✓	
Chen et al. (2023b)	CoNLL-2003	CrossNER	74.25		✓	
Dou et al. (2023)	OntoNotes−5.0	NCBI	86.42		✓	
Li et al. (2020b)	Multiple	BioNLP13PC	85.11		✓	
Ma et al. (2022c)	CoNLL-2003	OntoNotes−5.0	79.54		✓	
Hu et al. (2023)	CoNLL-2003	CrossNER	73.82		✓	
Zhang et al. (2022c)	CoNLL-2003	CrossNER	79.52			✓
Hu et al. (2024)	Multiple	CrossNER	77.82			✓

DA data augmentation, *DAdapt* domain adaptation, *TD* task decomposition

7.3.1 Instance-based transfer

Instance-based transfer generally refers to annotation projection using parallel corpora. Parallel corpora consist of texts in source and target languages that are translations of each other. By aligning named entities in parallel texts, the model can transfer knowledge from one language to another. Ni et al. (2017) introduced two weakly supervised methodologies for cross-lingual NER, without human annotation in the target language. The first approach involves automatically labelled NER data generation for the target language through annotation projection on comparable corpora. The second method involves projecting distributed word representations (word embeddings) from the target language to a source language. However, parallel data are hard to obtain and researchers utilise machine translation systems to generate parallel corpora. Mayhew et al. (2017) employed a cheap lexicon-based translation technique to generate training data in the target language by translating the source data. The lexicon comprises entries that encompass word-to-word translations, as well as word-to-phrase, phrase-to-word, and phrase-to-phrase translations. Xie et al. (2018) trained separate word embeddings using monolingual corpora.

The embedding from the source and target languages are mapped to a common space. For translating a word we look at the nearest neighbour in the English language. The named entity tag for the English word is used as a label for training an NER model in the target language. Jain et al. (2019) proposed an entity-projection system leveraging machine translation twice: first, to translate sentences and then to translate and match entities using orthographic, phonetic similarity, and distributional statistics without parallel corpora. However, this approach suffered from noisy pseudo-labels generated during the automatic labelling process. Other researchers explored methods to mitigate this problem. Zhou et al. (2022c) proposed two consistency training methods based on translation-based consistency training on unlabelled target-language data and dropout-based consistency training on labelled source language data for consistent predictions between tokens in the original sentence and their projection in the translated sentence. Ma et al. (2023c) introduced CoLaDa, a Collaborative Label Denoising Framework consisting of a model-collaboration-based denoising scheme and an instance-collaboration-based strategy to improve label consistency in token neighbourhoods within the representation space.

7.3.2 Model-based transfer

The frequency of named entities in a new language is very low. Hence, model-based (Wu and Dredze 2019; Wu et al. 2020c) transfer methods use a pre-trained model such as multilingual BERT that has been trained on over 104 different languages and hence contains language-independent features. Karthikeyan et al. (2019) conducted an empirical investigation into the role of various components within M-BERT concerning its cross-lingual capabilities. They examined how linguistic characteristics of languages, model architecture, and learning objectives influence its performance. For Wikipedia, we can match the important entity mentions in a new language to an English document instead of translating the entire document which is extremely time consuming (Tsai et al. 2016). When provided with a mention (substring) from a document composed in a foreign language, cross-lingual wikification aims to identify the corresponding title in the English Wikipedia. Recently, Wu et al. (2021) combined model-based and instance-based learning by dividing the dataset into smaller groups based on similarity or rules.

7.3.3 Representation alignment

Representation alignment is a relationship between the labels and the a representation matrix. Here, instead of using the original labels we project them on to the principal components of the data resulting in pseudo labels. There are two main methods for representation alignment based on Contrastive learning and Adversarial training. Contrastive learning is a technique that can be applied to cross-lingual NER to align both semantic and token-level representations across diverse languages. To alleviate the noisy pseudo-labelled target language data during self-training, Zhou et al. (2023a) introduced ContProto that utilises contrastive self-training, which enhances span representations through supervised contrastive learning as shown in Eq. 21 where A(i) ≡ {1, 2,..., 2 m}, and P(i) ≡ {p ϵ A(i): y_i = y_p} are indices of the positive sample set consisting of spans sharing the same label as s_i and ζ_i, ζ_p and ζ_a are the projected representations. Subsequently, prototype-based pseudo-labelling, gradually refine the quality of pseudo labels using prototype learning. Mo et al. (2024) proposed Multi-view Contrastive Learning for Cross-lingual NER (MCL-NER) which applies contrastive learning between source, codeswitched, and target sentences and contrasts among token-to-token relations in Eq. 22 where R(.) earns relation representation, F(.) gets semantic representation, R(x_i, x_j) - R(y_a, y_b) is relation distance, F(x_i, x_j) - F(y_a, y_b) is representation distance.

$$L_{cont} = -\frac{1}{2m}\sum_{i=1}^{2m}\frac{1}{|P(i)|}\sum_{p\epsilon P(i)}\log\frac{exp(\zeta_i.\zeta_p/\tau)}{\sum_{a\epsilon A(i)} exp(\zeta_i.\zeta_a/\tau)} \tag{21}$$

$$\min(\mid R(x_i, x_j) - R(y_a, y_b) \mid + \mid F(x_i, x_j) - F(y_a, y_b) \mid) \tag{22}$$

Besides contrastive learning, adversarial training is also an effective technique to align cross-lingual word representations. For example, in Huang et al. (2019a) a discriminator is trained to predict the language of a word using word embeddings of the target language and a transformed embedding of the source language. Shared words are those that are incorrectly classified by the discriminator. Bari et al. (2020) presented an unsupervised cross-lingual NER model through word-level adversarial learning and augmented fine-tuning, employing parameter sharing and feature augmentation methods. Chen et al. (2019) proposed utilising multiple source languages to learn both language-invariant and language-specific features at the instance level through adversarial networks and mixture-of-experts models. Chen et al. (2021c) introduced AdvPicker, employing adversarial learning to identify language-independent pseudo-labelled data for training a proficient NER model in a target language. Other methods were also proposed for cross-lingual alignment. Huang et al. (2023) introduced PRAM, a prototype-based representation alignment model which align entity representations, predictions and languages using a training objective, Attribution-Prediction Consistency (APC).

7.3.4 Knowledge distillation

Some researchers utilise knowledge distillation to transfer knowledge from the source language to target language for cross-lingual NER. Wu et al. (2020b) studied the challenging scenario of cross-lingual NER, wherein there's no labelled data in the target language. Here a student model in a target language is trained to mimic the probability of each token in a teacher model trained using labelled data. For this we train the student model such that we can minimise the mean square error between the student and teacher model for each sentence averaged over all the tokens. To incorporate rich and complementary information lying in the intermediate layers of PLM, Ma et al. (2022b) introduced the Mixture of Short-channel Distillers (MSD) method for Zero-shot cross-lingual NER. Firstly, Mixture of Distillers is implemented to establish multiple channels between corresponding layers of the teacher and student encoders. Secondly, domain information is transferred between the teacher and student models during the distillation process. Li et al. (2022d) was the first to introduce a similarity metric model as an auxiliary task to improve the cross-lingual NER performance on the target domain with knowledge distillation. An entity recogniser and a similarity evaluator are first trained in parallel as two teachers from the source domain and used to supervise training for student model via knowledge distillation. Liang et al. (2021) introduced Reinforced Iterative knowledge distillation for cross-lingual NER to make good use of rich unlabelled data in target languages. They use a policy network predicting the usefulness of unlabelled examples, selectively incorporating them into the distillation process. Additionally, the student model from the previous round becomes the teacher model. Ge et al. (2023) proposed an unsupervised prototype knowledge distillation network (ProKD) using a contrastive learning-based prototype alignment method to achieve class feature alignment by adjusting the distance among prototypes in the source and target languages, boosting the teacher network's capacity to acquire language-independent knowledge. In addition, ProKD introduces a prototypical self-training method to learn the intrinsic structure of the language by retraining the student network on the target data using samples' distance information from prototypes, thereby enhancing the student network's ability to acquire language-specifc knowledge. Ge et al. (2024) proposed an discrepancy and uncertainty aware Denoising Knowledge Distillation model (DenKD) to reduce noise in pseudo-labels. Discrepancy-aware denoising representation learning method optimise the class representations of the target language produced by the teacher network, thus enhancing the quality of pseudo labels and reducing noisy predictions. Uncertainty-aware denoising method quantify the pseudo-label noise and adjust the focus of the student network on different samples during knowledge distillation, thereby mitigating the noise's adverse effects.

7.3.5 Meta learning

Meta learning is a technique that can be applied in data scarcity scenarios in cross-lingual NER. Wu et al. (2020d) leveraged MAML to tackle cross-lingual NER tasks in zero/low resource scenarios. They constructed a set of pseudo-meta-NER tasks using the labelled data from the source language and propose a meta-learning algorithm to find a good model parameter initialisation that could adapt to new tasks quickly. When it comes to the adaptation phase, each test example is regarded as a new task, build a pseudo training set for it, and fine-tune the meta-trained model before testing.

7.3.6 Summary

Table 12 shows a summary of cross-lingual NER categorised by author and different techniques. We categorise the techniques used for cross-lingual NER tasks into five main techniques including instance-based transfer, model-based transfer, representation alignment, knowledge distillation and meta learning. Most of the papers surveyed in cross-lingual NER task are based on instance-based transfer and representation alignment followed by knowledge distillation. The results for majority of papers are presented based on CoNLL-2002 and CoNLL-2003 dataset where the source language is English and target language is Spanish. Karthikeyan et al. (2019) and Ge et al. (2024) experimented with LORELEI and Wikiann dataset respectively. We also included the F1 score achieved for each dataset.

7.4 Zero-shot NER

Zero-shot NER task refers to learning NER models from entity classes in training data and predict target entity classes that are absent from the training dataset. The main techniques used for zero-shot NER are prompt-based methods and other techniques.

7.4.1 Prompt based methods

Prompt-based methods are commonly used in zero-shot NER. Xie et al. (2023) proposed a self-improving framework, which uses an unlabelled corpus to stimulate the self-learning ability of LLMs for zero-shot NER. Firstly, LLM is used to make predictions on the unlabelled corpus. Next, reliable demonstrations from the self-annotated set were selected using various methods. Lastly, inference is conducted on the test query via in-context learning with the selected pseudo demonstrations. For zero shot NER and RE, Lv et al. (2023) proposed a novel Discriminative Soft Prompts (DSP) approach which reformulates zero-shot

Table 12 Summary for cross-lingual NER tasks (Source language: English, Target Language: Spanish) on CoNLL-2002 and CoNLL-2003 dataset, categorised by dataset, F1 score, and different techniques (* indicate zero shot cross-lingual task)

Paper	F1 score	Inst	Model	RA	KD	MetaL
Ni et al. (2017)	65.18	✓				
Mayhew et al. (2017)	65.18	✓				
Xie et al. (2018)	72.37	✓				
Jain et al. (2019)	73.5	✓				
Zhou et al. (2022c)	80.50	✓				
Ma et al. (2023c)	82.70	✓				
Wu and Dredze (2019)	72.6 *		✓			
Tsai et al. (2016)	60.55 *		✓			
Karthikeyan et al. (2019)	64.8 (LORELEI)		✓			
Wu et al. (2020c)	76.75		✓			
Wu et al. (2021)	79.31	✓	✓		✓	
Zhou et al. (2023a)	85.02			✓		
Mo et al. (2024)	79.2			✓		
Huang et al. (2019a)	86.41			✓		
Bari et al. (2020)	75.93			✓		
Chen et al. (2019)	73.5 *			✓		
Chen et al. (2021c)	79.00			✓		
Li et al. (2022d)	81.82			✓		
Huang et al. (2023)	82.06 *			✓		
Ma et al. (2022b)	81.92				✓	
Liang et al. (2021)	77.84				✓	
Ge et al. (2023)	79.53				✓	
Ge et al. (2024)	84.68 (Wikiann)				✓	
Wu et al. (2020b)	76.94				✓	
Wu et al. (2020d)	76.75					✓

Inst instance transfer, *Model* model transfer, *RA* representation alignment, *KD* knowledge distillation, *DA* data augmentation, *MetaL* meta learning

tasks into token discrimination tasks without having to construct verbalisers. A soft prompt co-reference strategy is designed to improve inference speed.

7.4.2 Other approaches

Besides prompt-based methods, other techniques are also proposed for zero-shot NER. Aly et al. (2021) proposed several architectures for zero-shot NERC based on cross-attention between the sentence and the entity type descriptions using transformers combined with pre-training. Some authors explore zero-shot NER for cross-lingual applications. Eronen et al. (2023) explored transfer language selection based on linguistic similarities for zero-shot cross-lingual NER.

7.4.3 Summary

Table 13 shows a summary of the zero-shot NER tasks based on the author and different techniques. We categorise the techniques used for zero-shot NER tasks into prompt-based

Table 13 Summary for zero-shot NER tasks categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	Prompt	Others
Xie et al. (2023)	CoNLL-2003	74.99	✓	
Lv et al. (2023)	OntoNotes-ZS	31.6	✓	
Eronen et al. (2023)	Wikiann (German)	82.7		✓
Aly et al. (2021)	OntoNotes-ZS	45		✓

methods and other techniques. Most of the papers surveyed for the low resource NER task are based on prompt-based methods. We also included the dataset used for the target domain for each paper and the F1 score achieved.

7.5 Few-shot NER

Due to the emergence of knowledge from various domains, it is challenging to manually annotate named entities on a large scale, which sometimes requires domain expertise. Few-shot NER involves studying NER systems that could learn unseen entity types with few examples, reducing the effort required to manually annotate entities. There are several techniques used for Few-shot NER including meta learning, entity knowledge and contrastive learning.

7.5.1 Meta learning

Metric-based meta learning involves using a distance metric or similarity function to measure similarity among distinct examples or data points. Examples includes Matching Network (Vinyals et al. 2016), prototypical network (Snell et al. 2017), and the Relation Network (Sung et al. 2018). Fritzler et al. (2019) tackle NER task using Prototypical Network by learning intermediate representations of words that cluster well in named entity classes. The class protoypes c_k are computed as shown in Eq. 23 where S_k is the set of objects from S that belong to this class and f_θ is a function that maps the input texts to the M-dimensional space. In order to classify an unseen example x, x is mapped to the M-dimensional space using f_θ and then assigned to a class whose prototype is closer to the representation of x. The distance $d(f_\theta(x), c_k)$ is calculated for every k. The measure of similarity of x to k is defined as $l_i = -d(f_\theta(x), c_k)$. Finally, these similarities are converted to a distribution over classes using the softmax function. Several researchers incorporated entity knowledge into metric-based meta learning. As Prototypical networks typically suffer from roughly estimated label dependency and closely distributed prototypes, Ji et al. (2022) proposed EP-Net, an Entity-level Prototypical Network enhanced by dispersedly distributed prototypes. EP-Net builds entity-level prototypes and considers text spans to be candidate entities without label dependency. Wang et al. (2022b) proposed a span-based prototypical network (SpanProto) that tackles few-shot NER via a two-stage approach, including span extraction and mention classification. However, the decoding process requires careful handling of overlapping spans due to the nature of span enumeration. Fang et al. (2023) introduced MANNER for few-shot cross-domain NER by utilising representations from the support set and memory to infer prototype distributions using optimal transport. Subsequently, these prototypes are employed in the entity typing module to predict entity types for sentences in the query set. Additionally, a span detection module is used to predict positional tags within query sentences. The final label is derived by combining the predicted entity types and position tags. Feng et al. (2024) proposed a method that extracts type-agnostic span representations using a sequence labelling model, refines class prototypes with a triaffine transformation integrating textual hierarchy and local–global features, and employs taxonomy-instance contrastive learning to align entity spans with branch descriptions while minimising noise in class prototypes.

$$c_k = \frac{1}{\|S_k\|} \sum_i f_\theta(x_i) \tag{23}$$

Optimisation-based meta-learning involves explicitly learning an update rule or weight initialisation to facilitate rapid learning during meta-testing. Andrychowicz et al. (2016) and Ravi and Larochelle (2016) focused on training recurrent neural networks to improve the direction of vanilla gradient descent for improved optimisation outcomes. MAML, introduced by Finn et al. (2017), optimises model parameters to discover an optimal starting point, allowing the model to quickly and effectively adapt to new unseen tasks. MAML trains a model via an inner loop (task-specific adaptation) and an outer loop (meta-update across tasks) during meta-training, so that at meta-testing, the model can quickly adapt to new tasks using only a few updates on support data. Approaches such as FOMAML (Finn et al. 2017)) leveraged first-order derivatives to reduce the memory consumption associated with high-order derivative calculations. Some researchers propose to combine metric-based meta learning with optimisation-based meta learning. Ma et al. (2022e) introduced a decomposed meta-learning technique addressing few-shot NER by sequentially addressing few-shot span-detection using MAML and few-shot entity typing using MAML-enhanced prototypical networks, MAML-ProtoNet. To avoid handling overlapping spans, few-shot span detection is modelled as a sequence labelling problem. Only detected entity spans are fed to the typing model for entity class inference, and hence eliminating the problem of noisy “O" prototype.

