Natural Language Processing (2024), 1–21
doi:10.1017/nlp.2024.37

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Statistical dataset evaluation: A case study on named entity recognition

Chengwen Wang[1], Qingxiu Dong[2], Xiaochen Wang[2] and Zhifang Sui[2]

[1]School of International Cultural Exchange, Central University of Finance and Economics, Beijing, China and [2]MOE Key Lab of Computational Linguistics, School of Computer Science, Peking University, Beijing, China
Corresponding author: Zhifang Sui; Email: szf@pku.edu.cn

(Received 17 December 2023; revised 9 May 2024; accepted 10 May 2024)

Abstract

Datasets serve as crucial training resources and model performance trackers. However, existing datasets have exposed a plethora of problems, inducing biased models and unreliable evaluation results. In this paper, we propose a model-agnostic dataset evaluation framework for automatic dataset quality evaluation. We seek the statistical properties of the datasets and address three fundamental dimensions: reliability, difficulty, and validity, following a Classical Test Theory (CTT). Taking the named entity recognition (NER) datasets as a case study, we introduce nine statistical metrics for a statistical dataset evaluation framework. Specifically, we investigate the reliability of a NER dataset with three metrics, including Redundancy, Accuracy, and Leakage Ratio. We assess the dataset difficulty through four metrics: Unseen Entity Ratio, Entity Ambiguity Degree, Entity Density, and Model Differentiation. For validity, we introduce the Entity Imbalance Degree and Entity-Null Rate to evaluate the effectiveness of the dataset in assessing language model performance. Experimental results validate that our evaluation framework effectively assesses various aspects of the dataset quality. Furthermore, we study how the dataset scores on our statistical metrics affect the model performance and appeal for dataset quality evaluation or targeted dataset improvement before training or testing models.

Keywords: Dataset evaluation framework; named entity recognition; reliability; difficulty; validity

1. Introduction

Recently, a large number of models have made breakthroughs in various datasets of natural language processing (NLP) (Kenton and Toutanova 2019; Liu *et al.* 2019). Meanwhile, an increasing number and variety of NLP datasets are proposed for model training and evaluation (Malmasi *et al.* 2022; Yin *et al.* 2017; Srivastava *et al.* 2022).

However, despite datasets significantly impacting model development and assessment (Bommasani *et al.* 2021), their quality is seldom systematically verified. Recent literature has indicated various quality issues within NLP datasets, for example, label mistakes (Wang *et al.* 2019). Datasets with quality issues frequently give rise to model shortcuts (Gururangan *et al.* 2022; Poliak *et al.* 2018) or induce incorrect conclusions (Goyal *et al.* 2022; Rashkin *et al.* 2023).

In this paper, we aim to answer two primary questions: (1) How to evaluate dataset quality in a *model-agnostic* manner? A comprehensive dataset quality evaluation is crucial for selecting adequate training resources. Furthermore, when there are discrepancies in model performance across different datasets, an unbiased evaluation of dataset quality can serve as a reliable arbitrator. (2) How do the statistical scores on dataset properties affect the model performance? The insights

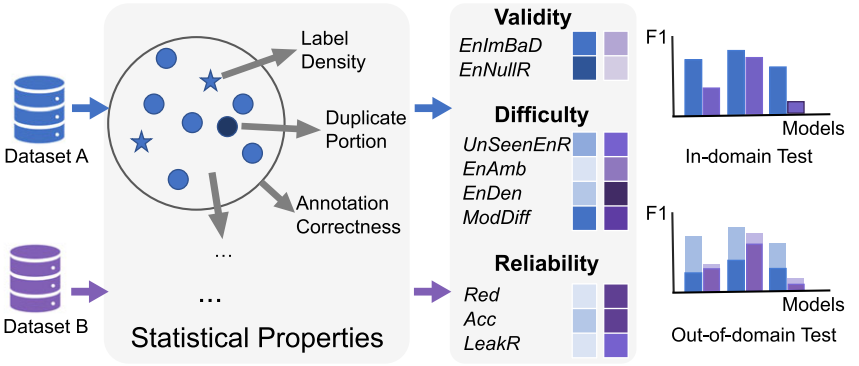

Figure 1. Our statistical dataset evaluation framework based on the Classical Testing Theory. We introduce nine quality evaluation metrics from three dimensions: reliability, difficulty, and validity. The dataset scores on the metrics have a significant impact on the models (trained on this dataset) in many aspects, such as the average performance and the out-of-domain robustness.

gained from this will guide improvements in dataset quality, which is crucial for developing effective and unbiased models.

To this end, we introduce a dataset evaluation framework (Figure 1) and take the named entity recognition (NER) datasets as a case study. Inspired by Classical Test Theory (CTT) (Novick 1966) in psychometrics, our dataset evaluation framework includes three key dimensions: reliability, difficulty, and validity. Reliability reflects how credible the dataset is, difficulty represents dataset difficulty and differentiation for models, and validity means how well the dataset fits the motivation and goal of the task. Following this framework, we introduce nine metrics under the three dimensions for the statistical properties of NER datasets and assess the quality of ten widely used NER datasets.

Extensive experimental results validate that our evaluation metrics derived from the dataset properties are highly correlated with the performance of NER models and human evaluation results. The evaluation results enhance our comprehension of the datasets and bring some novel insights. For example, one of the most widely used English NER datasets, *CoNLL03* (Sang and De Meulder 2003), is far less challenging (0.43, 0.30, and 2.63 points lower on the Unseen Entity Ratio, Entity Ambiguity Degree, and Model Differentiation metrics, respectively) than *WNUT16* (Strauss *et al.* 2016), which has received less attention previously. In addition, by controlled dataset adjustment (Sec. 6.4), we find the dataset quality on the statistical metrics, including Unseen Entity Ratio, Entity Ambiguity Degree, and Entity-Null Rate, affects the NER model performance significantly.

We believe that statistical dataset evaluation provides a direct and comprehensive reflection of the dataset quality. And we recommend dataset quality evaluation before training or testing models for a better understanding of tasks and data for other tasks in NLP.