Other meta-learning techniques for NER include adaptive sample re-weighting and meta-function pretraining for NER. Wang et al. (2021d) introduced a meta self-training framework that uses a minimal amount of manually annotated labels to train neural sequence models. Meta-learning aids in adaptive sample re-weighting to reduce error propagation from noisy pseudo-labels during self-training. Chen et al. (2023a) trained PLM to enhance their in-context NER capabilities by optimising them using a meta-function loss. This approach ensures that the extractor F, constructed implicitly through instruction and demonstration, closely approximates an explicitly fine-tuned surrogate golden extractor. By further optimising PLM with extraction loss, the method enables effective identification and classification of entities within textual contexts for in-context NER tasks.

7.5.2 Entity knowledge

Few-shot NER techniques are challenging in adapting to new entity types and are prone to the so-called negative transfer problem. This problem can be mitigated by integrating label knowledge into few-shot NER systems providing the model additional signal and enriched prior knowledge. Several researchers tried to incorporate label descriptions for few-shot learning. Wang et al. (2021c) decomposed the NER task into two sub-tasks: span detection and entity class inference. The span detection module, which is class-agnostic, identifies spans regardless of entity class, allowing knowledge transfer across classes. The entity class inference module then uses the detected spans with the natural language descriptions of entity classes to determine their semantic relationships. Ma et al. (2022a) proposed a neural architecture consisting of two BERT encoders, one to encode text tokens and another one to encode each of the labels in natural language format. The model then learns to match the representations of named entities computed by the first encoder with the label representations computed by the second encoder. Other researchers explore generating label knowledge to incorporate into the main NER task. Chen et al. (2022a) introduced Self-describing Networks (SDNet), a Seq2Seq network designed to handle two sequential generation tasks: 1) Mention describing, which generates descriptions for the concepts of mentions, and 2) Entity generation, which adaptively produces entity mentions corresponding to the desired novel types one by one. Using SDNet, NER can be performed directly through the entity generation process by incorporating type descriptions into its prompt. Lai et al. (2022) introduced a two-stage model called PCBERT designed for Chinese few-shot NER. This model comprises two main components: Parent (P-BERT) and Child (C-BERT). During the prompt-tuning stage, P-BERT is trained on the label extension dataset to generate the label extension features for C-BERT. In the subsequent fine-tuning stage, P-BERT remains frozen while C-BERT is fine-tuned. Other researchers integrate entity information into pretraining tasks to aid in downstream NER tasks. Dong et al. (2023) presented a Multi-Task Semantic Decomposition Framework via Joint Task-specific Pre-training (MSDP) which introduces two novel pre-training tasks: Demonstration-based MLM and Class Contrastive Discrimination. These tasks effectively integrate entity boundary information and improve entity representation in PLM. For the downstream main task, they proposed a multi-task joint optimisation framework using a semantic decomposing method, which helps the model combine two distinct types of semantic information for entity classification.

7.5.3 Contrastive learning

Contrastive representation learning framework (Le-Khac et al. 2020) typically consisting of (query, key), similarity distribution, model, encoder, transform head and contrastive loss. This methodology facilitates the creation of distinct entity prototypes for each entity for a few shot NER. Das et al. (2022) proposed CONTaiNER, a contrastive learning technique in NER that enhances the inter-token distance by optimising a generalised objective, distinguishing between token categories through Gaussian-distributed embeddings. Other researchers tried to integrate entity type names into contrastive learning to generate more accurate and consistent prototypes. Huang et al. (2022a) developed COPNER by introducing class-specific words into prompts, which serve as supervision signals for contrastive learning to optimise token representations and as metric references for distance-metric inference on test samples. Li et al. (2023b) introduced TadNER, leveraging entity type names to resolve false span detection and unstable prototypes in two-stage prototypical networks. TadNER employs a type-aware span filtering strategy during span detection, which filters out erroneous spans and a type-aware contrastive learning strategy to create more precise and consistent prototypes for type classification. To extend nested NER to few shot setting, Xu et al. (2023b) proposed a span-based method based on Focusing, brIdging and prompTing (FIT) without using source domain data. The focusing and bridging components effectively identify precise candidate spans. Then the prompting component utilises the distinctive characteristics of nested entities by employing soft prompts and contrastive learning to classify spans. In their work on few-shot nested NER, Ming et al. (2024) introduced a Global-Biaffine span representation to model the global dependency information of each entity span. They also employ a novel positive-enhanced contrastive loss function to maximise the utility of specific positive samples in contrastive learning. Finally, they use nearest neighbour inference to identify and predict unlabelled entities.

7.5.4 Summary

Table 14 shows a summary of the few-shot NER tasks based on author and different techniques. We categorise the techniques used for few-shot NER tasks into three main techniques including meta learning, entity knowledge and contrastive learning. Most of the papers surveyed for the few-shot NER task are based on meta learning, followed by entity knowledge. We also included the dataset used for each paper and the F1 score achieved for 1-shot vs 5-shots settings.

7.6 Multi-modal NER

Multi-modal NER leverages various modalities beyond text to improve accuracy, especially when additional modalities such as images offer valuable cues for model predictions in social media domains. There are different techniques employed in multi-modal NER which includes multi-modal fusion, multi-task learning, multi-modal alignment, prompt-based methods and external knowledge.

Table 14 Summary for few shot NER tasks categorised by dataset, F1 score (1 shot & 5 shots) and different techniques

Paper	Dataset	1 Shot	5 Shots	MetaL	EK	ContL
Yang and Katiyar (2020)	CONLL-2003	62.3	75.2	✓		
Fritzler et al. (2019)	OntoNotes−5.0	-	-	✓		
Ji et al. (2022)	Few-NERD-Intra	25.8	36.4	✓		
Wang et al. (2022b)	CONLL-2003	47.70	61.88	✓		
Fang et al. (2023)	CoNLL-2003	49.06	64.84	✓		
Feng et al. (2024)	CONLL-2003	-	86.7	✓		
Ma et al. (2022e)	Few-NERD-Intra	43.50	56.84	✓		
Wang et al. (2021d)	CONLL-2003	-	76.65 (10-shot)	✓		
Chen et al. (2023a)	CONLL-2003	57.40	63.45	✓		
Wang et al. (2021c)	CONLL-2003	-	71.1		✓	
Ma et al. (2022a)	CONLL-2003	68.4	76.6		✓	
Chen et al. (2022a)	CONLL-2003	-	71.4		✓	
Dong et al. (2023)	Few-NERD-Intra	47.13	64.69	✓		
Lai et al. (2022)	Ontonotes 5.0	-	-	✓		
Das et al. (2022)	Few-NERD-Intra	33.82	47.51			✓
Huang et al. (2022a)	Few-NERD-Intra	59.56	62.37		✓	✓
Li et al. (2023b)	Few-NERD-Intra	55.44	60.87		✓	✓
Xu et al. (2023b)	ACE2005	-	37.74			✓
Ming et al. (2024)	GENIA	28.36	42.25	✓		✓

MetalL meta learning, *EK* entity knowledge, *ContL* contrastive learning

7.6.1 Multi-modal fusion

Multi-modal fusion involves fusing data from multiple modalities such as text and image to improve the predictive capabilities of a model. Depending on where this fusion occurs within the processing pipeline, it can be broadly categorised into early fusion, late fusion, and intermediate fusion. We will be mainly focusing on early fusion and Intermediate fusion. Early fusion, also called data-level fusion, involves merging modality embeddings into a singular feature representation before feeding it into the model. However, this method may fail to capture the complementary information of multiple modalities and could lead to data redundancy. To address this issue, the early fusion approach is often combined with feature extraction methods such as PCA and autoencoder. Moon et al. (2018) introduced a modality attention module at the input of the NER network. This module calculated a weighted combination of different modalities, including word embeddings, character embeddings, and visual features.

Intermediate fusion involves merging modality information after obtaining high-dimensional embeddings for each modality, and then using an intermediate layer for fusion. The effectiveness of this fusion method depends on the design of the model. Several researchers apply techniques such as gating to eliminate noise in multi-modal representations that hurt multimodal NER performance after multi-modal fusion. Zhang et al. (2018) devised an adaptive co-attention network (ACN) layer between the LSTM and CRF layers. Within the ACN, a gated multimodal fusion module was implemented to acquire a fusion vector that incorporates both textual and visual features. A filtration gate was also introduced to evaluate the utility of the fusion feature in inproving the tagging accuracy of individual tokens. Zheng et al. (2024) introduced DPE-MNER, a model that dynamically integrates multi-modal representations through a structured approach. It decomposes fusion into hierarchical layers, prioritises integration based on specific needs, and explicitly models cross-modal relevance to remove irrelevant information. Several researchers empolyed prefix tuning to fuse multi-modalities. Chen et al. (2022d) proposed HVPNeT which leverages visual prefix-guided fusion mechanism to concatenate object-level visual representation as the prefix of each self-attention layer in BERT and a dynamic gate for each layer to aggregate hierarchical multi-scaled visual features as visual prefix. Chen et al. (2022c) proposed a hybrid transformer, MKGformer, for multi-level fusion of visual and text representation via coarse-grained prefix-guided interaction and fine-grained correlation-aware fusion modules. Other researchers proposed graph-based fusion methods. Zhang et al. (2021a) proposed UMGF which constructs a unified multi-modal graph using both the input sentence and the image with multiple graph-based multi-modal fusion layers.

7.6.2 Multi-task learning

Researchers incorporate multi-task learning to alleviate visual semantic bias in multimodal NER. To reduce visual bias, Yu et al. (2020b) jointly trained a purely text-based entity span detection as an auxiliary module, and a Unified Multimodal Transformer to guide the final predictions with the entity span predictions. Lu et al. (2022) proposed FMIT for MNER which transform the fine-grained semantic representation of the vision and text into a unified lattice structure and leverage entity boundary detection as an auxiliary task to alleviate visual bias. Wang et al. (2022e) proposed ITA, which jointly train NER tasks with cross-view alignment then minimises the KL divergence between cross-modal input view and textual input view to reduce visual semantic bias. Chen et al. (2022d) proposed a Hierarchical Visual Prefix fusion NeTwork (HVPNeT) which jointly train multi-modal NER and multi-modal RE tasks to obtain multimodal features with strong generalisation ability.

Multi-task learning is also used to reduce irrelevant semantics in visual features by leveraging text-image relation prediction tasks. Sun et al. (2020) introduced RIVA, a pre-trained multimodal NER model that is trained on text-image relation prediction and next-word prediction tasks. Sun et al. (2021b) proposed RpBERT which uses a multi-task algorithm to train on the MNER datasets using text-image relation prediction and multi-model NER tasks. Other researchers leverage multi-task learning to optimise visual features. Zhou et al. (2022a) introduced the SMVAE model, which uses two VAEs specialised for each modality to capture their respective latent representations. These representations, derived from the VAEs, are used in label prediction through the product-of-experts (PoE) method (Hinton 2002) on the latent representations of both modalities. Jia et al. (2023) proposed MNER-QG that jointly performs MRC-based multi-modal NER and query grounding. To perform the query grounding task, they use manual annotations and weak supervisions that are obtained through training a highly flexible visual grounding model with transfer learning. Chen et al. (2023c) introduced a multi-task multi-modal learning framework that distinguishes between shared and task-specific features. Their approach enhances Multi-modal NER by incorporating cross-modal auxiliary tasks based on Cross-modal Matching and Cross-modal Mutual Information Maximisation to boost MNER performance.

7.6.3 Multi-modal alignment

Alternative studies suggest aligning features from textual and visual modalities for multi-modal NER. Multi-modal alignment involves identifying relationships and correspondences between sub-components of instances across two or more modalities. Multi-modal alignment can be classified into two categories: implicit and explicit. Explicit alignment involves directly aligning sub-components between modalities. For example, this could involve aligning the recipe steps with their corresponding instructional video segments. Implicit alignment serves as an intermediate step, often latent, for another task. For example, image retrieval based on text description may involve an alignment step between words and image regions. In this section, we will mainly discuss implicit alignment.

Several studies propose attention mechanisms to implicitly align different modalities such as image and text. Tian et al. (2021) proposed a Hierarchical Self-adaptation Network (HSN) for Multi-modal NER in social media. Their method involves a Cross-modal Interaction Module to enhance semantic interactions between different modalities via Multi-head Hierarchical Attention (MHA) and a Self-adaptive Multi-modal Integration module to handle missing or mismatched modalities. Other researchers propose to use contrastive learning to implicitly align different modalities in multi-modal NER. Xu et al. (2022) introduced a matching and alignment framework (MAF) for Multi-modal NER. Firstly, a cross-modal matching (CM) module computes the similarity score between the text and the image. Next, a cross-modal alignment (CA) module improves the consistency of representations between the two modalities through contrastive learning. Guo et al. (2023) proposed a Multi-Grained Interaction Contrastive Learning (MGICL) framework. MGICL operates by segmenting data into various granularities: sentence and word token levels for text, and image and object levels for images. Next, multi-grained contrastive learning is performed across different modalities. Lastly, a visual gate control mechanism is used to dynamically select relevant visual information, thereby mitigating the impact of visual noise. To mitigate biases from discrepancies in the quantity and entity types of visual objects, Zhang et al. (2023g) introduced a de-bias contrastive learning model to achieve implicit alignment across modalities by enhancing the learning process within a shared latent semantic space for text and images. This approach employs de-bias contrastive learning, integrating a hard sample mining strategy and using a de-biased contrastive loss function. Bao et al. (2023) proposed MPMRC-MNER, a Multi-modal Prompt-based MRC based framework to implicitly align between text and image, leveraging multi-modal prompt, prompt-aware attention and contrastive learning.

Using MRC queries and query grounding, prior information can be obtained about entity types and image regions. Jia et al. (2023) proposed MNER-QG that jointly performs MRC-based multi-modal NER and query grounding. To perform the query grounding task, they used manual annotations and weak supervisions that are obtained through training a highly flexible visual grounding model with transfer learning. Lu et al. (2022) proposed a Flat Multi-modal Interaction Transformer (FMIT) for MNER that uses noun phrases in sentences and general domain words to obtain visual cues using visual grounding. Researchers also explored a new MNER task called Grounded Multi-modal NER (GMNER) which aims to identify named entities, entity types and their corresponding visual regions. Yu et al. (2023) first introduced a Grounded Multi-modal NER (GMNER) task and a Hierarchical Index generation framework named H-Index, which generates the entity-type-region triples in a hierarchical manner using a sequence-to-sequence model. Li et al. (2024) introduced RiVEG, a framework for GMNER that operates in two distinct stages. In the first stage, the NER and Expansion stage, RiVEG leverages auxiliary refined knowledge from LLMs to enhance MNER performance. This stage also guides LLMs in converting named entities into named entity referring expressions. In the stage of Named Entity Grounding, RiVEG reformulates the entire Entity grounding (EG) task as a union of Visual Entailment (VE) and Visual Grounding (VG).

Other methods were proposed to align multiple modalities for Multi-modal NER. Zheng et al. (2021) presented AGBAN that uses adversarial training to align entity-related features from both visual objects and textual content. Wang et al. (2022e) proposed ITA, which first converts image into visual contexts in textual space and concatenates NER texts with visual contexts as a new cross-modal input view. Cross-view alignment then minimises the KL divergence between the cross-modal input view and the textual input view. Mai et al. (2024) proposed a dynamic graph construction framework (DGCF). They designed a similarity vector-based text-image matching inference strategy to capture both overall and local matching relations between text and images, with the overall matching determining the proportion of visual information retained. Following this, they developed a multi-modal dynamic graph interaction module to construct a dynamic cross-modal graph and a semantic graph. Finally, a CRF layer is used to predict entity labels.

7.6.4 Prompt-based methods

Prompt-based methods are widely used for multi-modal NER tasks. Several researchers incorporated visual prefix into each self-attention layer to guide the fusion process. Chen et al. (2022c) proposed a hybrid transformer, MKGformer that utilises multi-level fusion, which integrates visual and text representation through coarse-grained prefix-guided interaction and fine-grained correlation-aware fusion modules. Chen et al. (2022d) proposed a Hierarchical Visual Prefix fusion NeTwork (HVPNeT). It leverages visual prefix-guided fusion mechanism to concatenate object-level visual representation as the prefix of each self-attention layer in BERT and a dynamic gate for each layer to aggregate hierarchical multi-scaled visual features as visual prefix.

Other prompt-based methods were proposed for multi-modal NER. Cai et al. (2023a) explored few shot multi-modal NER using in-context learning (ICL) consisting of three components. Retrieve example module, which use k-nearest neighbours of text and image to select examples. Demonstration designing module, which includes instruction construct and demonstration construct, and Predict module, which applies an LLM to generate prediction results without training. Zhuang et al. (2023) introduced a prompt network tailored for MNER tasks (P-MNER). To mitigate noise originating from irrelevant image regions, a visual feature extraction model (FRR) using FasterRCNN and ResNet leverages fine-grained visual features to enhance MNER tasks. Additionally, a text correction fusion module (TCFM) is used to counteract visual bias during modal fusion within the model by continuously integrating the original text features with the fusion features to iteratively correct the fusion features.

7.6.5 External knowledge

Some authors leverage external knowledge for Multi-modal NER. Wang et al. (2022d) proposed a Multi-modal Retrieval based framework (MoRe) consisting of a text retrieval module and an image-based retrieval module. The retrieval results are sent to the textual and visual models respectively for predictions. Finally, a Mixture of Experts (MoE) module combines the predictions from the two models. As traditional models often exhibit poor performance with unseen entities, Ok et al. (2024) proposed SCANNER, which comprises the Span Candidate Detection Module and the Entity Recognition Module. The Span Candidate Detection Module first identifies potential entity candidates from the input text. For each extracted candidate entity, SCANNER then uses a range of knowledge sources, including Wikipedia, image captioners, and object knowledge extractors to gather relevant knowledge for NER predictions.

Researchers also leverage ChatGPT as an implicit knowledge base for multi-modal NER. Li et al. (2023a) introduced PGIM, a two-stage framework designed to improve entity prediction efficiency by using ChatGPT as an implicit knowledge base. The framework includes a Multi-modal Similar Example Awareness module, which identifies relevant examples from a set of predefined artificial samples. These examples are integrated into a structured prompt template specific to MNER, guiding ChatGPT to generate refined auxiliary knowledge. The acquired knowledge is then combined with the original text and passed into a downstream model for additional processing.

7.6.6 Summary

Table 15 shows a summary of multi-modal NER categorised by author and different techniques. We categorise the techniques used for multi-modal NER tasks into five main

Table 15 Summary for multi-modal NER tasks categorised by dataset, F1 score and different techniques (* indicate semi-supervised settings)

Paper	Dataset	F1 score	Fusion	MTL	Align	Prompt	EK
Moon et al. (2018)	SnapCaptions	52.4	✓				
Zhang et al. (2018)	Twitter 2015	70.69	✓				
Zheng et al. (2024)	Twitter-2015	77.56	✓				
Chen et al. (2022d)	Twitter-2015	75.32	✓	✓		✓	
Chen et al. (2022c)	Twitter-2017	87.49	✓			✓	
Zhang et al. (2021a)	Twitter 2015	74.85	✓				
Yu et al. (2020b)	Twitter 2015	73.41		✓			
Lu et al. (2022)	Twitter 2015	76.25		✓		✓	
Wang et al. (2022e)	Twitter 2015	76.01		✓	✓		
Sun et al. (2020)	Twitter 2015	71.5		✓			
Sun et al. (2021b)	Twitter 2015	74.4		✓			
Zhou et al. (2022a)	Twitter-2015	61.65*		✓			
Jia et al. (2023)	Twitter 2015	74.94		✓	✓		
Chen et al. (2023c)	Twitter-2015	74.39		✓			
Tian et al. (2021)	Twitter 2015	74.18			✓		
Xu et al. (2022)	Twitter 2015	73.42			✓		
Guo et al. (2023)	Twitter 2015	80.18			✓		
Zhang et al. (2023g)	Twitter 2015	75.28			✓		
Bao et al. (2023)	Twitter 2015	76.26			✓		
Yu et al. (2023)	Twitter-GMNER	79.73			✓		
Li et al. (2024)	Twitter 2015	79.44			✓		
Zheng et al. (2021)	Twitter 2015	73.25			✓		
Mai et al. (2024)	Twitter-2015	75.13			✓		
Cai et al. (2023a)	Twitter 2015	56.99 (16-shots)				✓	
Zhuang et al. (2023)	Twitter-2015	79.43				✓	
Wang et al. (2022d)	Twitter-2015	79.21					✓
Ok et al. (2024)	Twitter 2015	85.73					✓
Li et al. (2023a)	Twitter 2015	79.33					✓

Fusion multi-modal fusion, *MTL* multi-task learning, *Align* multi-modal alignment, *Prompt* prompt-based methods, *EK* external knowledge

techniques including multi-modal fusion, multi-task learning, multi-modal alignment, prompt-based methods and external knowledge. Most of the papers surveyed for multi-modal NER are based on multi-modal alignment, followed by multi-modal fusion and multi-task learning. We also included the dataset used for each paper and the F1 score achieved.

7.7 Visually-rich document NER

Visually-rich document NER (Vrd-NER) aims to extract entities from documents such as forms, receipts, and invoices. It can be considered as a subset of multi-modal NER as it involves different modalities such as text, image and layout. Techniques used in Vrd-NER include convolution-based methods, graph neural network, pretrained encoder models and pretrained encoder-decoder models.

7.7.1 Convolution-based methods

Early approaches are based on convolution-based methods that utilise a fully convolutional encoder-decoder network as the model architecture for Vrd-NER. Katti et al. (2018) introduced Chargrid, a new text representation that maintains the 2D layout of a document by encoding each page as a two-dimensional character grid. They propose a general document understanding pipeline for structured documents using this representation, which employs a fully convolutional encoder-decoder network to predict segmentation masks and bounding boxes.

7.7.2 Graph neural networks

Several approaches integrated Graph neural network for Vrd-NER. Yu et al. (2021) proposed PICK consisting of an encoder, graph module and decoder. The encoder encodes the text using transformer and encodes the image using CNN to obtain the text embedding and image embedding respectively. The text embedding and image embedding is then combined and fed to the graph module together with the bounding boxes to learn the richer graph embedding representation of nodes. The Bi-LSTM CRF decoder then uses the graph embeddings with the encoder representation to jointly perform sequence tagging. Tang et al. (2021) proposed MatchVIE that uses a multi-feature extraction backbone and two branches for relevancy evaluation and entity recognition. The text embedding, position embedding and image embedding are fed to a multi-feature extraction backbone and using multi-head attention to extract the token features. Relevancy evaluation branch is used to understand the relevant relationships between text segments using a graph neural network. Lastly, an entity recognition branch uses the context vector as input to predict the probability of each class. To address the reading order issue, Zhang et al. (2023a) introduced Token Path Prediction (TPP), a simple prediction head to predict entity mentions as token sequences within documents. TPP models the document layout as a complete directed graph of tokens and predicts token paths within the graph as entities.

7.7.3 Pretrained encoder models

Pretrained Encoder-based models typically utilise a pretrained transformer encoder backbone BERT-base or ERNIE base and continually pretrained on two or more of the different modalities such as text, images, layout. Appalaraju et al. (2021) proposed Docformer, a multi-model transformer pretrained in self-supervised manner, combining text, vision and spatial features using a multi-modal self-attention layer. Docformer utilises an encoder-only transformer architecture with a ResNet50 CNN backbone to extract visual features. The text, visual and spatial features are untied and pass through the multi-modal self-attention separately in each transformer layer. Li et al. (2021c) proposed SelfDoc, a task-agnostic pre-training for document image understanding. It uses a document object detector based on Faster RCNN to detect document object proposals and uses adaptive pooling on each RoI head to generate visual features. OCR is used to obtain the text in each proposal and a pretrained sentence BERT converts the text in each proposal to a feature vector. The language and vision features are processed separately using single-modality encoders. The output of the text and visual encoder is fed into a cross-modality encoder to model the agreement of text and vision and discover inner relationships between modalities. Hong et al. (2022) proposed BERT Relying on spatiality (BROS) that uses text and layout to extract information from documents. It relies solely on the text and layout (spatial) information without relying on the visual features. Xu et al. (2020) proposed LayoutLM that uses document layout information in addition to the text information during pretraining. Image features can be added to include text, visual elements during fine-tuning. The vanilla LayoutLM (Xu et al. 2020) model used visual embeddings in the fine-tuning stage and absolute 2D position embedding. Later, Xu et al. (2021c) proposed LayoutLMv2 which integrates visual information in the pre-training stage and uses the transformer to learn the cross-modality interaction across different modalities. Huang et al. (2022b) proposed LayoutLMv3, a multimodal transformer that uses unified text-image masking to learn cross-modal representation. Gu et al. (2021) proposed Unified Pre-training Framework for Document Understanding (UDoc), a unified pre-training framework that extends the transformer to take multimodal embeddings as input for document understanding. UDoc consists of four components feature extraction, feature embedding, multi-layer gated cross-attention encoder and pretraining tasks. To address multilingual document understanding, Wang et al. (2022a) proposed Language-independent Layout Transformer (LiLT) for structured document understanding. The text and layout information are first decoupled and jointly optimised during pre-training and then re-coupled for fine-tuning. This allows LiLT to be pre-trained on structured documents of a single language and then directly fine-tuned on other languages.

Text extracted from documents using OCR typically exhibits noisy improper reading order. Gu et al. (2022) proposed a layout-aware multimodal network XYLayoutLM which can capture and leverage rich layout information from proper reading orders produced using Augmented XY Cut. Peng et al. (2022) introduced ERNIE-Layout, a pre-training method that incorporates text, layout, and image features. Initially, input sequences are reorganised in the serialisation stage. Subsequently, a reading order prediction task is used to teach the model the correct sequence for document comprehension. To further refine the model's awareness of layout, a spatial-aware disentangled attention mechanism is integrated into the multi-modal transformer, along with a replaced regions prediction task during pretraining.

7.7.4 Pretrained encoder-decoder models

Pretrained Encoder-decoder models utilise a encoder-decoder model architecture e.g. T5-base or BART-base backbone. Kim et al. (2022) proposed Donut, an OCR free Visual document understanding model consisting of an encoder and decoder. The encoder consists of a Swin Transformer that breaks up the image into non-overlapping patches and passes through a shifted-based multi-head self-attention module and two-layer multi-layer perceptron for each image patch. The output of the encoder serves as the input to the decoder. The decoder uses a BART model which generates the token sequence. Tang et al. (2023b) proposed UDOP which unifies vision, text and layout and different document tasks. UDOP model architecture consists of a unified vision, text and layout encoder and vision-text-layout decoder. The unified encoder and text-layout decoder use the T5 model to generate text and layout in a sequence-to-sequence manner. The vision decoder uses a masked auto-encoder (MAE) to generate image pixels given text and layout.

7.7.5 Other approaches

Other approaches integrated a text reading module at the beginning of the processing workflow for end-to-end information extraction from visually-rich documents. Wang et al. (2021a) introduced VIES consisting of a shared backbone with three branches for text detection, recognition and information extraction. The text detection branch consists of a Mask-RCNN that localise the text within the document. The recognition branch uses an encoder to extract the input feature sequence and an LSTM decoder with attention to generate the output text sequence. The information extraction branch then uses the visual and semantic features together with the spatial features to fuse these features together using the adaptive feature fusion module (AFFM). The fused features are re-couple to combine global and local information and then fed to Bi-LSTM CRF for sequence labelling.

7.7.6 Summary

Table 16 shows a summary of visually rich document (VrD) NER tasks based on author and different techniques. We categorise the techniques used for VrD NER task into 4 main techniques including convolution-based methods, graph neural networks, pre-trained encoder models and pretrained encoder-decoder models and other approaches. Most of the papers surveyed for the VrD NER task are based on pre-trained encoder models followed by graph neural networks and convolution-based methods. We also included the dataset used for each paper and the F1 score achieved.

Table 16 Summary for VrD-NER tasks categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	Conv	Graph	PEM	PEDM	Others
Katti et al. (2018)	Invoices	61.99 *	✓				
Zhang et al. (2023a)	FUNSD-r	80.40		✓			
Yu et al. (2021)	SROIE	96.1		✓			
Tang et al. (2021)	FUNSD	81.33		✓			
Appalaraju et al. (2021)	CORD	96.99			✓		
Li et al. (2021c)	FUNSD	83.36			✓		
Hong et al. (2022)	FUNSD	84.52			✓		
Xu et al. (2020)	FUNSD	79.27			✓		
Xu et al. (2021c)	FUNSD	84.2			✓		
Huang et al. (2022b)	FUNSD	92.08			✓		
Gu et al. (2021)	FUNSD	87.96			✓		
Wang et al. (2022a)	FUNSD	88.41			✓		
Gu et al. (2022)	XFUND	82.04			✓		
Peng et al. (2022)	FUNSD	93.12			✓		
Tang et al. (2023b)	FUNSD	91.62				✓	
Kim et al. (2022)	CORD	84.1				✓	
Wang et al. (2021a)	SROIE	96.12					✓

Conv convolution-based methods, *Graph* graph neural network, *PEM* pretrained encoder model, *PEDM* pretrained encoder-decoder model

* is based on the average accuracy metric

7.8 Fine-grained NER

Coarse-grained named entity often comprises less than 18 named entity categories. Coarse-grained NER datasets include CoNLL-2003 (Tjong Kim Sang and De Meulder 2003), ACE-2004 (Doddington et al. 2004), ACE-2005 (Doddington et al. 2004)), and OntoNotes−5.0 (Weischedel et al. 2013). For example, CoNLL-2003 dataset includes four coarse-grained NE categories: Person, Location, Organisation and Miscellaneous (Tjong Kim Sang and De Meulder 2003). In contrast, fine-grained NER datasets comprise hundreds of NE categories, which are fine-grained classification of coarse-grained categories. For example, Sekine (2008) further segregated coarse-grained entity category (Organisation) into its fine-grained categories such as Political Party, Military, Sports Organisation, Show Organisation.