2. Related work

2.1 Issues in NLP datasets

Recent works have shown that NLP datasets have a number of quality problems, for example, label mistakes (Wang *et al.* 2019), entity missing[a] (Tejaswin *et al.* 2021), and unwanted biases resulting from the annotation process (Kaushik and Lipton 2018; Nadeem *et al.* 2021). For instance, Wang *et al.* (2019) identified a notable 5.38 percent rate of label mistakes in the CoNLL03 NER dataset,

[a]The target summary contains entities (names, dates, events, etc.) that are absent from the source.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

a concerning figure for a widely used benchmark in NLP research. Tejaswin *et al.* (2021) manually checked 600 randomly selected instances from three sources: CNN/DailyMail (Hermann *et al.* 2015; Nallapati *et al.* 2016), Gigaword (Rush *et al.* 2015), and XSum (Narayan *et al.* 2018), which are datasets commonly used for text summarization tasks. Their analysis revealed a significant proportion of instances with issues of Entity Missing and Evidence Missing[b] in these datasets. This indicates that the target summaries often contained entities or concepts absent from the source texts, raising questions about the accuracy of these datasets.

Furthermore, studies by Sugawara *et al.* (2020) and Gururangan *et al.* (2022) suggest that performance metrics on certain machine reading comprehension and natural language inference datasets might be artificially inflated. This is attributed to models exploiting spurious correlations rather than truly understanding the underlying language structures, resulting in poor generalization when applied to real-world scenarios.

An equally significant concern in NLP dataset construction is data leakage, particularly test-train overlap, which poses substantial risks to model evaluation. Studies like Lewis *et al.* (2021) reveal that a considerable portion of test data may mirror the training set, risking models' overfitting to the data rather than generalizing, thus inflating performance scores. Larson *et al.* (2023) echoes this sentiment, highlighting similar concerns in document classification realms. These studies collectively call for improved dataset division methods and robust validation techniques to mitigate data leakage and truly measure a model's generalization capabilities on unseen data.

However, most works focus on a specific issue of the datasets, and most issues are highly related to the model training process. Inspired by CTT, we built our dataset quality evaluation framework from reliability, difficulty, and validity dimensions. And we developed metrics for assessing the quality of datasets under the above three dimensions in conjunction with NER task characteristics and experimentally validated the effectiveness of our metrics in dataset evaluation.

2.2 Data-centric AI

In the contemporary landscape of NLP and machine learning, the pivotal role of datasets has increasingly been acknowledged. The seminal work by Ng *et al.* (2021) has galvanized the shift toward a data-centric AI paradigm, underscoring the potential of enhancing data quality to achieve superior model performance over merely refining algorithms. This approach dovetails with the initiatives like the NHS's Data Quality Maturity Index Methodology,[c] which provides a structured framework to assess and improve the quality of data in healthcare, a sector that greatly benefits from NLP technologies.

Simultaneously, the introduction of DataCLUE by Xu *et al.* (2021) marks a significant stride in this domain, offering the first benchmark specifically tailored for evaluating data-centric approaches in NLP. This benchmark aligns with tools such as the Data Quality for AI Tool (Jariwala *et al.* 2022) provided by IBM, which facilitates exploratory data analysis through its API, thereby enabling a more rigorous and systematic enhancement of datasets.

Moreover, a comprehensive review (Zha *et al.* 2023) offers a detailed exploration of the need for data-centric AI, addressing the methodological pivot from a model-centric to a data-centric perspective in AI research. This survey highlights the indispensable need for high-quality data to train robust machine learning models, especially in domains where data are prone to noise, sparsity, and bias.

In light of these developments, our proposed evaluation metrics aim to contribute to the ongoing efforts of dataset quality improvement. These metrics are designed to facilitate both automatic and semi-automatic enhancements of datasets, ensuring that the data used to train NLP models

[b] Evidence Missing: The target summary is based on concepts which are absent from the source.

[c] https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality/data-quality-maturity-index-methodology

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

are of the highest fidelity and thus capable of driving the performance of these models to new heights. The systematic application of such metrics can significantly streamline the process of data quality assurance, making it more tractable for researchers and practitioners to achieve data excellence in AI systems.

The cumulative effect of these methodologies and tools signifies a transformative movement in AI research, where data are no longer a passive element but a dynamic and critical component of the AI development lifecycle. As this data-centric ethos permeates the field, it is anticipated that future advancements in NLP will be increasingly driven by innovations in data quality management, thereby catalyzing a new era of AI systems that are both powerful and reliable.

3. Classical Test Theory

Human tests or exams usually follow strict testing theories, such as CTT (Novick 1966), a statistical framework to measure the quality of the exams. According to CTT, a thorough and systematic evaluation should consider three dimensions: reliability, difficulty, and validity.

In this paper, we introduce **CTT for Dataset Evaluation**. Adapting traditional CTT to dataset evaluation, we specified the definitions of reliability, difficulty, and validity as follows:

- **Reliability** measures the trustworthiness of the evaluation dataset. For instance, datasets with a high number of labeling errors lack sufficient confidence to evaluate the performance of different models.
- **Difficulty** is used to assess how the dataset differentiates between various models and human–machine performance in terms of difficulty.
- **Validity** aims to evaluate how well the dataset effectively measures the capability of models.

4. Dataset quality evaluation framework

Following CTT for Dataset Evaluation, we build our statistical dataset evaluation framework[d] and apply it to NER datasets. It includes nine fundamental metrics of the statistic properties in the NER datasets. In this section, we introduce the definitions and the mathematical formulations of the proposed metrics.

For a dataset[e] D with n instances, let $(x^{(i)}, y^{(i)})$ represent the i-th instance ($i = 1, 2, \ldots, n$). The input sequence $x^{(i)}$ consists of $m^{(i)}$ tokens, and the output sequence $y^{(i)}$ consists of $m^{(i)}$ entity values. Let $\mathscr{C}$ represent the entity types in D (including "Not an entity"), and each entity type $c_j \in \mathscr{C}, j \in 1, 2, \ldots, v$, where v represents the total number of entity types. We use Te, Tr, De to represent the test set, the training set, and the development set, respectively. The function $e(y_D)$ is defined to obtain a set of entity values in the set of $y^{(i)}$ of D, y, and sometimes we omit D for simplification.