Mai et al. (2018) first performed an empirical study between FG-NER models for English and Japanese and showed that an LSTM+CNN+CRF model works well for English FG-NER but does not work well for Japanese due to a large number of character types. To address this problem, they removed the CNN layer in the model and used dictionary and category embeddings. Fine-grained NER are typically modeled using a hierarchical-based approach.

7.8.1 Hierarchical-based approach

Due to the hierarchical relationship between coarse-grained and fine-grained entities, it is intuitive to leverage this relationship for fine-grained NER. Lee et al. (2023) proposed a fine-grained NER model using a Fine-to-Coarse(F2C) mapping matrix to leverage the hierarchical structure between fine-grained and coarse-grained entities explicitly. An inconsistency filtering technique is proposed to remove coarse-grained entities that are inconsistent with fine-grained entity types. Ma et al. (2023b) introduced C2FNER task which aims to train models to quickly adapt from coarse annotations to recognising fine-grained classes with limited samples. During coarse-grained training, a Cluster-based Prototype Margin Loss is utilised to learn discriminative representations grouped by clusters, which aids in fine-grained learning. For fine-grained few-shot learning, a Prototype Retrieval algorithm is employed to fetch representative clusters for each fine class, followed by Mixture Prototype Learning to enhance fine-grained representations.

7.8.2 Summary

Table 17 shows a summary of the fine-grained NER tasks based on author and hierarchical-based. We also included the dataset used for each paper and the F1 score achieved.

7.9 Active learning NER

Active learning (AL) is a technique that maximise performance gains while minimising the number of labelled samples. Its focus lies in selecting the most informative samples from the unlabelled dataset and presenting them to an oracle (such as a human annotator) for labelling. This approach aims to reduce labelling costs while maintaining performance standards. Deep active learning (DeepAL) merges deep learning with active learning, capitalising on the strengths of both domains. In DeepAL, the parameters of the deep learning

Table 17 Summary for fine-grained NER tasks categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	Hierarchical	Other
Lee et al. (2023)	OntoNotes+CoNLL'03+Few-NERD	57.18	✓	
Ma et al. (2023b)	Few-NERD	41.62 (5-shots)	✓	
Mai et al. (2018)	FG-NER (English)	83.14		✓

model are initialised or pre-trained using labelled data, while unlabelled samples are utilised to extract features through the deep learning model. Subsequently, samples are chosen based on a designated query strategy, and their labels are obtained from the oracle to form a new labelled training set. The deep learning model is then updated and trained simultaneously using this augmented dataset. This iterative process continues until the labelling budget is exhausted or predefined termination conditions are met. Broadly, the DeepAL framework can be divided into two components: the AL query strategy applied to the unlabelled dataset and the training method for the deep learning model.

7.9.1 Multi-task learning

Automated medical NER identifies specific entities like diseases or medications in medical texts, while normalisation (NEN) matches these entities to standard identifiers. It has great value in the medical domain, e.g., medical report generation (Mei et al. 2024) and diagnostics (Wu et al. 2023b). Zhao et al. (2019) suggested jointly modelling NER and NEN using multi-task learning to leverage task relationships. Active learning is commonly used and trained in a semi-supervised manner to reduce labelling costs. However, the existing multi-task active learning models do not take the influence of task-specific features and the diversity constraint into account. To address this, Zhou et al. (2021a) proposed MTAAL, a multi-task adversarial active learning model for medical NER and normalisation. It comprises four components: shared encoder, task private decoders, task discriminator, and diversity discriminator. Adversarial learning keeps the effectiveness of multi-task learning module and active learning module. The task discriminator eliminates the influence of irregular task-specific features. And the diversity discriminator exploits the heterogeneity between samples to meet the diversity constraint.

7.9.2 Other approaches

Besides multi-task learning, Liu et al. (2022b) introduced an uncertainty-based active learning strategy called the lowest token probability (LTP) based on BERT-CRF which selects the tokens whose probability under the most likely tag sequence y^* is lowest as shown in Eq. 24. This method involves combining the input and output of a CRF to identify informative instances. The advantage of LTP is it does not show bias towards longer sequences and does not require model adjustments. Moscato et al. (2024) introduced Active Learning-based Data Augmentation for NER (ALDANER) that applied active learning to data augmentation to prioritise selecting informative samples from an augmented dataset while reducing the effects of noisy annotations generated during data augmentation.

$$\phi^{LTP}(x) = 1 - min_{y_i^* \epsilon y^*} P(y_i^* | x_i; A) \tag{24}$$

7.9.3 Summary

Table 18 shows a summary of active learning NER categorised by author and different techniques. We categorise the techniques used for active learning NER task into multi-task learning techniques and other approaches. Most of the papers surveyed in active learning NER task are based on multi-task learning. We also included the dataset used for each paper and the F1 score achieved.

7.10 Continual learning NER

Continual learning refers to incrementally learning new information from a non-stationary stream of data. There are three main types of continual learning in NER which includes task-incremental, domain-incremental and class-incremental learning. Task incremental learning (TIL) is a category of continual learning that seeks to train a single network for multiple tasks (one after another), where training data for each task are only available during the training of that task. Domain incremental learning (DIL) aims to adapt to a sequence of domains with access to only a small subset of data (i.e. memory) from previous domains. Qin and Joty (2021) proposed LFPT5, a unified framework for Lifelong Few-shot Language Learning (LFLL) based on prompt tuning of T5 for sequence labelling, text classification and text generation. LFPT5 generates pseudo-labelled samples of previously learned domains and later gets trained on those samples to alleviate forgetting of previous knowledge as it learns the new domain. In addition, a KL divergence loss is minimised to achieve label consistency between the previous and current model. Class incremental learning (CIL) focuses on learning a model that continuously learns new classes in a sequential manner without forgetting old ones. It is relevant in real-world settings, such as voice-enabled assistants, where there is a frequent introduction of novel named entity types. Monaikul et al. (2021) first proposed a class-incremental NER where a teacher NER model transfers its knowledge to a student model through knowledge distillation, preserving knowledge on old entities while learning new entity types. They propose two methods for addressing class-incremental NER: ExtendNER and AddNER. ExtendNER involves expanding the classifier's dimension when new classes are introduced, whereas AddNER adjusts to new classes by incorporating a separate classifier for each new category. Named entities are often identified together with

Table 18 Summary for active learning NER task categorised by dataset, F1 score, and different techniques

Paper	Dataset	F1 score	MTL	Others
Liu et al. (2022b)	OntoNotes−5.0 Chinese	66.6		✓
Zhou et al. (2021a)	BC5CDR	86.0	✓	
Moscato et al. (2024)	CoNLL-2003	85.8		✓

MTL multi-task learning, *KD* knowledge distillation, *Span* span-based, *ContL* contrastive learning

the "O" (others) class which outnumber the actual entity classes. Ma et al. (2023a) discovered that severe confusion between “O" and entities affects the model’s ability to learn new classes for Class-incremental NER. To address this, an entity-aware contrastive learning method is proposed that adaptively detects entity clusters in "O" and two effective distance-based relabelling strategies for better learning the old classes. Among the three types of continual learning approaches for NER, CIL is the most widely explored by NER researchers. Class-incremental NER typically employ techniques such as knowledge distillation and span-based methods which will be elaborated subsequently.

7.10.1 Knowledge distillation

Class-incremental NER suffers from the backward incompatibility problem where previously learned entity type’s mentions may appear in the samples trained in the current task but without the relevant annotations. These false negative labels will inevitably force models to forget old knowledge to fit the new conflicting one, also known as catastrophic forgetting. Knowledge distillation is an effective technique used to mitigate catastrophic forgetting during continual learning. Xia et al. (2022) presented a new two-stage continual NER framework, Learn-and-Review (L&R). In the learning stage, prior knowledge from a teacher model is distilled to a student model using the current dataset. During the review stage, data augmentation is performed and augmented dataset is then used to distil both new knowledge from the updated student model and prior knowledge from the teacher model, resulting in an improved student model. Zhang et al. (2023c) proposed Decomposing Logits Distillation by decomposing the predicted logit into two terms that measure the likelihood that an input token belonging to a specific entity type or not and explicitly constrain each term. In contrast, traditional Logits distillation only preserves the sum of these two terms without considering the change in each component, which is more inferior in retaining old knowledge and mitigating catastrophic forgetting. Wang et al. (2022c) proposed few shot class-incremental learning for NER. They generate synthetic data of the old classes using the trained NER model and further distils the NER model from previous steps with both synthetic data, and real data from the current training set. Zhang et al. (2023b) proposed a pooled feature distillation loss that carefully trade-off between retaining knowledge of old entity types and acquiring new ones and a confidence-based pseudo-labelling for the non-entity type, i.e., predicting entity types using the old model to handle the semantic shift of the non-entity type. An adaptive re-weighting type-balanced learning strategy is introduced to handle the issue of biased type distribution. Zheng et al. (2022) proposed a unified causal framework by retrieving and distilling the causality from both new entity types and Other-class followed by curriculum learning to reduce the effect of label noise. A self-adaptive weight is introduced for balancing the causal effects between new entity types and Other-Class.

7.10.2 Span-based

An often neglected problem in class-incremental NER is the forward incompatibility present in prior sequence labelling modelling. This problem occurs when the non-entity mentions learning currently may belong to a certain entity type to be learned in future tasks. To solve this issue, researchers replaced sequence labelling with a span-based model for class-incremental NER. Span-based models are found to be forward compatible as it converts the NER into a binary classification problem, which reduces interference in future tasks. Zhang and Chen (2023) introduced SpanKL, a Span-based model that uses knowledge distillation to retain acquired knowledge by employing independent modelling at both the span and entity levels. To better address the token-noise problem in continual NER, Chen and He (2023) proposed SKD-NER, another span-based model that uses knowledge distillation (KD) to retain memory and employ reinforcement learning during the KD process to optimise soft labelling and distillation losses generated by the teacher model.

7.10.3 Summary

Table 19 shows a summary of the continual learning NER task based on author and different techniques. We categorise the techniques used for continual learning NER task into knowledge distillation, span-based and other approaches. Most of the papers surveyed for continual learning NER task are based on knowledge distillation followed by span-based approach. We also included the dataset used for each paper and the F1 score achieved.

Table 19 Summary for continual learning NER task categorised by dataset, F1 score, and different techniques (* indicate Macro F1 scores)

Paper	Dataset	F1 score	KD	Span	Others
Monaikul et al. (2021)	CoNLL-2003	87.0*	✓		
Xia et al. (2022)	CoNLL-2003	85.74*	✓		
Zhang et al. (2023c)	CoNLL-2003	79.54*	✓		
Wang et al. (2022c)	CoNLL-2003	65.12 (10-shot)	✓		
Zhang et al. (2023b)	OntoNotes−5.0	66.27*	✓		
Zheng et al. (2022)	OntoNotes−5.0	60.52*	✓		
Zhang and Chen (2023)	OntoNotes−5.0	89.78*	✓	✓	
Chen and He (2023)	OntoNotes−5.0	88.17	✓	✓	
Ma et al. (2023a)	Few-NERD	48.11*			✓
Qin and Joty (2021)	CoNLL-2003+OntoNotes−5.0	47.59			✓

MTL multi-task learning, *KD* knowledge distillation, *Span* span-based, *ContL* contrastive learning

7.11 Open vocabulary NER

Traditional supervised learning can only recognise a fixed number of entity types observed based on their supervised labels. However, novel entity types continually emerged in real-world scenarios. Hence, it is a non-trivial problem to build NER models that are capable of classifying novel entity types on the fly at inference time. Open-vocabulary NER requires that the trained NER model to be capable of recognising entities in any novel type by their textual names or descriptions. For example, in the sentence “Barack Obama was born in Honolulu, Hawaii", the corresponding type description for “Barack Obama" could be “A politician" is a person active in party politics, or a person holding or seeking an elected". Recently, there has been an increased interest by researchers in exploring Open Vocabulary NER. Zhou et al. (2023b) explored targeted distillation with mission-focused instruction tuning by using ChatGPT for distillation into much smaller UniversalNER models for open NER. Zaratiana et al. (2024) proposed GLiNER which uses Bidirectional Language Models for open NER. Their model consists of a pre-trained textual encoder, a span representation module and an entity representation module. Finally, a matching score between entity representations and span representations is computed. Jin et al. (2023) proposed open-vocabulary NER (OVNER) as a semantic matching task and proposed a novel and scalable two-stage method called Context-Type SemAntiC Alignment and FusiOn (CACAO). In the pre-training stage, Dual Encoder is pre-train on context-type pairs using contrastive learning. In the fine-tuning stage, Cross-Encoder is fine-tuned on base types with human supervision.

7.11.1 Summary

Table 20 shows a summary of the open vocabulary NER task based on author and different techniques. Techniques used in open-vocabulary NER includes knowledge distillation, span-based and contrastive learning. We also included the dataset used for each paper and the F1 score achieved.

8 Experiments and results

Twitter data related to the 2024 Queensland election was scrapped, resulting in 1,321 tweets annotated with various entity types, including person, geographical location, political event, organisation, political party, government sector, and government. The dataset was divided into training, validation, and test sets with splits of 70%, 15%, and 15%, respectively. NER was performed on these annotated entities. Two models were used for fine-tuning the NER task: a vanilla BERT (Devlin et al. 2019) model and the GLiNER (Zaratiana et al. 2024)

Table 20 Summary for open vocabulary NER task categorised by dataset, F1 score, and different techniques (* indicate Macro F1 scores)

Paper	Dataset	F1 score	KD	Span	ContL
Zhou et al. (2023b)	Pile-NER	55.6	✓		
Zaratiana et al. (2024)	Pile-NER	60.9		✓	
Jin et al. (2023)	OntoNotes−5.0	45.1*			✓

MTL multi-task learning, *KD* knowledge distillation, *Span* span-based, *ContL* contrastive learning

model. The vanilla BERT (Devlin et al. 2019) model is a pre-trained language model based on the bidirectional encoder from transformers. In contrast, GLiNER (Zaratiana et al. 2024) employs a BERT-like bidirectional transformer encoder capable of identifying any entity type by calculating a matching score between entity representations and span representations.

For BERT (Devlin et al. 2019) training, a total of 10 epochs was conducted with a learning rate of 5×10^{-5}, a training batch size of 8, and the Adam (Kingma and Ba 2015) optimiser. In contrast, GLiNER (Zaratiana et al. 2024) training involved 4 epochs with a learning rate of 5×10^{-6}, a training batch size of 8, and AdamW (Loshchilov and Hutter 2017) optimiser. The results are compared in Table 21, focusing on the overall micro precision, micro recall, and micro F1 score, as well as the micro precision, micro recall, and micro F1 scores for individual entities. The vanilla BERT (Devlin et al. 2019) model achieved an average micro precision of 81.0, micro recall of 85.0, and micro F1 of 83.0. GLiNER (Zaratiana et al. 2024) outperformed BERT (Devlin et al. 2019), achieving an average micro precision of 82.0, micro recall of 85.9, and micro F1 of 83.9, attributed to the incorporation of entity types into the prompt.

To determine the optimal parameters in BERT, a trial-and-error approach on a validation set is used. Figure 8 shows the F-measure plotted against different parameter settings, namely (a) number of training samples, (b) batch size, (c) learning rate and (d) epochs. In Fig. 8a as expected, we find that the validation F-measure increases with the percentage of training samples used. From the available samples following a heuristic approach, 70% are used for training, 15% for validation and 15% are used for testing. In Fig. 8b, the effect of increasing the batch size from 1 to 10 is considered. Since a pre-trained model is fine-tuned, a batch size of 1 does not significantly differ in F-measure from a batch size of 10. There is a slight improvement when the batch size was 5. Hence, a batch size of 8 for training is optimal for this dataset.

The effect of the learning rate on the F-measure is considered in Fig. 8c Here a slight improvement when the learning rate is reduced from 1e5 to 5e4 is observed. However, a higher learning rate of 0.001 failed to train the model. This may be because there are not enough training samples for all types of entities, resulting in over-fitting. Following previous authors and this graph, the learning rate is empirically set to 5e6, and it is increased

Table 21 NER experiment results for BERT and GLiNER for various entity types and their average scores

Model	Entity	Precision	Recall	F1
BERT	Person	74.0	76.0	75.0
	Geographical location	86.0	91.0	89.0
	Political event	88.0	93.0	91.0
	Organisation	65.0	67.0	66.0
	Political party	90.0	93.0	91.0
	Government sector	31.0	45.0	37.0
	Government	53.0	53.0	53.0
	Average (Micro)	81.0	85.0	83.0
GLiNER	Person	80.9	83.2	82.0
	Geographical location	79.3	91.3	84.9
	Political event	89.2	94.4	91.7
	Organisation	61.9	41.9	50.0
	Political party	82.0	94.8	87.9
	Government sector	23.8	50.0	32.3
	Government	72.7	27.6	40.0
	Average (Micro)	82.0	85.9	83.9

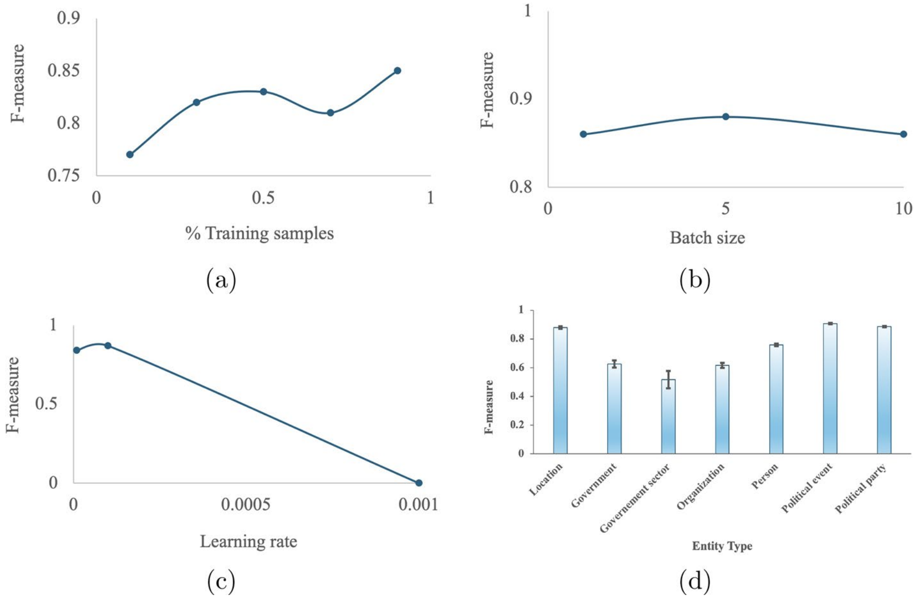

Fig. 8 Effect of parameters in BERT on F-measure **a** percentage of training samples, **b** batch size, **c** learning rate, **d** variance across number of epochs

slightly with each epoch. Lastly, the optimal number of training epochs is determined. Figure 8d shows that some entity types showed a large variance in F-measure as the number of epochs increases from 5 to 15. However, the average F-measure over all entities did not change. To have good accuracy on all entity types, the number of epochs is set to 10.

9 Challenges for NER

9.1 Data labelling

Labelling data stands as a crucial stage in supervised NER, involving annotators assigning entity types to individual words in a given text, like person, organisation, or location. However, this process is often time-consuming, labour-intensive, and expensive.The challenges in NER data labelling encompass several aspects. An issue arises from unclear entity boundaries, causing discrepancies among annotators in interpreting entities, such as “Mr. Steve Jobs" and “Steve Jobs" within the same sentence. Additionally, entities might assume different types based on contextual variations, leading to ambiguity. For instance, “mouse" in the context of technology relates to “electronics" while the same word in an animal context pertains to the “rodent" category. Moreover, accurately distinguishing named entity types may require domain-specific expertise, especially in specialised fields like medicine, where specific knowledge is required.

9.2 Fine grained NER

Traditional NER systems typically categorise entities into a small set of coarse types, often fewer than ten defined categories. For instance, in the CoNLL-2003 NER task, there exist four main categories: Person, Location, Organisation, and Miscellaneous (Tjong Kim Sang and De Meulder 2003). In contrast, Fine-grained NER pursues the identification and classification of a more extensive range of entity categories, potentially numbering in the hundreds, with more specific subcategories within the standard types. An example of coarse-grained types includes Person, Organisation, and Location, while fine-grained types further segregate Person entity into Actor, Athlete, and Politician entities. Challenges associated with fine-grained NER include a scarcity of annotated data for fine-grained entity types, resulting in difficulties when training accurate models. Additionally, distinguishing between fine-grained entity types presents challenges, even for human annotators.

9.3 Low resource NER

Low resource NER involves using available data and models from a language with abundant resources (e.g. English) to address NER tasks in a typically more resource-scarce language. NER for low-resource languages encounters significant hurdles, primarily due to the scarcity of annotated data crucial for training and evaluating NER models. The process of creating annotated data is time-consuming, expensive, and requires linguistic expertise. In addition, diverse languages, domains, genres, and tasks often require varying annotation schemes, increasing the complexity and diversity of NER data. Another challenge lies in the transferability and generalisability of NER models for low-resource languages, affecting their usability and scalability. Model transferability involves a model's ability to perform well in languages, domains, or tasks different from its training data. Model generalisation pertains to a model's performance on unseen or new data. However, transferring and generalising NER models for low-resource languages is complex and may involve hurdles like cross-lingual learning, domain adaptation, and zero-shot learning.

9.4 Class imbalance

NER encounters a class imbalance when there is an unequal distribution of entities in a dataset which is more pronounced in class-incremental learning scenarios. Specifically, the "others" class, which does not align with any specific entity category, vastly outnumbers actual entity classes. This class imbalance can pose training challenges in backpropagation, where the "others" class gradient dominates during the model training process. Consequently, various methods are used to address this problem. An approach involves data sampling to consider class imbalance. Oversampling involves selecting more sentences containing minority classes, while undersampling involves fewer samples from majority classes. Weighted random sampling automatically selects sentences from both minority and majority classes based on sample weights. Another method modifies the loss function to address the class imbalance. Techniques such as weighted cross entropy loss, focal loss (Lin et al. 2017) and dice loss (Li et al. 2020e) modify the loss function to address class imbalance concerns in traditional cross entropy loss. To address the issue of class imbalance in low-resource settings, Nguyen et al. (2023a) suggested a method where the conventional multi-class NER tagging problem is reframed into a dual-task approach: predicting entity tokens and beginning entity tokens. They optimised their NER model by maximising the AUC score.

9.5 Explainable NER

Explainable AI aims to ensure that AI systems are transparent, understandable, and trustworthy in their decision-making processes. Despite the remarkable results achieved by LLM-based deep learning models, these models are still black boxes that are neither interpretable nor explainable. There are several explainable AI methods developed that were used for NLP tasks such as LIME (Ribeiro et al. 2016), SHAP (Lundberg and Lee 2017), Layer-wise Relevance Propagation (LRP) (Bach et al. 2015) and Integrated Gradients (Sundararajan et al. 2017). However, NER remains an understudied task in terms of explainability. Being one of the first who studied explainable NER, Zugarini and Rigutini (2023) propose SAGE, Semantic-Aware Global Explanations for NER which is a post-hoc method to produce highly interpretable global rules extracted using data mining to explain NLP classifiers. They also compare SAGE against other Explainable AI methods such as LIME and Decision Trees. Zhang et al. (2023h) introduced E-NER, a trustworthy framework, which enhances Evidential Deep Learning by incorporating two uncertainty-guided loss terms and implementing uncertainty-guided training techniques.

10 Conclusion

This survey provides a comprehensive overview of NER, covering various aspects of the field. We first introduce NER, background and theoretical research for NER. Next, we present a more comprehensive taxonomy for NER compared to previous surveys. Different learning methods for NER are surveyed including rule-based, supervised, semi-supervised, weakly-supervised and unsupervised NER. Moving on, different modelling paradigms for NER are covered. We also present the NER datasets used for different NER tasks and NER evaluation metrics. Furthermore, the survey delves into detail for the different NER tasks. Finally, the survey also presents the common challenges faced in NER.

One of the biggest challenges in NER is to disambiguate an entity based on context, for example “Washington" can be the name of a person or a place. Existing pre-trained word vectors are based on co-occurrence data from social media articles. However, to determine the context of named entities in a sentence we could further explore the cosine angle between the two entities. For example, in the context of elections when the party name “Labor" is followed by the social issue “Women" in a tweet then we can disambiguate it from other words such as ‘act of doing physical work’.

Another challenge during training is the unequal number of entities in the datasets. For example, in the election data we see more entities from the ‘person’ class compared to ‘social issue’. Another problem may be that there are more occurrences of a particular ‘politician’ compared to others. To overcome this problem, we feel that instead of a unified sequence model, we need to model each named entity type as a separate state or mode. Lastly, word vectors for several named entities may not be available in the pre-trained word vectors. This problem is particularly amplified in new languages and domains. For each sentence, we can achieve greater accuracy by considering the overall sentiment of the tweet and the magnitude of known word vectors. Our future work will focus on developing a framework for NER based on these conclusions.

11 Contributions

Wei Liang wrote the original draft of the manuscript and conducted experiments. Iti Chaturvedi created the election dataset and helped draft the response to the reviewer's comments. Amber Hogarth studied the parameter tuning on election tweets and also proofreading of the article. Mao Rui helped with revising and editing the manuscript. Erik Cambria supervised the work and edited the final manuscript.

Author contributions Wei Liang wrote the original draft of the manuscript and conducted experiments. Iti Chaturvedi created the election dataset and helped draft the response to the reviewer's comments. Amber Hogarth studied the parameter tuning on election tweets and also proofreading of the article. Mao Rui helped with revising and editing the manuscript. Erik Cambria supervised the work and edited the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. Open Access funding enabled and organized by CAUL and its Member Institutions. This work is partially supported by College of Science and Engineering, James Cook University, Australia. Open access publishing facilitated by James Cook University, as part of the Springer - James Cook University agreement via the Council of Australian University Librarians. This work is also partially supported by the College of Computing and Data Science, Nanyang Technological University, Singapore.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Aguilar G, Monroy APL, González FA, et al (2018) Modeling noisiness to recognize named entities using multitask neural networks on social media. In: HLT-NAACL, pp 1401–1412

Akkasi A, Varoğlu E (2017) Improving biochemical named entity recognition using PSO classifier selection and Bayesian combination methods. IEEE/ACM Trans Comput Biol Bioinf 14(6):1327–1338

Alam F, Islam MA (2020) A proposed model for Bengali named entity recognition using maximum entropy Markov model incorporated with rich linguistic feature set. In: ICCA, pp 1–6

Aly R, Vlachos A, McDonald R (2021) Leveraging type descriptions for zero-shot named entity recognition and classification. In: ACL-IJCNLP, pp 1516–1528

Andrychowicz M, Denil M, Colmenarejo SG, et al (2016) Learning to learn by gradient descent by gradient descent. In: NIPS, pp 3988–3996

Appalaraju S, Jasani B, Kota BU, et al (2021) DocFormer: end-to-end transformer for document understanding. In: ICCV, pp 973–983

Arora J, Park Y (2023) Split-NER: Named entity recognition via two question-answering-based classifications. In: ACL, pp 416–426

Bach S, Binder A, Montavon G et al (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):e0130140

Bai Y, Wang Y, Xia B, et al (2020) Adversarial named entity recognition with POS label embedding. In: IEEE IJCNN, pp 1–8

Bao X, Tian M, Zha Z, et al (2023) MPMRC-MNER: A unified MRC framework for multimodal named entity recognition based multimodal prompt. In: CIKM, pp 47–56

Bari MS, Joty S, Jwalapuram P (2020) Zero-resource cross-lingual named entity recognition. In: AAAI, pp 7415–7423

Bengio Y, Louradour J, Collobert R, et al (2009) Curriculum learning. In: ICML, pp 41–48

Bojanowski P, Grave E, Joulin A et al (2017) Enriching word vectors with subword information. TACL 5:135–146

Bowman S, Angeli G, Potts C, et al (2015) A large annotated corpus for learning natural language inference. In: EMNLP, pp 632–642

Buciluǎ C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: KDD, pp 535–541

Cai J, Huang S, Jiang Y, et al (2023b) Improving low-resource named entity recognition with graph propagated data augmentation. In: ACL, pp 110–118

Cai C, Wang Q, Liang B, et al (2023a) In-context learning for few-shot multimodal named entity recognition. In: EMNLP, pp 2969–2979

Cambria E, Mao R, Chen M et al (2023) Seven pillars for the future of artificial intelligence. IEEE Intell Syst 38(6):62–69

Chaturvedi I, Cambria E, Welsch RE et al (2018) Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf Fus 44:65–77

Chen Z, Guo C (2022) A pattern-first pipeline approach for entity and relation extraction. Neurocomputing 494:182–191

Chen Z, Zhang Y, Mi S (2023) Assisting multimodal named entity recognition by cross-modal auxiliary tasks. Pattern Recogn Lett 175:52–58

Chen S, Aguilar G, Neves L, et al (2021b) Data augmentation for cross-domain named entity recognition. In: EMNLP, pp 5346–5356

Chen P, Ding H, Araki J, et al (2021a) Explicitly capturing relations between entity mentions via graph neural networks for domain-specific named entity recognition. In: ACL-IJCNLP, pp 735–742

Chen X, Hassan A, Awadalla HH, et al (2019) Multi-source cross-lingual model transfer: Learning what to share. In: ACL, pp 3098–3112

Chen Y, He L (2023) SKD-NER: Continual named entity recognition via span-based knowledge distillation with reinforcement learning. In: EMNLP, pp 6689–6700

Chen W, Jiang H, Wu Q, et al (2021c) AdvPicker: Effectively leveraging unlabeled data via adversarial discriminator for cross-lingual NER. In: ACL-IJCNLP, pp 743–753

Chen X, Li L, Deng S, et al (2022b) LightNER: A lightweight tuning paradigm for low-resource NER via pluggable prompting. In: COLING, pp 2374–2387

Chen X, Li L, Qiao S, et al (2023b) One model for all domains: collaborative domain-prefix tuning for cross-domain NER. In: IJCAI, pp 5030–5038

Chen J, Liu Q, Lin H, et al (2022a) Few-shot named entity recognition with self-describing networks. In: ACL, pp 5711–5722

Chen J, Lu Y, Lin H, et al (2023a) Learning in-context learning for named entity recognition. In: ACL, pp 13661–13675