4.1 Metrics under reliability

The metrics under reliability aim to evaluate how accurate and trustworthy a dataset is, including **Redundancy**, **Accuracy**, and **Leakage Ratio**. Reliability metrics—Redundancy, Accuracy, and Leakage Ratio—are key elements in assessing a dataset's trustworthiness. The evaluation of

[d] Our framework is fundamentally designed to be applicable at a macroscopic level across various NLP tasks by advocating for dataset quality assessment through measures of reliability, validity, and difficulty. However, at a more granular level, it is specifically optimized for sequence tagging tasks.

[e] Usually, the dataset includes the training set, the development set, and the test set.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Redundancy aims to uncover duplicate information within the dataset, which is crucial for ensuring consistency in results as it aids in securing an unbiased representation of data. By manually verifying Accuracy, we can assess the dataset's capability in accurately reflecting real-world information, a fundamental basis for reliable outcomes. The detection of the Leakage Ratio prevents the spillover of knowledge from test data to training data, essential for measuring the model's true performance. Together, these metrics form the cornerstone of dataset reliability, ensuring the effectiveness of NLP modeling.

Redundancy measures the proportion of duplicate instances in a dataset D. A lower Redundancy value is better as it indicates fewer duplicates and, therefore, a higher diversity in the data. It is calculated by dividing the number of instances appearing more than once by the dataset's total number of instances:

$$\mathrm{Red}(D) = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left[\left(x^{(i)}, y^{(i)}\right) = \left(x^{(j)}, y^{(j)}\right)\right]}{n} \tag{1}$$

In the case of **Accuracy**, a higher value is preferred because it reflects the proportion of correctly annotated instances, suggesting a more reliable dataset. Accuracy aims to evaluate the annotation correctness of the dataset and can be calculated as follows:

$$\delta\left(x^{(i)}, y^{(i)}\right) = \begin{cases} 1, & \text{if } y^{(i)} \text{ is accurate for } x^{(i)}, \\ 0, & \text{else} \end{cases} \tag{2a}$$

$$\mathrm{Acc}(D) = \frac{\sum_{i=1}^{n} \delta(x^{(i)}, y^{(i)})}{n} \tag{2b}$$

We recommend selecting 100 instances from each dataset split and inviting at least three professional linguists to annotate the Accuracy. To evaluate inter-rater reliability, we compute the Cohen Kappa coefficient (Cohen 1960) pairwise among three annotators, subsequently averaging these values. A mean Kappa exceeding 0.75 indicates substantial rater agreement, ensuring annotation reliability.

Leakage Ratio is a critical metric used to assess the extent of data leakage between different dataset partitions, specifically how many instances in the test set (Te) have incorrectly appeared in the training set (Tr) or development set (De). A lower Leakage Ratio is indicative of better dataset partitioning as it suggests that there is minimal to no overlap between the sets, which is essential for preventing models from merely memorizing specific instances instead of learning to generalize. The Leakage Ratio is defined as:

$$\mathrm{LeakR}(D) = \frac{\sum_{i=1}^{|Te|} \left[\left(Te^{(i)} \in Tr\right) \text{ or } \left(Te^{(i)} \in De\right)\right]}{|Te|} \tag{3}$$

4.2 Metrics under difficulty

We propose four metrics under difficulty to assess how challenging the datasets are, including three intrinsic metrics (**Unseen Entity Ratio**, **Entity Ambiguity Degree**, and **Text Complexity**) and one extrinsic metric (**Model Differentiation**). These difficulty metrics assess a dataset's challenge level for NLP models. Unseen Entity Ratio tests generalization by measuring novel entities, pushing models beyond their training. Entity Ambiguity Degree and Text Complexity challenge models with varied entity types and dense entity arrangements, requiring nuanced interpretation. Model Differentiation shows a dataset's power to separate model performances, testing robustness. Together, they define the dataset's challenge in terms of generalization, ambiguity, density, and differentiation, fitting the difficulty dimension.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

The **Unseen Entity Ratio** quantifies the proportion of new entities in the test set labels that are not present in the training set, promoting the model's ability to generalize. A higher Unseen Entity Ratio is desirable as it indicates a greater challenge for the model to recognize entities it has not encountered during training. The calculation is as follows:

$$\text{UnSeenEnR}(D) = \frac{|e\left(y_{Te}\right) \setminus e(y_{Tr})|}{|e(y_{Te})|} \tag{4}$$

Entity Ambiguity Degree is mainly used to measure how many entities are labeled with more than one kind of entities types. For example, if "apple" is labeled as "Fruit" in one instance and labeled as "Company" in another instance, then there is a conflict in D. A higher Entity Ambiguity Degree represents a more challenging dataset because it indicates more instances where an entity is labeled with different types, thereby confusing NER models. We introduce $e^*(D)$ to represent the number of conflict entities in dataset D and obtain the Entity Ambiguity Degree by:

$$\text{EnAmb}(D) = 1 - \frac{e^*(D)}{n} \tag{5}$$

Text Complexity measures the average Entity Density in sentences within the dataset. Higher Text Complexity signals a more difficult dataset because it implies that sentences are densely packed with entities, requiring more nuanced understanding and recognition by the model. It is formulated as:

$$\text{EnDen}(D) = \sum_{i=1}^{n} \frac{|e\left(y^{(i)}\right)|}{nm^{(i)}} \tag{6}$$

Model Differentiation evaluates the dataset's ability to distinguish the performance of different models. A higher Model Differentiation value is better as it indicates that the dataset can effectively reveal differences in model performances, making it a useful tool for benchmarking. It is determined using the standard deviation of the scores of k different models:

$$\text{ModDiff}(D) = \text{Std}(\theta_1, \theta_2, \cdots, \theta_k) \tag{7}$$

We recommend using the top five model scores on the dataset[f] for ModDiff calculation.

4.3 Metrics under validity

The metrics under validity, for example, **Entity Imbalance Degree** and **Entity-Null Rate** for NER datasets, are mainly proposed to evaluate the effectiveness of the dataset in evaluating the model's ability on the specific task. Validity metrics like Entity Imbalance Degree and Entity-Null Rate assess if a dataset can effectively evaluate a model's task-specific abilities. Entity Imbalance Degree checks for equal entity representation, ensuring models learn without bias—a key for valid evaluations. Entity-Null Rate measures how rich the dataset is in entity examples, vital for testing model learning depth. Both metrics directly contribute to assessing a dataset's ability to provide a fair and thorough evaluation of model performance, embodying the essence of validity.

Entity Imbalance Degree mainly measures the unevenness of the distribution of different entities in D. A lower Entity Imbalance Degree is better as it indicates a more balanced distribution of entity types, which is desirable for ensuring that the model is equally exposed to all categories and does not develop a bias toward the more frequent ones. Specifically, we use standard deviation to

[f]The Paperswithcode website regularly updates the scores of leading models on benchmark NER datasets. We selected the top five performing models based on their evaluation scores available on the site.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 1. Statistical evaluation of ten NER datasets

	Reliability			Difficulty				Validity	
	Red ↓	Acc ↑	LeakR ↓	UnSeenEnR ↑	EnAmb ↑	EnDen ↑	ModDiff ↑	EnImBaD ↓	EnNullR ↓
CLUENER	**0.00**	0.86	**0.00**	0.37	0.80	0.26	**4.58**	0.04	0.00
OntoNotes4	0.02	0.98	0.04	0.47	**2.54**	**1.02**	–	0.13	0.46
MSRA	**0.00**	0.99	**0.00**	0.28	1.16	0.17	0.38	0.11	0.41
PeopleDaily	**0.00**	0.96	**0.00**	0.22	1.73	0.65	–	0.11	0.40
Resume	**0.00**	**1.00**	0.01	0.46	0.29	0.25	0.41	0.17	0.17
Weibo	0.05	0.98	0.17	0.56	0.92	0.55	0.90	0.27	0.44
WikiAnn	0.03	0.89	0.13	0.55	1.53	0.74	–	**0.02**	**0.00**
CoNLL03	0.05	0.96	0.03	0.46	0.35	0.28	0.24	0.06	0.20
WNUT16	0.01	0.97	**0.00**	**0.89**	0.65	0.51	2.87	0.08	0.56
OntoNotes5	0.01	0.91	0.03	0.28	0.76	0.36	0.64	0.06	0.55

–indicates that the dataset and evaluation model scores have not been found on the Paperswithcode website, so the model discrimination of this dataset cannot be calculated. The upper rows are Chinese NER datasets, and the lower rows are English NER datasets. ↑indicates that the larger the value, the better the quality of the dataset on this metric. ↓indicates that the lower the value, the better the quality of the dataset on this metric.

quantify the degree of dispersion of the distribution of all the different types of entities $\mathscr{C}$ in the dataset:[g]

$$\text{EnImBaD}(\mathscr{D}) = Std\left(P_{y_D}(c_1), P_{y_D}(c_2), \cdots, P_{y_D}(c_v)\right) \tag{8}$$

Entity-Null Rate evaluates the proportion of instances in the dataset that do not contain any entity. A lower Entity-Null Rate is preferred because it suggests that the dataset contains a richer set of examples for the model to learn from, with more instances that include entity information. The Entity-Null Rate is defined as:

$$\zeta(y^{(i)}) = \begin{cases} 1, & \text{if } y^{(i)} \text{ has no entity,} \\ 0, & \text{else} \end{cases} \tag{9a}$$

$$\text{EnNullR}(D) = \frac{\sum_{i=1}^{n} \zeta(y^{(i)})}{n} \tag{9b}$$

5. Statistical dataset evaluation for NER

To validate our statistical dataset evaluation methods, we assess the quality of ten widely used NER datasets, including three English NER datasets and seven Chinese NER datasets. The evaluation results for ten NER datasets are shown in Table 1. Figure 2 presents the evaluation results of WNUT16, CoNLL03, Resume, and MSRA under different dimensions and metrics.

5.1 Datasets

We provide the basic information about the datasets in Table 2.

[g]Each $c_j \in \mathscr{C}, j \in 1, 2, \ldots, v$ represents a specific entity type, and $P_{y_D}(c_j)$ denotes the probability that c_j appears within all output entities in dataset D. This probability is computed by dividing the frequency of c_j in y_D by the total entity count in y_D.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 2. Standard named entity recognition dataset statistics.

Dataset	Lang	#Tags	Source
CLUENER	Zh	10	THUCNEWS
OntoNotes 4	Zh	4	News, Broadcast etc.
MSRA	Zh	3	News
PeopleDaily	Zh	3	News
Resume	Zh	8	Sina Finance
Weibo	Zh	4	Sina microblog
WikiAnn	Zh	3	Wikipedia
CoNLL03	En	4	Reuters News
WNUT16	En	10	User-generated web text
OntoNotes 5	En	18	Broadcast etc.

Zh and En mean Chinese and English, respectively. It is important to note that OntoNotes 4 has four common tags in the Chinese dataset, although OntoNotes 4 has a total of eighteen tags (for the English dataset).

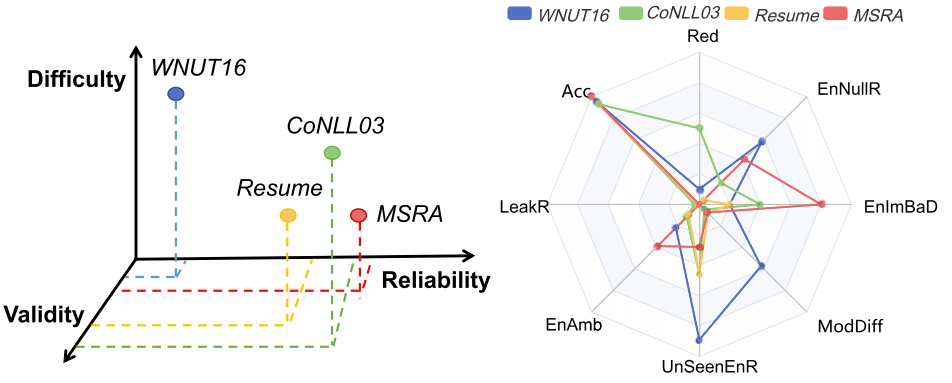

Figure 2. Evaluation results of WNUT16, CoNLL03, Resume, and MSRA under different dimensions and metrics. The abbreviations and corresponding full names of the metrics are presented in Sec. 4.

English NER datasets include the following: *CoNLL03 NER* (Sang and De Meulder 2003) is a classical NER evaluation dataset consisting of 1,393 English news articles. *WNUT16 NER* (Strauss *et al.* 2016) is provided by the second shared task at WNUT-2016 and consists of social media data from Twitter. *OntoNotes5* (Weischedel *et al.* 2013) is a multi-genre NER dataset collected from broadcast news, broadcast conversation, weblogs, and magazine genre, which is a widely cited English NER dataset.