Chen J, Wang Z, Tian R, et al (2020a) Local additivity based data augmentation for semi-supervised NER. In: EMNLP, pp 1241–1251

Chen Y, Wu C, Qi T, et al (2020b) Named entity recognition in multi-level contexts. In: AACL-IJCNLP, pp 181–190

Chen X, Zhang N, Li L, et al (2022c) Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In: SIGIR, pp 904–915

Chen X, Zhang N, Li L, et al (2022d) Good visual guidance make a better extractor: hierarchical visual prefix for multimodal entity and relation extraction. In: NAACL, pp 1607–1618

Clark K, Luong MT, Manning CD, et al (2018) Semi-supervised sequence modeling with cross-view training. In: EMNLP, pp 1914–1925

Cui L, Wu Y, Liu J, et al (2021) Template-based named entity recognition using BART. In: ACL-IJCNLP, pp 1835–1845

Dai X, Adel H (2020) An analysis of simple data augmentation for named entity recognition. In: COLING, pp 3861–3867

Dai X, Karimi S, Hachey B, et al (2020) An effective transition-based model for discontinuous NER. In: ACL, pp 5860–5870

Das SSS, Katiyar A, Passonneau RJ, et al (2022) CONTaiNER: Few-shot named entity recognition via contrastive learning. In: ACL, pp 6338–6353

Derczynski L, Bontcheva K, Roberts I (2016) Broad twitter corpus: a diverse named entity recognition resource. In: COLING, pp 1169–1179

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional Transformers for language understanding. In: HLT-NAACL, pp 4171–4186

Ding B, Liu L, Bing L, et al (2020) DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In: EMNLP, pp 6045–6057

Ding N, Xu G, Chen Y, et al (2021) Few-NERD: A few-shot named entity recognition dataset. In: ACL-IJCNLP, pp 3198–3213

Dixit K, Al-Onaizan Y (2019) Span-level model for relation extraction. In: ACL, pp 5308–5314

Doddington GR, Mitchell A, Przybocki MA, et al (2004) The automatic content extraction (ace) program-tasks, data, and evaluation. In: LREC, pp 837–840

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

Dong G, Wang Z, Zhao J, et al (2023) A multi-task semantic decomposition framework with task-specific pre-training for few-shot NER. In: CIKM, pp 430–440

Dou C, Sun X, Wang Y, et al (2023) Domain-adapted dependency parsing for cross-domain named entity recognition. In: AAAI, pp 12737–12744

Ehrmann M, Hamdi A, Pontes EL et al (2023) Named entity recognition and classification in historical documents: a survey. ACM Comput Surv 56(2):1–47

Eronen J, Ptaszynski M, Masui F (2023) Zero-shot cross-lingual transfer language selection using linguistic similarity. Inf Process Manag 60(3):103250

Fang J, Wang X, Meng Z, et al (2023) MANNER: A variational memory-augmented model for cross domain few-shot named entity recognition. In: ACL, pp 4261–4276

Fei H, Ji D, Li B, et al (2021) Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In: AAAI, pp 12785–12793

Feng J, Xu G, Wang Q et al (2024) Note the hierarchy: taxonomy-guided prototype for few-shot named entity recognition. Inf Process Manag 61(1):103557

Fetahu B, Fang A, Rokhlenko O, et al (2022) Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In: HLT-NAACL, pp 2777–2790

Finkel JR, Manning CD (2009) Nested named entity recognition. In: EMNLP, pp 141–150

Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: ICML, pp 1126–1135

Fisher J, Vlachos A (2019) Merge and label: a novel neural network architecture for nested ner. In: ACL, pp 5840–5850

Florian R, Ittycheriah A, Jing H, et al (2003) Named entity recognition through classifier combination. In: HLT-NAACL, pp 168–171

Fritzler A, Logacheva V, Kretov M (2019) Few-shot classification in named entity recognition task. In: SAC, pp 993–1000

Fu Y, Tan C, Chen M, et al (2021) Nested named entity recognition with partially-observed treecrfs. In: AAAI, pp 12839–12847

Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey. Artif Intell Rev 47(1):1–66

Ge L, Hu C, Ma G, et al (2023) ProKD: An unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition. In: AAAI, pp 12818–12826

Ge L, Hu C, Ma G, et al (2024) Discrepancy and uncertainty aware denoising knowledge distillation for zero-shot cross-lingual named entity recognition. In: AAAI, pp 18056–18064

Geng R, Chen Y, Huang R et al (2023) Planarized sentence representation for nested named entity recognition. Inf Process Manag 60(4):103352

Gligic L, Kormilitzin A, Goldberg P et al (2020) Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. Neural Netw 121:132–139

Gu J, Kuen J, Morariu VI, et al (2021) UniDoc: Unified pretraining framework for document understanding. In: NIPS, pp 39–50

Gu Z, Meng C, Wang K, et al (2022) XYLayoutLM: Towards layout-aware multimodal networks for visually-rich document understanding. In: CVPR, pp 4573–4582

Guo A, Zhao X, Tan Z, et al (2023) MGICL: multi-grained interaction contrastive learning for multimodal named entity recognition. In: CIKM, pp 639–648

Gupta P, Schütze H, Andrassy B (2016) Table filling multi-task recurrent neural network for joint entity and relation extraction. In: COLING, pp 2537–2547

He K, Mao R, Huang Y, et al (2023) Template-free prompting for few-shot named entity recognition via semantic-enhanced contrastive learning. In: IEEE transactions on neural networks and learning systems, pp 1–13

He H, Sun X (2017) A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media. In: AAAI, pp 3216–3222

He Y, Tang B (2022) SetGNER: General named entity recognition as entity set generation. In: EMNLP, pp 3074–3085

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. In: NIPS workshop

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hong T, Kim D, Ji M, et al (2022) BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In: AAAI, pp 10767–10775

Huang H, Lei M, Feng C (2021) Hypergraph network model for nested entity mention recognition. Neurocomputing 423:200–206

Huang Z, Chen K, He J, et al (2019b) ICDAR2019 competition on scanned receipt ocr and information extraction. In: IEEE ICDAR, pp 1516–1520

Huang Y, He K, Wang Y, et al (2022a) COPNER: Contrastive learning with prompt guiding for few-shot named entity recognition. In: COLING, pp 2515–2527

Huang L, Ji H, May J (2019a) Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In: HLT-NAACL, pp 3823–3833

Huang J, Li C, Subudhi K, et al (2021b) Few-shot named entity recognition: an empirical baseline study. In: EMNLP, pp 10408–10423

Huang Y, Liu W, Zhang X, et al (2023) PRAM: An end-to-end prototype-based representation alignment model for zero-resource cross-lingual named entity recognition. In: ACL, pp 3220–3233

Huang Y, Lv T, Cui L, et al (2022b) LayoutLMv3: Pre-training for document ai with unified text and image masking. In: MM, pp 4083–4091

Hu A, Dou Z, Wen Jr (2019) Document-level named entity recognition by incorporating global and neighbor features. In: CCIR, p 79–91

Hu J, Guo D, Liu Y, et al (2023) A simple yet effective subsequence-enhanced approach for cross-domain NER. In: AAAI, pp 12890–12898

Hu X, Hong Z, Jiang Y, et al (2024) Three heads are better than one: improving cross-domain NER with progressive decomposed network. In: AAAI, pp 18261–18269

Hu H, Liu T, Feng H, et al (2021) A multi-task framework for named entity recognition. In: IEEE CISAI, pp 51–56

Hu J, Zhao H, Guo D, et al (2022) A label-aware autoregressive framework for cross-domain NER. In: NAACL, pp 2222–2232

Iovine A, Fang A, Fetahu B, et al (2022) CycleNER: An unsupervised training approach for named entity recognition. In: WWW, pp 2916–2924

Jain A, Paranjape B, Lipton ZC (2019) Entity projection via machine translation for cross-lingual NER. In: EMNLP-IJCNLP, pp 1083–1092

Jaume G, Ekenel HK, Thiran JP (2019) FUNSD: A dataset for form understanding in noisy scanned documents. In: IEEE ICDARW, pp 1–6

Jia C, Liang X, Zhang Y (2019) Cross-domain NER using cross-domain language modeling. In: ACL, pp 2464–2474

Jia M, Shen L, Shen X, et al (2023) MNER-QG: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In: AAAI, pp 8032–8040

Jia C, Zhang Y (2020) Multi-cell compositional LSTM for NER domain adaptation. In: ACL, pp 5906–5917

Jie Z, Lu W (2019) Dependency-guided LSTM-CRF for named entity recognition. In: EMNLP-IJCNLP, pp 3862–3872

Ji B, Li S, Gan S, et al (2022) Few-shot named entity recognition with entity-level prototypical network enhanced by dispersedly distributed prototypes. In: COLING, pp 1842–1854

Jin Z, Cao P, He Z, et al (2023) Alignment precedes fusion: open-vocabulary named entity recognition as context-type semantic matching. In: EMNLP, pp 14616–14637

Ji Z, Xia T, Han M, et al (2021) A neural transition-based joint model for disease named entity recognition and normalization. In: ACL-IJCNLP, pp 2819–2827

Ju M, Miwa M, Ananiadou S (2018) A neural layered model for nested named entity recognition. In: HLT-NAACL, pp 1446–1459

Kambar MEZN, Esmaeilzadeh A, Heidari M (2022) A survey on deep learning techniques for joint named entities and relation extraction. In: IEEE AIIoT, pp 218–224

Karimi S, Metke-Jimenez A, Kemp M et al (2015) Cadec: A corpus of adverse drug event annotations. J Biomed Inform 55:73–81

Karthikeyan K, Wang Z, Mayhew S, et al (2019) Cross-lingual ability of Multilingual BERT: an empirical study. In: ICLR

Katiyar A, Cardie C (2018) Nested named entity recognition revisited. In: HLT-NAACL

Katti AR, Reisswig C, Guder C, et al (2018) Chargrid: Towards understanding 2d documents. In: EMNLP, pp 4459–4469

Ke W, Tian Z, Liu Q, et al (2023) Towards incremental NER data augmentation via syntactic-aware insertion transformer. In: ICJAI, pp 5104–5112

Khalid MA, Jijkoun V, De Rijke M (2008) The impact of named entity normalization on information retrieval for question answering. In: ECIR, Springer, pp 705–710

Kim JD, Ohta T, Tateisi Y et al (2003) GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics 19:i180–i182

Kim H, Kim H (2024) Recursive label attention network for nested named entity recognition. Expert Syst Appl 249:123657

Kim G, Hong T, Yim M, et al (2022) OCR-free document understanding Transformer. In: ECCV, pp 498–517

Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: ICLR

Kong L, Dyer C, Smith NA (2016) Segmental recurrent neural networks. In: ICLR

Küçük D, Arıcı N, Küçük D (2017) Named entity recognition in Turkish: approaches and issues. In: NLDB, pp 176–181

Kuru O, Can OA, Yuret D (2016) CharNER: Character-level named entity recognition. In: COLING, pp 911–921

Lai P, Ye F, Zhang L, et al (2022) PCBERT: Parent and child BERT for Chinese few-shot NER. In: COLING, pp 2199–2209

Lample G, Ballesteros M, Subramanian S, et al (2016) Neural architectures for named entity recognition. In: HLT-NAACL, pp 260–270

Lee DH, Kadakia A, Tan K, et al (2022) Good examples make a faster learner: simple demonstration-based learning for low-resource NER. In: ACL, pp 2687–2700

Lee S, Oh S, Jung W (2023) Enhancing low-resource fine-grained named entity recognition by leveraging coarse-grained datasets. In: EMNLP, pp 3269–3279

Le-Khac PH, Healy G, Smeaton AF (2020) Contrastive representation learning: a framework and review. IEEE Access 8:193907–193934

Lewis M, Liu Y, Goyal N, et al (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL, pp 7871–7880

Li J, Sun A, Han J et al (2020) A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng 34(1):50–70

Li F, Wang Z, Hui SC et al (2021) A segment enhanced span-based model for nested named entity recognition. Neurocomputing 465:26–37

Liang T, Xia C, Zhao Z et al (2023) Transferring from textual entailment to biomedical named entity recognition. IEEE/ACM Trans Comput Biol Bioinf 20(4):2577–2586

Liang S, Gong M, Pei J, et al (2021) Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In: KDD, pp 3231–3239

Liang C, Yu Y, Jiang H, et al (2020) BOND: BERT-assisted open-domain named entity recognition with distant supervision. In: KDD, pp 1054–1064

Liao W, Veeramachaneni S (2009) A simple semi-supervised algorithm for named entity recognition. In: HLT-NAACL, pp 58–65

Li J, Fei H, Liu J, et al (2022b) Unified named entity recognition as word-word relation classification. In: AAAI, pp 10965–10973

Li X, Feng J, Meng Y, et al (2020d) A unified MRC framework for named entity recognition. In: ACL, pp 5849–5859

Li P, Gu J, Kuen J, et al (2021c) SelfDoc: Self-supervised document representation learning. In: CVPR, pp 5652–5660

Li D, Hu B, Chen Q (2022a) Prompt-based text entailment for low-resource named entity recognition. In: COLING, pp 1896–1903

Li Z, Hu C, Guo X, et al (2022d) An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In: ACL, pp 170–179

Li Q, Ji H (2014) Incremental joint extraction of entity mentions and relations. In: ACL, pp 402–412

Li F, Lin Z, Zhang M, et al (2021a) A span-based model for joint overlapped and discontinuous named entity recognition. In: ACL-IJCNLP, pp 4814–4828

Li J, Li H, Pan Z, et al (2023a) Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In: EMNLP, pp 2787–2802

Li J, Li H, Sun D, et al (2024) LLMs as bridges: Reformulating grounded multimodal named entity recognition. In: ACL

Li Y, Liu L, Shi S (2020f) Empirical analysis of unlabeled entity problem in named entity recognition. In: ICLR

Li B, Liu S, Sun Y, et al (2020a) Recursively binary modification model for nested named entity recognition. In: AAAI, pp 8164–8171

Lin BY, Lee DH, Shen M, et al (2020) TriggerNER: Learning with entity triggers as explanations for named entity recognition. In: ACL, pp 8503–8511

Lin BY, Lu W (2018) Neural adaptation layers for cross-domain named entity recognition. In: EMNLP, pp 2012–2022

Lin TY, Goyal P, Girshick R, et al (2017) Focal loss for dense object detection. In: ICCV, pp 2999–3007

Li J, Shang S, Shao L (2020b) MetaNER: Named entity recognition with meta-learning. In: WWW, pp 429–440

Li Y, Shetty P, Liu L, et al (2021d) BERTifying the hidden Markov model for multi-source weakly supervised named entity recognition. In: ACL-IJCNLP, pp 6178–6190

Li Y, Song L, Zhang C (2022c) Sparse conditional hidden Markov model for weakly supervised named entity recognition. In: KDD, pp 978–988

Li J, Sun Y, Johnson RJ, et al (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database: the journal of biological databases and curation 2016

Li X, Sun X, Meng Y, et al (2020e) Dice loss for data-imbalanced NLP tasks. In: ACL, pp 465–476

Liu M, Tu Z, Zhang T et al (2022) LTP: A new active learning strategy for CRF-based named entity recognition. Neural Process Lett 54(3):2433–2454

Liu P, Guo Y, Wang F et al (2022) Chinese named entity recognition: the state of the art. Neurocomputing 473:37–53

Liu Q, Mao R, Geng X et al (2023) Semantic matching in machine reading comprehension: an empirical study. Inf Process Manag 60(2):103145

Liu J, Chen Y, Xu J (2022a) Low-resource ner by data augmentation with prompting. In: IJCAI, pp 4252–4258