Chinese NER datasets consist of the following: *CLUENER* (Xu *et al.* 2020), a well-defined NER dataset, includes finer-grained entity types beyond standard ones (person, organization, and location), such as Company, Game, and Book. *OntoNotes4* (Weischedel *et al.* 2011) is copyrighted by Linguistic Data Consortium[h] (LDC), a large manual annotated database containing various fields with structural information and shallow semantics. *MSRA* (Levow 2006) is a large NER dataset

[h]https://www.ldc.upenn.edu/

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

in the field of news, containing distinctive text structure characteristics. *PeopleDaily NER*[i] is a very classic benchmark dataset to evaluate different NER models. *Resume NER* (Zhang and Yang 2018) features resumes of senior executives from Chinese stock market companies, with a high annotator agreement of 97.1 percent. It includes 1027 randomly selected summaries annotated for 8 entity types using the YEDDA system (Yang *et al.* 2018). They randomly select 1027 resume summaries and manually annotate 8 types of named entities with YEDDA system (Yang *et al.* 2018). The inter-annotator agreement is 97.1 percent. *Weibo NER* (Peng and Dredze 2015; He and Sun 2017) is sourced from the Sina Weibo social media platform. *WikiAnn* (Pan *et al.* 2017) is a Chinese part of a multilingual NER dataset from Wikipedia articles.

5.2 Settings

According to the metrics we proposed in Sec. 4, we calculate the statistical scores for each dataset. Specifically, we average the scores of the training, the development, and the test split of the datasets for Redundancy, Accuracy, Entity Ambiguity Degree, Entity Density, Entity Imbalance Degree, and Entity-Null Rate, respectively. For Leakage Ratio, Unseen Entity Ratio, and Model Differentiation, we only calculate the scores on the specific splits involved according to Sec. 4.1 and Sec. 4.2.

5.3 Dataset reliability

5.3.1 Annotation Accuracy

Accuracy scores quantitatively inform us that we cannot take it for granted that all benchmark datasets are reliable.

We observe that *CLUENER* has the lowest Accuracy score. In particular, it has 0.17 (17 percent) errors in its development set (shown in Table 3). Conversely, the other datasets (e.g., *Resume* and *WNUT16*) have a relatively high Accuracy score for both Chinese and English NER datasets.

5.3.2 Leakage Ratio

The dataset's shortcomings (under the reliability dimension) can be effectively revealed by the Leakage Ratio. Given the Leakage Ratio results, we are surprised to find that *Weibo* and *WikiAnn* have serious data leakage issues.

As shown in Table 1 and Fig. 3, 0.17 (17 percent) and 0.13 (13 percent) of the instances in the test set of *Weibo* and *WikiAnn* have appeared in their corresponding training or development sets, respectively.

5.3.3 Overall reliability

Combining several metrics under the reliability dimension in Table 1, we can conclude that *Resume* and *MSRA* maintain high reliability.

In specific, there is no data redundancy in *Resume* and *MSRA*. That is to say, the instances of each part of the dataset are unique and non-repeating. Additionally, they achieve the highest Accuracy scores and hardly show data leakage problems, with a Leakage Ratio of 0.01 (1 percent) and 0.00 (0 percent), respectively.

5.4 Dataset difficulty

5.4.1 Unseen Entity Ratio

Results on Unseen Entity Ratio (UnSeenEnR) demonstrate the generalization ability of NER models on unseen entities.

The evaluation results show that *Weibo* and *WNUT16* are more difficult in terms of UnSeenEnR because their test sets have a 0.56 (56 percent) and a 0.89 (89 percent) ratio of entities

[i] https://github.com/zjy-ucas/ChineseNER

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 3. Results of metrics (except Leakage Ratio) under the reliability dimension of the NER datasets.

Dataset	Lang	Split	Red	Acc
CLUENER	Zh	train	0.00	0.89
		dev	0.00	0.83
OntoNotes4	Zh	train	0.02	0.98
		dev	0.03	0.97
		test	0.01	0.98
MSRA	Zh	train	0.00	0.99
		test	0.00	1.00
PeopleDaily	Zh	train	0.00	0.96
		dev	0.00	0.94
		test	0.00	0.97
Resume	Zh	train	0.00	1.00
		dev	0.00	1.00
		test	0.00	1.00
Weibo	Zh	train	0.08	0.96
		dev	0.03	0.98
		test	0.03	0.99
WikiAnn	Zh	train	0.04	0.90
		dev	0.03	0.88
		test	0.03	0.90
CoNLL03	En	train	0.06	0.93
		dev	0.03	0.96
		test	0.05	0.98
WNUT16	En	train	0.04	0.95
		dev	0.00	0.98
		test	0.00	0.97
OntoNotes5	En	train	0.01	0.90
		dev	0.01	0.88
		test	0.01	0.95

Red and Acc denote Redundancy and Accuracy, respectively. Zh and En mean Chinese and English, respectively.

that have not appeared in training, respectively. *WikiAnn* is the Chinese dataset only second to *Weibo* that can better evaluate the generalization ability of NER Models. Conversely, *PeopleDaily NER* and *OntoNotes5* are suboptimal for evaluating model generalization ability. Our experimental results in Sec. 6.5 reveals that model trained on them are more likely to perform better on seen entities compared to those that have not appeared in the training set.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

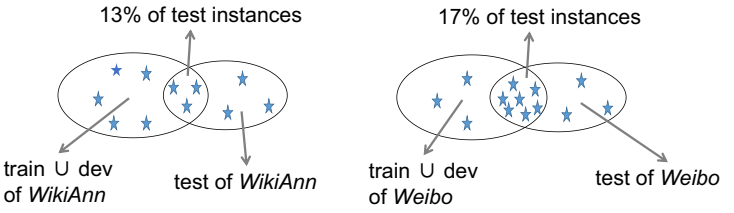

Figure 3. Leakage Ratio values of *WikiAnn* and *Weibo*. It is observed that 0.13 (13 percent) and 0.17 (17 percent) of the instances in the test set of *WikiAnn* and *Weibo*, respectively, appear in their corresponding training or development sets.

5.4.2 Entity Ambiguity Degree

Entity Ambiguity Degree (EnAmb) captures observable variation in the information complexity of datasets.

Given our findings, *OntoNotes 4* and *WNUT16* are the Chinese and English NER datasets with the highest Entity Ambiguity Degree, respectively, which means that they are more difficult for models to accurately predict entity types. Consistent with our conclusion (in Sec. 6.5), Bernier-Colborne and Langlais (2020) also argue that SOTA models cannot (or are not able) deal well with the entities labeled differently in different contexts.

5.4.3 Model Differentiation

Extrinsic evaluation metrics, such as Model Differentiation, are also necessary for evaluating the difficulty of datasets.

Unlike those intrinsic evaluation metrics (e.g., Entity Ambiguity Degree), **Model Discrimination** (ModDiff) aims to assess the dispersion of model scores on a unified benchmark dataset. That is to say, a more difficult dataset should have a clear distinction between models with different abilities. As shown in Table 1, *CLUENER* and *WNUT16* are Chinese and English datasets that can better distinguish model performance, respectively.

5.4.4 Overall difficulty

WNUT16 is a more difficult benchmark for English NER as a whole.

Although *WNUT16* has fewer citations than *CoNLL03* and *OntoNotes5*, as demonstrated in Table 1, *WNUT16* has a higher Entity Ambiguity Degree and Unseen Entity Ratio than the other two English NER datasets. Meanwhile, we find that the model performance gap on *WNUT16* is large, indicating that it is more difficult and can effectively distinguish models with different performances.

5.5 Dataset validity

5.5.1 Entity Imbalace Degree

Datasets with uneven distribution of entity types may not effectively evaluate the ability of models on the long-tailed instances.

Intuitively, the model does not perform as well on those long-tailed entity types as other entities. We observe that *Weibo* achieves the highest Entity Imbalance Degree (EnImBaD) by a large margin, indicating that its distribution of entity types is heavily uneven. Therefore, datasets with severely uneven distribution of entity types can only evaluate the performance of the models on a large number of distributed entity types.

5.5.2 Entity-Null Rate

Surprisingly, there are a large number of instances without any entities in many datasets such as *OntoNotes4*, *MSRA*, *WNUT16,* and *OntoNotes5*.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Although certain naturally distributed texts will contain some sentences without named entities, a high number of entity-free samples in a NER dataset makes it impossible to give a sufficient number of instances for NER model validation.

5.5.3 Overall validity

In general, *CoNLL03* is the English NER dataset with the highest validity. As shown in Table 1, *CoNLL03* has the lowest Entity-Null Rate (EnNullR), indicating that it can intensively test the entity recognition capabilities of NER models.

6. How do dataset properties affect model performance?

To validate the metrics and results under our statistical evaluation framework[j] and to further investigate how the statistical metric scores on dataset properties affect the model performance, we conduct controlled dataset adjustment in this section.

6.1 Models

For experiments on Chinese NER datasets, we use three models: 1) **Lattice-LSTM** (Zhang and Yang 2018), based on LSTM networks (Chiu and Nichols 2016), which automatically identifies key words from the context; 2) **Flat-Lattice** (Li *et al.* 2020), which converts the lattice structure into a flat structure; and 3) **Roberta** (Liu *et al.* 2019), a transformer-based pretrained model which removes the next sentence predict task in BERT.

For the English datasets, we also take three models, including: 1) **LSTM CRF** (Lample *et al.* 2016), a traditional model based on the bidirectional LSTM with conditional random fields (CRF); 2) **LUKE** (Yamada *et al.* 2020), which provides new pretrained contextualized representations of words and entities by predicting masked words and entities in entity-annotated corpus based on the bidirectional transformer (Vaswani *et al.* 2017); and 3) **W2NER** (Li *et al.* 2020), which converts NER to word–word relationship classification and models the neighboring relations between entity words with Next-Neighboring-Word (NNW) and Tail-Head-Word (THW) relations.

6.2 Experiment settings

All the experiments are done on the NVIDIA RTX 2080 GPU and 3090 GPU and evaluated by seqeval.[k] Specifically, we utilize Micro F1 scores to measure the performance of the NER model. For the experiment with Train-Dev Dataset Adjustment (Sec. 6.4), we report the averaged results and variances over three random seeds.

6.2.1 Hyperparameters

In our research, we concentrated on refining model parameters and embedding techniques to boost performance. We chose a non-BERT variant of the Flat-Lattice model, which we enhanced with a CRF layer on Roberta. We also utilized the most effective version of LSTM CRF, notable for its use of pretrained word embeddings, character-level word modeling, and an optimized dropout rate.

Consistent with observations by Lample *et al.* (2016), we found that models using pretrained word embeddings typically surpass those with randomly initialized embeddings. Thus, we experimented with various word embedding methods to cover a broad range of approaches, as elaborated in 6.2.2.

[j] We provide additional validation details of the metrics within our statistical evaluation framework in the supplementary appendix document.

[k] https://github.com/chakki-works/seqeval

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 4. Chinese NER model replication results.

	Lattice-LSTM		Flat-Lattice		Roberta	
	ori.	repro.	ori.	repro.	ori.	repro.
MSRA	93.18	93.12	94.35	94.06	–	94.57
OntoNotes4	73.88	73.43	75.70	75.84	–	80.30
Resume	94.46	94.46	94.93	95.11	–	96.19
Weibo	58.79	56.49	63.42	57.92	–	67.92

repro. denotes reproduction. - denotes that the authors of the literature we cited did not experiment on that dataset. And ori. denotes original paper results.

Our study utilized various optimization algorithms. For instance, AdamW optimizer (Loshchilov and Hutter 2017) was used for models like W2NER, Roberta, and LUKE. In contrast, models such as Lattice-LSTM, LSTM CRF, and Flat-Lattice were fine-tuned using stochastic gradient descent (SGD). Notably, both LUKE and W2NER models were further improved by combining AdamW with a learning rate warmup and linear decay strategy. LUKE also incorporated early stopping based on the development set performance. The specific hyperparameters for these models can be found in Appendix.

6.2.2 Word embeddings

- **Static Word Embeddings**: In the realm of static word embeddings, Lattice-LSTM utilizes its unique word,[l] character, and character bigram embeddings.[m] However, since LSTM CRF's own pretrained embedding was unavailable, we opted for common-crawl vectors from FastText.[n] Similarly, Flat-Lattice employed the same pretrained embeddings as Lattice-LSTM.