Liu J, Fei H, Li F, et al (2024) TKDP: Threefold knowledge-enriched deep prompt tuning for few-shot named entity recognition. In: IEEE transactions on knowledge and data engineering, pp 1–14

Liu K, Hu Q, Liu J, et al (2017) Named entity recognition in Chinese electronic medical records based on CRF. In: IEEE WISA, pp 105–110

Liu B, Lee WS, Yu PS, et al (2002) Partially supervised classification of text documents. In: ICML, pp 387–394

Liu L, Shang J, Ren X, et al (2018) Empower sequence labeling with task-aware neural language model. In: AAAI, pp 5253–5260

Liu Z, Xu Y, Yu T, et al (2021) CrossNER: Evaluating cross-domain named entity recognition. In: AAAI, pp 13452–13460

Liu T, Yao JG, Lin CY (2019) Towards improving neural named entity recognition with gazetteers. In: ACL, pp 5301–5307

Li Y, Yu Y, Qian T (2023b) Type-aware decomposed framework for few-shot named entity recognition. In: EMNLP, pp 8911–8927

Lobov V, Ivoylova A, Sharoff S (2022) Applying natural annotation and curriculum learning to named entity recognition for under-resourced languages. In: COLING, pp 4468–4480

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. In: ICLR

Lou C, Yang S, Tu K (2022) Nested named entity recognition as latent lexicalized constituency parsing. In: ACL, pp 6183–6198

Luan Y, He L, Ostendorf M, et al (2018a) Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: EMNLP, pp 3219–3232

Luan Y, He L, Ostendorf M, et al (2018b) Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: EMNLP, pp 3219–3232

Luan Y, Wadden D, He L, et al (2019) A general framework for information extraction using dynamic span graphs. In: HLT-NAACL, pp 3036–3046

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: NIPS, pp 4768–4777

Lu D, Neves L, Carvalho V, et al (2018) Visual attention model for name tagging in multimodal social media. In: ACL, pp 1990–1999

Luoma J, Pyysalo S (2020) Exploring cross-sentence contexts for named entity recognition with BERT. In: COLING, pp 904–914

Luo Y, Xiao F, Zhao H (2020a) Hierarchical contextualized representation for named entity recognition. In: AAAI, pp 8441–8448

Luo Y, Zhao H (2020) Bipartite flat-graph network for nested named entity recognition. In: ACL, pp 6408–6418

Luo Y, Zhao H, Zhan J (2020b) Named entity recognition only from word embeddings. In: EMNLP, pp 8995–9005

Lu W, Roth D (2015) Joint mention extraction and classification with mention hypergraphs. In: EMNLP, pp 857–867

Lu J, Zhang D, Zhang J, et al (2022) Flat multi-modal interaction transformer for named entity recognition. In: COLING, pp 2055–2064

Lv B, Liu X, Dai S, et al (2023) DSP: Discriminative soft prompts for zero-shot entity and relation extraction. In: ACL, pp 5491–5505

Ma JY, Chen B, Gu JC, et al (2022b) Wider & closer: mixture of short-channel distillers for zero-shot cross-lingual named entity recognition. In: EMNLP, pp 5171–5183

Ma J, Ballesteros M, Doss S, et al (2022a) Label semantics for few shot named entity recognition. In: ACL, pp 1956–1971

Ma R, Chen X, Lin Z, et al (2023a) Learning “o” helps for learning more: handling the unlabeled entity problem for class-incremental ner. In: ACL, pp 5959–5979

Ma X, Hovy E (2016) End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. In: ACL, pp 1064–1074

Mai W, Zhang Z, Li K et al (2024) Dynamic graph construction framework for multimodal named entity recognition in social media. IEEE Trans Comput Soc Syst 11(2):2513–2522

Mai K, Pham TH, Nguyen MT, et al (2018) An empirical study on fine-grained named entity recognition. In: COLING, pp 711–722

Ma T, Jiang H, Wu Q, et al (2022e) Decomposed meta-learning for few-shot named entity recognition. In: ACL, pp 1584–1596

Ma R, Lin Z, Chen X, et al (2023b) Coarse-to-fine few-shot learning for named entity recognition. In: ACL, pp 4115–4129

Mao R, Liu Q, He K et al (2023) The biases of pre-trained language models: an empirical study on prompt-based sentiment analysis and emotion detection. IEEE Trans Affect Comput 14(3):1743–1753

Mao R, He K, Zhang X et al (2024) A survey on semantic processing techniques. Inf Fus 101:101988

Mao R, Chen G, Zhang X, et al (2024a) GPTEval: A survey on assessments of ChatGPT and GPT-4. In: LREC-COLING, p 7844–7866

Mao R, He K, Ong CB, et al (2024b) MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling. In: ACL, pp 9891–9908

Marrero M, Urbano J, Sánchez-Cuadrado S et al (2013) Named entity recognition: fallacies, challenges and opportunities. Comput Stand Interfaces 35(5):482–489

Ma R, Tan Y, Zhou X, et al (2022c) Searching for optimal subword tokenization in cross-domain NER. In: IJCAI, pp 4289–4295

Ma T, Wu Q, Jiang H, et al (2023c) CoLaDa: A collaborative label denoising framework for cross-lingual named entity recognition. In: ACL, pp 5995–6009

Mayhew S, Tsai CT, Roth D (2017) Cheap translation for cross-lingual named entity recognition. In: EMNLP, pp 2536–2545

Ma R, Zhou X, Gui T, et al (2022d) Template-free prompt tuning for few-shot NER. In: HLT-NAACL, pp 5721–5732

Mei X, Mao R, Cai X, et al (2024) Medical report generation via multimodal Spatio-temporal fusion. In: ACM MM

Merdjanovska E, Aynetdinov A, Akbik A (2024) NoiseBench: Benchmarking the impact of real label noise on named entity recognition. In: EMNLP, pp 18182–18198

Miller A, Fisch A, Dodge J, et al (2016) Key-value memory networks for directly reading documents. In: EMNLP, pp 1400–1409

Ming H, Yang J, Gui F et al (2024) Few-shot nested named entity recognition. Knowl-Based Syst 293:111688

Miwa M, Bansal M (2016) End-to-end relation extraction using LSTMs on sequences and tree structures. In: ACL, pp 1105–1116

Miwa M, Sasaki Y (2014) Modeling joint entity and relation extraction with table representation. In: EMNLP, pp 1858–1869

Monaikul N, Castellucci G, Filice S, et al (2021) Continual learning for named entity recognition. In: AAAI, pp 13570–13577

Moon S, Neves L, Carvalho V (2018) Multimodal named entity recognition for short social media posts. In: HLT-NAACL, pp 852–860

Morwal S, Jahan N, Chopra D (2012) Named entity recognition using hidden Markov model (HMM). Int J Natl Lang Comput 1(4)

Moscato V, Postiglione M, Sansone C et al (2023) TaughtNet: Learning multi-task biomedical named entity recognition from single-task teachers. IEEE J Biomed Health Inform 27(5):2512–2523

Moscato V, Postiglione M, Sperlí G (2023) Few-shot named entity recognition: definition, taxonomy and research directions. ACM Trans Intell Syst Technol 14(5):1–46

Moscato V, Postiglione M, Sperlì G, et al (2024) Aldaner: Active learning based data augmentation for named entity recognition. Knowl-Based Syst, p 112682

Mowery DL, Velupillai S, South BR, et al (2014) Task 2: ShARe/CLEF eHealth evaluation lab 2014. In: CLEF

Mo Y, Yang J, Liu J, et al (2024) MCL-NER: Cross-lingual named entity recognition via multi-view contrastive learning. In: AAAI, pp 18789–18797

Muis AO, Lu W (2016) Learning to recognize discontiguous entities. In: EMNLP, pp 75–84

Nasar Z, Jaffry SW, Malik MK (2021) Named entity recognition and relation extraction: state-of-the-art. ACM Comput Surv 54(1):1–39

Navarro DF, Ijaz K, Rezazadegan D et al (2023) Clinical named entity recognition and relation extraction using natural language processing of medical free text: a systematic review. Int J Med Inf 177:105122

Nguyen NT, Miwa M, Ananiadou S (2023b) Span-based named entity recognition by generating and compressing information. In: EACL, pp 1984–1996

Nguyen ND, Tan W, Du L, et al (2023a) AUC maximization for low-resource named entity recognition. In: AAAI, pp 13389–13399

Ni J, Dinu G, Florian R (2017) Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In: ACL, pp 1470–1480

Nie Y, Tian Y, Song Y, et al (2020) Improving named entity recognition with attentive ensemble of syntactic information. In: EMNLP, pp 4231–4245

Ok H, Kil T, Seo S, et al (2024) SCANNER: Knowledge-enhanced approach for robust multi-modal named entity recognition of unseen entities. In: HLT-NAACL, pp 7718–7730

Peng Q, Zheng C, Cai Y et al (2021) Unsupervised cross-domain named entity recognition using entity-aware adversarial training. Neural Netw 138:68–77

Peng Q, Pan Y, Wang W, et al (2022) ERNIE-layout: layout knowledge enhanced pre-training for visually-rich document understanding. In: EMNLP, pp 3744–3756

Pennington J, Socher R, Manning CD (2014) GloVe: Global vectors for word representation. In: EMNLP, pp 1532–1543

Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. In: HLT-NAACL, pp 2227–2237

Popovski G, Kochev S, Korousic-Seljak B, et al (2019) FoodIE: A rule-based named-entity recognition method for food information extraction. In: ICPRAM, pp 915–922

Pradhan S, Elhadad N, South BR, et al (2013) Task 1: ShARe/CLEF eHealth evaluation lab 2013. In: CLEF

Qian S, Jin W, Chen Y, et al (2023) A survey on multimodal named entity recognition. In: ICIC, pp 609–622

Qiao B, Zou Z, Huang Y et al (2022) A joint model for entity and relation extraction based on BERT. Neural Comput Appl 34:3471–3481

Qin C, Joty S (2021) LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of T5. In: ICLR

Quimbaya AP, Múnera AS, Rivera RAG et al (2016) Named entity recognition over electronic health records through a combined dictionary-based approach. Procedia Comput Sci 100:55–61

Qu X, Zeng J, Liu D, et al (2023) Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble. In: AAAI, pp 13501–13509

Radford A, Narasimhan K, Salimans T, et al (2018) Improving language understanding by generative pre-training. preprint

Raffel C, Shazeer N, Roberts A et al (2020) Exploring the limits of transfer learning with a unified text-to-text Transformer. J Mach Learn Res 21(140):5485–5551

Rahimi A, Li Y, Cohn T (2019) Massively multilingual transfer for ner. In: ACL, pp 151–164

Ramage D, Hall D, Nallapati R, et al (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP, pp 248–256

Ravi S, Larochelle H (2016) Optimization as a model for few-shot learning. In: ICLR

Rei M (2017) Semi-supervised multitask learning for sequence labeling. In: ACL, pp 2121–2130

Reich A, Chen J, Agrawal A, et al (2022) Leveraging expert guided adversarial augmentation for improving generalization in named entity recognition. In: ACL, pp 1947–1955

Ren Z, Qin X, Ran W (2023) SLNER: Chinese few-shot named entity recognition with enhanced span and label semantics. Appl Sci 13(15):8609

Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" Explaining the predictions of any classifier. In: KDD, pp 1135–1144

Ritter A, Clark S, Etzioni O, et al (2011) Named entity recognition in tweets: an experimental study. In: EMNLP, pp 1524–1534

Rojas M, Bravo-Marquez F, Dunstan J (2022) Simple yet powerful: an overlooked architecture for nested named entity recognition. In: COLING, pp 2108–2117

Roth D, Yih Wt (2004) A linear programming formulation for global inference in natural language tasks. In: HLT-NAACL, pp 1–8

Sarawagi S, Cohen WW (2004) Semi-Markov conditional random fields for information extraction. In: NIPS, pp 1185–1192

Sekine S (2008) Extended named entity ontology with attribute information. In: LREC

Shang J, Liu L, Gu X, et al (2018) Learning named entity tagger using domain-specific dictionary. In: EMNLP, pp 2054–2064

Shen Y, Song K, Tan X, et al (2023a) DiffusionNER: Boundary diffusion for named entity recognition. In: ACL, pp 3875–3890

Shen Y, Tan Z, Wu S, et al (2023b) PromptNER: Prompt locating and typing for named entity recognition. In: ACL, pp 12492–12507

Shen Y, Wang X, Tan Z, et al (2022) Parallel instance query network for named entity recognition. In: ACL, pp 947–961

Shibuya T, Hovy E (2020) Nested named entity recognition via second-best sequence learning and decoding. Trans Assoc Comput Linguist 8:605–620

Snell J, Swersky K, Zemel R (2017) Prototypical networks for few-shot learning. In: NIPS, pp 4080–4090

Sohrab MG, Miwa M (2018) Deep exhaustive model for nested named entity recognition. In: EMNLP, pp 2843–2849

Song S, Shen F, Zhao J (2024) Ropda: Robust prompt-based data augmentation for low-resource named entity recognition. AAAI 38(17):19017–19025

Straková J, Straka M, Hajic J (2019) Neural architectures for nested ner through linearization. In: ACL, pp 5326–5331

Sui Y, Bu F, Hu Y, et al (2022) Trigger-GNN: A trigger-based graph neural network for nested named entity recognition. In: IEEE IJCNN, pp 01–08

Sun C, Yang Z, Wang L et al (2021) Biomedical named entity recognition using BERT in the machine reading comprehension framework. J Biomed Inform 118:103799

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: ICML, pp 3319–3328

Sung F, Yang Y, Zhang L, et al (2018) Learning to compare: relation network for few-shot learning. In: CVPR, pp 1199–1208

Sun L, Wang J, Su Y, et al (2020) RIVA: a pre-trained tweet multimodal model based on text-image relation for multimodal NER. In: COLING, pp 1852–1862

Sun L, Wang J, Zhang K, et al (2021b) RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In: AAAI, pp 13860–13868

Tang X, Xia D, Li Y, et al (2023a) A survey of low-resource named entity recognition. In: ISCC, pp 246–260

Tang G, Xie L, Jin L, et al (2021) MatchVIE: Exploiting match relevancy between entities for visual information extraction. In: IJCAI, pp 1039–1045

Tang Z, Yang Z, Wang G, et al (2023b) Unifying vision, text, and layout for universal document processing. In: CVPR, pp 19254–19264

Tan C, Qiu W, Chen M, et al (2020a) Boundary enhanced neural span classification for nested named entity recognition. In: AAAI, pp 9016–9023

Tan C, Qiu W, Chen M, et al (2020b) Boundary enhanced neural span classification for nested named entity recognition. In: AAAI, pp 9016–9023

Tan Z, Shen Y, Zhang S, et al (2021) A sequence-to-set network for nested named entity recognition. In: IJCAI, pp 3936–3942

Tian Y, Sun X, Yu H et al (2021) Hierarchical self-adaptation network for multimodal named entity recognition in social media. Neurocomputing 439:12–21

Tjong Kim Sang EF, De Meulder F (2003) Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: HLT-NAACL, pp 142–147

Tjong Kim Sang EF (2002) Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: COLING, pp 155–158

Tsai CT, Mayhew S, Roth D (2016) Cross-lingual named entity recognition via wikification. In: CoNLL, pp 219–228

Ugawa A, Tamura A, Ninomiya T, et al (2018) Neural machine translation incorporating named entity. In: COLING, pp 3240–3250

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: NIPS, pp 6000–6010

Veena G, Kanjirangat V, Gupta D (2023) Agroner: An unsupervised agriculture named entity recognition using weighted distributional semantic model. Expert Syst Appl 229:120440