- **Dynamic Word Embeddings**: Dynamic word embeddings represent a significant advancement over static embeddings, as they are context-sensitive and capable of capturing varying meanings of words in different contexts. Our approach prominently featured BERT-based embeddings, known for their extensive integration of grammatical, lexical, and semantic information. LUKE, for instance, introduced new pretrained contextualized representations of words and entities using Roberta. W2NER used bert-large-cased for English datasets and bert-base-chinese for Chinese datasets, taking advantage of Roberta's refined capabilities as an optimized version of BERT.

6.3 Model replication results

We replicated six NER models in accordance with the experimental setup, and the results of the model replication are presented in Table 4 and 5.

6.4 Controlled dataset adjustment

To investigate how statistical properties affect model performance, we conducted controlled dataset adjustments: 1) we modified the test set to create two new sets (of the same size) with distinct statistical values for specific metrics (i.e., Test Dataset Adjustment). 2) Similarly, we

[l] https://github.com/jiesutd/RichWordSegmentor
[m] https://github.com/jiesutd/LatticeLSTM
[n] https://fasttext.cc/docs/en/english-vectors.html

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 5. English NER model replication results.

	LSTM CRF		W2NER		LUKE	
	ori.	repro.	ori.	repro.	ori.	repro.
CoNLL03	83.63	83.61	93.07	92.02	94.30	94.2
WNUT16	–	26.04	–	45.81	–	56.99
OntoNotes5	–	80.14	90.50	84.92	–	87.27

repro. denotes reproduction. - denotes that the authors of the literature we cited did not experiment on that dataset. And ori. denotes original paper results.

adjusted the training and development sets to form new sets (of the same size) with distinguishable metrics values (i.e., Train-Dev Dataset Adjustment).

6.4.1 Test Dataset Adjustment

We adjusted the test set for three metrics: Leakage Ratio, Unseen Entity Ratio, and Entity Ambiguity Degree. This led to two new test sets with distinct statistical values for these metrics. For example, as for the Unseen Entity Ratio, we adjusted the test set to construct two new test sets, one with an Unseen Entity Ratio of 0.80 (80 percent) and the other with an Unseen Entity Ratio of 0.20 (20 percent), while ensuring that the two newly constructed test sets have the same number of instances.

6.4.2 Train-Dev Dataset Adjustment

Initially, we chose datasets with a high Entity-Null Rate (*WNUT16*, *OntoNotes5* for English; *Weibo*, *OntoNotes4* for Chinese). We then filtered the training and development sets to adjust the Entity-Null Rate to 0.20 (20 percent) and 0.80 (80 percent), ensuring equal numbers of instances in these subsets. Finally, we trained the data with various models before testing and comparing the results with the same test set.

6.5 Experiment results and analysis

- **Datasets with high Unseen Entity Ratio are more difficult for NER models:** Intuitively, those entities that were seen during training are less challenging for NER models compared to those that did not appear in the training set. Figure 4 supports our intuition. Models perform better on datasets with a lower proportion of unseen entities than on datasets with a relatively high proportion of unseen entities.

- **Entities with strong Entity Ambiguity Degree are indeed more likely to confuse the model:** We can infer from Figure 5 that datasets with a high Entity Ambiguity Degree are more challenging for the model. As for models tested on Chinese datasets, their average performance is 6.42 (F1) points higher on datasets with low entity ambiguity rates than on datasets with high entity ambiguity rates. The English NER model is more likely to be confused by entities with a high entity ambiguity rate and make wrong decisions.

- **The models exhibit improved performance with increased test set leakage, highlighting the necessity for enhanced generalization in NER models:** As shown in Table 6, three models (i.e., Lattice-LSTM, Flat-Lattice, and Roberta) consistently achieve better performance when the leakage rate of the test set is 0.80 (80 percent) than when it is 0.20

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

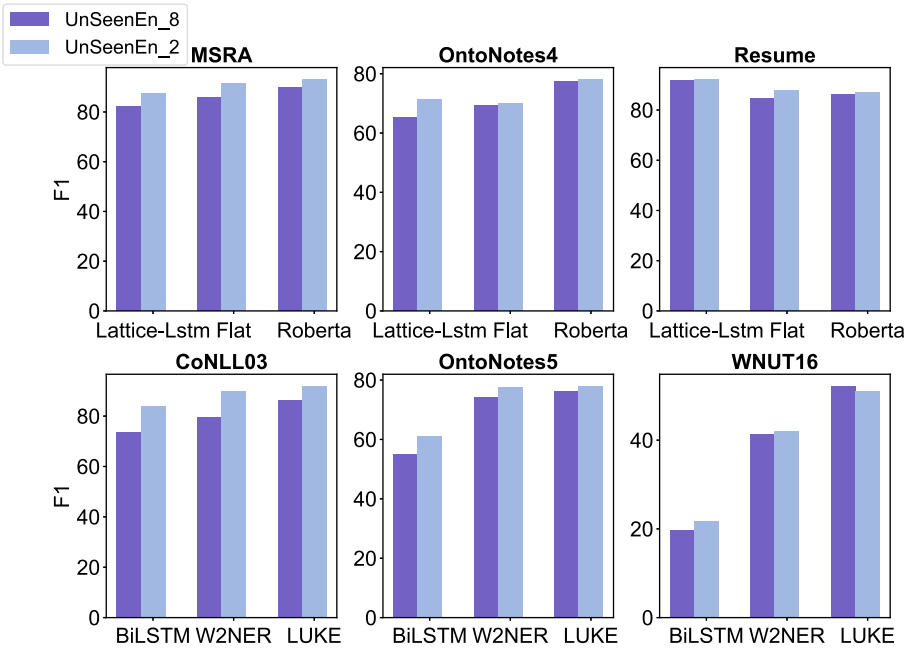

Figure 4. Model performance on NER datasets when the proportion of unseen entities (UnSeenEn) in the test set is 0.80 (UnSeenEn_8) and 0.20 (UnSeenEn_2), respectively.