Vinyals O, Blundell C, Lillicrap T, et al (2016) Matching networks for one shot learning. In: NIPS, pp 3637–3645

Wadden D, Wennberg U, Luan Y, et al (2019) Entity, relation, and event extraction with contextualized span representations. In: EMNLP-IJCNLP, pp 5784–5789

Walker C, Strassel S, Medero J, et al (2006) Ace 2005 multilingual training corpus. (No Title)

Wan Q, Wei L, Chen X et al (2021) A region-based hypergraph network for joint entity-relation extraction. Knowl-Based Syst 228:107298

Wang Y, Li Y, Zhu Z et al (2020) Adversarial learning for multi-task sequence labeling with attention mechanism. IEEE/ACM Trans Audio, Speech, Lang Process 28:2476–2488

Wang Y, Tong H, Zhu Z et al (2022) Nested named entity recognition: a survey. ACM Trans Knowl Discov Data 16(6):1–29

Wang X, Cai J, Jiang Y, et al (2022d) Named entity and relation extraction with multi-modal retrieval. In: EMNLP, pp 5925–5936

Wang Y, Chu H, Zhang C, et al (2021c) Learning from language description: low-shot named entity recognition via decomposed framework. In: EMNLP, pp 1618–1630

Wang X, Gui M, Jiang Y, et al (2022e) ITA: Image-text alignments for multi-modal named entity recognition. In: HLT-NAACL, pp 3176–3189

Wang X, Jiang Y, Bach N, et al (2021b) Improving named entity recognition by external context retrieving and cooperative learning. In: ACL-IJCNLP, pp 1800–1812

Wang J, Jin L, Ding K (2022a) LiLT: A simple yet effective language-independent layout Transformer for structured document understanding. In: ACL, pp 7747–7757

Wang Z, Karthikeyan K, Mayhew S, et al (2020d) Extending multilingual BERT to low-resource languages. In: EMNLP, pp 2649–2656

Wang Y, Li Y, Tong H, et al (2020b) HIT: Nested named entity recognition via head-tail pair and token interaction. In: EMNLP, pp 6027–6036

Wang J, Liu C, Jin L, et al (2021a) Towards robust visual information extraction in real world: new dataset and novel solution. In: AAAI, pp 2738–2745

Wang B, Lu W, Wang Y, et al (2018a) A neural transition-based model for nested mention recognition. In: EMNLP, pp 1011–1017

Wang S, Meng Y, Ouyang R, et al (2023) GNN-SL: Sequence labeling based on nearest examples via GNN. In: ACL, pp 12679–12692

Wang Y, Mukherjee S, Chu H, et al (2021d) Meta self-training for few-shot neural sequence labeling. In: KDD, pp 1737–1747

Wang Z, Qu Y, Chen L, et al (2018b) Label-aware double transfer learning for cross-specialty medical named entity recognition. In: HLT-NAACL, pp 1–15

Wang J, Shou L, Chen K, et al (2020a) Pyramid: A layered model for nested named entity recognition. In: ACL, pp 5918–5928

Wang J, Wang C, Tan C, et al (2022b) SpanProto: A two-stage span-based prototypical network for few-shot named entity recognition. In: EMNLP, pp 3466–3476

Wang R, Yu T, Zhao H, et al (2022c) Few-shot class-incremental learning for named entity recognition. In: ACL, pp 571–582

Weischedel R, Pradhan S, Ramshaw L, et al (2013) OntoNotes release 5.0. LDC2013T19, Philadelphia, Penn: Linguistic Data Consortium

Wu J, He K, Mao R et al (2023) MEGACare: Knowledge-guided multi-view hypergraph predictive framework for healthcare. Inf Fus 100:101939

Wu H, Ding R, Zhao H, et al (2023a) Adversarial self-attention for language understanding. In: AAAI, pp 13727–13735

Wu S, Dredze M (2019) Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In: EMNLP-IJCNLP, pp 833–844

Wu C, Ke W, Wang P, et al (2024) ConsistNER: Towards instructive ner demonstrations for llms with the consistency of ontology and context. In: AAAI, pp 19234–19242

Wu Q, Lin Z, Karlsson BF, et al (2021) UniTrans: unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In: IJCAI, pp 3926–3932

Wu Q, Lin Z, Karlsson B, et al (2020b) Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In: ACL, pp 6505–6514

Wu Q, Lin Z, Wang G, et al (2020c) Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In: AAAI, pp 9274–9281

Wu Q, Lin Z, Wang G, et al (2020d) Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In: AAAI, pp 9274–9281

Wu S, Shen Y, Tan Z, et al (2022) Propose-and-refine: a two-stage set prediction network for nested named entity recognition. In: IJCAI, pp 4418–4424

Wu C, Wu F, Qi T, et al (2020a) Named entity recognition with context-aware dictionary knowledge. In: CCL, pp 915–926

Xiang W, Wang B (2019) A survey of event extraction from text. IEEE Access 7:173111–173137

Xia Y, Wang Q, Lyu Y, et al (2022) Learn and review: enhancing continual named entity recognition via reviewing synthetic samples. In: ACL, pp 2291–2300

Xia C, Zhang C, Yang T, et al (2019) Multi-grained named entity recognition. In: ACL, pp 1430–1440

Xie T, Li Q, Zhang Y, et al (2023) Self-improving for zero-shot named entity recognition with large language models. In: HLT-NAACL, pp 583–593

Xie J, Yang Z, Neubig G, et al (2018) Neural cross-lingual named entity recognition with minimal resources. In: EMNLP, pp 369–379

Xu Y, Huang H, Feng C, et al (2021b) A supervised multi-head self-attention network for nested named entity recognition. In: AAAI, pp 14185–14193

Xu B, Huang S, Sha C, et al (2022) MAF: a general matching and alignment framework for multimodal named entity recognition. In: WSDM, pp 1215–1223

Xu L, Jie Z, Lu W, et al (2021a) Better feature integration for named entity recognition. In: HLT-NAACL, pp 3457–3469

Xu Y, Li M, Cui L, et al (2020) LayoutLM: Pre-training of text and layout for document image understanding. In: KDD, pp 1192–1200

Xu W, Li X, Zhang W, et al (2023a) From cloze to comprehension: retrofitting pre-trained masked language models to pre-trained machine reader. In: NIPS, pp 67309–67328

Xu Y, Xu Y, Lv T, et al (2021c) LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: ACL-IJCNLP, pp 2579–2591

Xu Y, Yang Z, Zhang L, et al (2023b) Focusing, bridging and prompting for few-shot nested named entity recognition. In: ACL, pp 2621–2637

Yamada I, Asai A, Shindo H, et al (2020) LUKE: Deep contextualized entity representations with entity-aware self-attention. In: EMNLP, pp 6442–6454

Yan Y, Cai B, Song S (2023a) Nested named entity recognition as building local hypergraphs. In: AAAI, pp 13878–13886

Yang Z, Ma J, Chen H et al (2022) Context-aware attentive multilevel feature fusion for named entity recognition. IEEE Trans Neural Netw Learn Syst 35(1):973–984

Yang Y, Chen W, Li Z, et al (2018) Distantly supervised NER with partial annotation learning and reinforcement learning. In: COLING, pp 2159–2169

Yang Y, Hu X, Ma F, et al (2023) Gaussian prior reinforcement learning for nested named entity recognition. In: IEEE ICASSP, pp 1–5

Yang Y, Katiyar A (2020) Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In: EMNLP, pp 6365–6375

Yang Z, Ma J, Chen H, et al (2021) HiTRANS: A hierarchical transformer network for nested named entity recognition. In: EMNLP, pp 124–132

Yang S, Tu K (2022) Bottom-up constituency parsing and nested named entity recognition with pointer networks. In: ACL, pp 2403–2416

Yan H, Gui T, Dai J, et al (2021) A unified generative framework for various NER subtasks. In: ACL-IJCNLP, pp 5808–5822

Yang L, Yuan L, Cui L, et al (2022a) FactMix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In: COLING, pp 5360–5371

Yan Z, Yang S, Liu W, et al (2023b) Joint entity and relation extraction with span pruning and hypergraph neural networks. In: EMNLP, pp 7512–7526

Ye Z, Ling ZH (2018) Hybrid semi-Markov CRF for neural sequence labeling. In: ACL, pp 235–240

Yu J, Bohnet B, Poesio M (2020a) Named entity recognition as dependency parsing. In: ACL, pp 6470–6476

Yu J, Jiang J, Yang L, et al (2020b) Improving multimodal named entity recognition via entity span detection with unified multimodal Transformer. In: ACL, pp 3342–3352

Yu J, Li Z, Wang J, et al (2023) Grounded multimodal named entity recognition on social media. In: ACL, pp 9141–9154

Yu W, Lu N, Qi X, et al (2021) Pick: processing key information extraction from documents using improved graph learning-convolutional networks. In: IEEE ICPR, pp 4363–4370

Zafarian A, Rokni A, Khadivi S, et al (2015) Semi-supervised learning for named entity recognition using weakly labeled training data. In: IEEE AISP, pp 129–135

Zaratiana U, Tomeh N, El Khbir N, et al (2023) Filtered semi-Markov CRF. In: EMNLP, pp 222–235

Zaratiana U, Tomeh N, Holat P, et al (2022) GNNer: Reducing overlapping in span-based ner using graph neural networks. In: ACL, pp 97–103

Zaratiana U, Tomeh N, Holat P, et al (2024) GLiNER: Generalist model for named entity recognition using bidirectional transformer. In: HLT-NAACL, pp 5364–5376

Zeng X, Li Y, Zhai Y, et al (2020) Counterfactual generator: a weakly-supervised method for named entity recognition. In: EMNLP, pp 7270–7280

Zhang Y, Zhang H (2023) FinBERT-MRC: financial named entity recognition using BERT under the machine reading comprehension paradigm. Neural Process Lett 55(6):7393–7413

Zhang Z, Zhang H, Wan Q et al (2022) LELNER: A lightweight and effective low-resource named entity recognition model. Knowl-Based Syst 251:109178

Zhang Y, Chen Q (2023) A neural span-based continual named entity recognition model. In: AAAI, pp 13993–14001

Zhang S, Cheng H, Gao J, et al (2022a) Optimizing Bi-encoder for named entity recognition via contrastive learning. In: ICLR

Zhang D, Cong W, Dong J, et al (2023b) Continual named entity recognition without catastrophic forgetting. In: EMNLP, pp 8186–8197

Zhang Q, Fu J, Liu X, et al (2018) Adaptive co-attention network for named entity recognition in tweets. In: AAAI, pp 5674–5681

Zhang C, Guo Y, Tu Y, et al (2023a) Reading order matters: information extraction from visually-rich documents by token path prediction. In: EMNLP, pp 13716–13730

Zhang Z, Hu M, Zhao S, et al (2023h) E-NER: Evidential deep learning for trustworthy named entity recognition. In: ACL, pp 1619–1634

Zhang X, Jiang Y, Wang X, et al (2022b) Domain-specific NER via retrieving correlated samples. In: COLING, pp 2398–2404

Zhang J, Liu X, Lai X, et al (2023d) 2INER: Instructive and in-context learning on few-shot named entity recognition. In: EMNLP, pp 3940–3951

Zhang M, Qiao X, Zhao Y, et al (2023f) SmartSpanNER: Making SpanNER robust in low resource scenarios. In: EMNLP, pp 7964–7976

Zhang D, Wei S, Li S, et al (2021a) Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: AAAI, pp 14347–14355

Zhang X, Yuan J, Li L, et al (2023g) Reducing the bias of visual objects in multimodal named entity recognition. In: WSDM, pp 958–966

Zhang D, Yu Y, Chen F, et al (2023c) Decomposing logits distillation for incremental named entity recognition. In: SIGIR, pp 1919–1923

Zhang X, Yu B, Liu T, et al (2021b) Improving distantly-supervised named entity recognition with self-collaborative denoising learning. In: EMNLP, pp 10746–10757

Zhang X, Yu B, Wang Y, et al (2022c) Exploring modular task decomposition in cross-domain named entity recognition. In: SIGIR, pp 301–311

Zhang J, Zhang Y, Chen Y, et al (2023e) Structure and label constrained data augmentation for cross-domain few-shot ner. In: EMNLP, pp 518–530

Zhang Z, Zhao Y, Gao H, et al (2024) Linkner: Linking local named entity recognition models to large language models using uncertainty. In: WWW, pp 4047–4058

Zhao S, Liu T, Zhao S, et al (2019) A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In: AAAI, pp 817–824

Zheng S, Hao Y, Lu D et al (2017) Joint entity and relation extraction based on a hybrid neural network. Neurocomputing 257:59–66

Zheng C, Wu Z, Wang T et al (2021) Object-aware multimodal named entity recognition in social media posts with adversarial learning. IEEE Trans Multimedia 23:2520–2532

Zheng C, Cai Y, Xu J, et al (2019) A boundary-aware neural model for nested named entity recognition. In: EMNLP-IJCNLP, pp 357–366

Zheng J, Liang Z, Chen H, et al (2022) Distilling causal effect from miscellaneous other-class for continual named entity recognition. In: EMNLP, pp 3602–3615

Zheng S, Wang F, Bao H, et al (2017b) Joint extraction of entities and relations based on a novel tagging scheme. In: ACL, pp 1227–1236

Zheng Z, Zhang Z, Wang Z, et al (2024) Decompose, prioritize, and eliminate: dynamically integrating diverse representations for multimodal named entity recognition. In: LREC-COLING, pp 4498–4508

Zhong X, Cambria E (2021) Time expression and named entity recognition, Springer

Zhong X, Cambria E, Hussain A (2022) Does semantics aid syntax? An empirical study on named entity recognition and classification. Neural Comput Appl 34(11):8373–8384

Zhou JT, Zhang H, Jin D, et al (2019) Dual adversarial neural transfer for low-resource named entity recognition. In: ACL, pp 3461–3471

Zhou B, Cai X, Zhang Y, et al (2021a) MTAAL: Multi-task adversarial active learning for medical named entity recognition and normalization. In: AAAI, pp 14586–14593

Zhou R, Li X, Bing L, et al (2022c) ConNER: Consistency training for cross-lingual named entity recognition. In: EMNLP, pp 8438–8449

Zhou R, Li X, Bing L, et al (2023a) Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In: ACL, pp 4018–4031

Zhou R, Li X, He R, et al (2022d) MELM: Data augmentation with masked entity language modeling for low-resource ner. In: ACL, pp 2251–2262

Zhou K, Li Y, Li Q (2022b) Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In: ACL, pp 7198–7211

Zhou W, Zhang S, Gu Y, et al (2023b) UniversalNER: Targeted distillation from large language models for open named entity recognition. In: ICLR

Zhou B, Zhang Y, Song K, et al (2022a) A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In: EMNLP, pp 6293–6302

Zhou X, Zhang X, Tao C, et al (2021b) Multi-grained knowledge distillation for named entity recognition. In: HLT-NAACL, pp 5704–5716

Zhuang W, Yijia Z, Kang A, et al (2023) P-MNER: Cross modal correction fusion network with prompt learning for multimodal named entity recognitiong. In: CCL, pp 689–700

Zhu E, Li J (2022) Boundary smoothing for named entity recognition. In: ACL, pp 7096–7108

Zhu E, Liu Y, Li J (2023) Deep span representations for named entity recognition. In: ACL, pp 10565–10582

Zhu W, Liu J, Xu J, et al (2021) Improving low-resource named entity recognition via label-aware data augmentation and curriculum denoising. In: CCL, pp 1131–1142

Zugarini A, Rigutini L (2023) SAGE: Semantic-aware global explanations for named entity recognition. In: IEEE IJCNN, pp 1–8