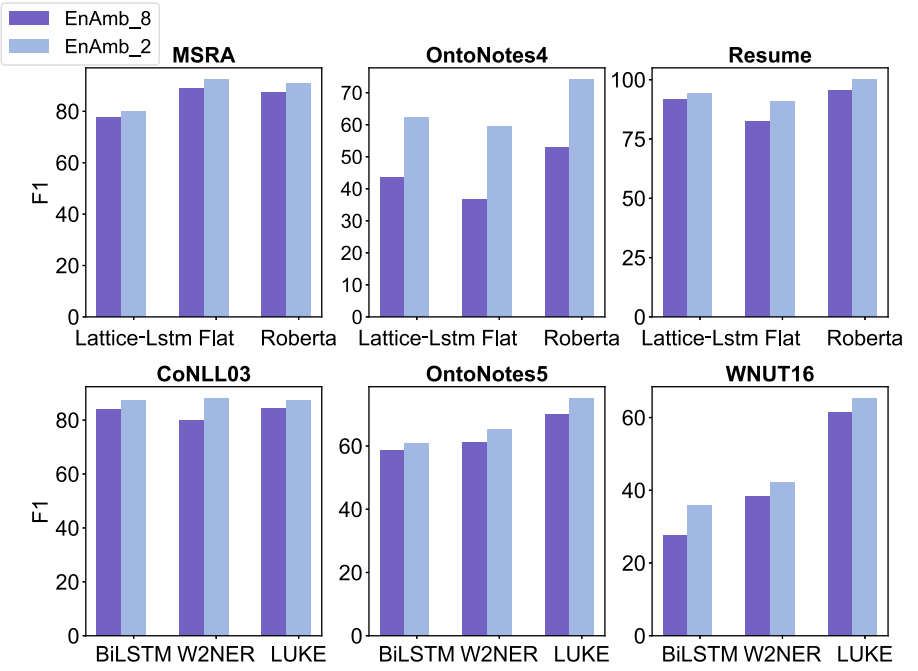

Figure 5. Model performance on NER datasets when the proportion of ambiguous entities in the test set is 0.80 (EnAmb_8) and 0.20 (EnAmb_2), respectively.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 6. Model performance when the proportion of leaked samples in the test set is 80 percent and 20 percent, respectively.

Dataset	Test	LSTM	Flat-Lattice	Roberta
Weibo	Leakage (80%)	**73.68**	**74.24**	**78.52**
	Leakage (20%)	64.86	48.65	70.40

LSTM represents Lattice-LSTM.

Table 7. Model performance on English datasets when the proportion of samples without entities in the training set and development set is 0.80 (80 percent), 0.20 (20 percent), 0.00 (0 percent), and original, respectively.

		W2NER		LSTM CRF	
Dataset	Train & Dev	avg.	std.	avg.	std.
OntoNotes5	EnNullR (0.80)	68.99	0.6663	64.80	0.0031
	EnNullR (0.20)	77.96	0.0134	75.01	0.0354
	EnNullR (0.00)	76.41	0.2037	74.75	0.0440
	Original	86.99	0.0672	80.91	9.8596
WNUT16	EnNullR (0.80)	49.66	1.3646	25.26	0.1722
	EnNullR (0.20)	54.22	0.5449	36.99	0.9092
	EnNullR (0.00)	52.36	0.3577	36.04	1.3302
	Original	55.49	2.1829	36.89	0.0900

(20 percent). In particular, we found that the performance of Flat-Lattice on the Weibo test set with a Leakage Ratio of 0.80 (80 percent) outperformed the 0.20 (20 percent) by a large margin, that is, 25.69 percent. We speculate that because the model has seen the leaked data in the test set during training, it performs better on the test set with a relatively high data leakage rate. Looking at the experimental results from another perspective, researchers need to pay more attention to improving the NER model's generalization ability.

- **Entity-Null Rate plays a small difference:** As shown in Tables 7 and 8, the F1 score of the training set and development set with EnNullR of 0.20 (20 percent) is better than 0.80 (80 percent). Therefore, we conclude that the contribution of instances without entities to the model is less than the instances with entities during training. However, are instances without any entities completely useless for model training? We delete all these instances and show the results in Tables 7 and 8. The performance of models trained on such datasets decreases, which indicates that the instances without entity are necessary, as they keep the distribution of the test set and training set relatively consistent.

7. Discussion

Our statistical evaluation framework can be used to analyze the factors that affect the dataset's quality and, furthermore, to build a higher-quality dataset in a targeted manner or augment the data with statistical improvement guidance. In this section, we take an initial step to analyze how the dataset construction process affects the statistical properties of datasets.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Table 8. Model performance on Chinese datasets when the proportion of samples without entities in the training set and development set is 0.80 (80 percent), 0.20 (20 percent), 0.00 (0 percent), and original, respectively.

		Lattice-LSTM		Flat-Lattice	
Dataset	Train & Dev	avg.	std.	avg.	std.
Weibo	EnNullR (0.80)	29.26	5.0456	30.96	0.3409
	EnNullR (0.20)	52.97	0.0408	53.55	2.7369
	EnNullR (0.00)	54.02	0.3600	55.89	6.2324
	Original	55.04	1.0192	57.92	–
MSRA	EnNullR (0.80)	80.51	0.0750	83.26	0.0157
	EnNullR (0.20)	91.11	0.0151	92.82	0.0259
	EnNullR(0.00)	91.94	0.0001	93.60	0.0097
	Original	92.50	0.0273	94.06	–

Table 9. Standard named entity recognition dataset construction method.

Dataset	Construction method
CLUENER	Distant supervision + human
OntoNotes4	Human annotation
MSRA	Human annotation
PeopleDaily	Human annotation
Resume	Human annotation
Weibo	Human annotation
WikiAnn	Cross-lingual name tagging framework
CoNLL03	Human annotation
WNUT16	Human annotation
OntoNotes5	Human annotation

As shown in Table 9, based on an overview of the literature that presented the ten NER datasets, we provide a summary of how they were built. We can see that all datasets were created manually, with the exception of *CLUENER* and *WikiAnn*. As for *CLUENER*, Xu *et al.* (2020) prelabel their dataset using the distant-supervised approach with a vocabulary and then manually check and modify some labels. *WikiAnn* is constructed using a cross-lingual name tagging framework based on a series of new Knowledge Base (KB) mining methods (Pan *et al.* 2017).

We observe from Table 1 that only two of the ten NER datasets, *CLUENER* and *WikiAnn*, had Acc scores below 0.90 (90 percent), indicating that the NER dataset, which was not totally created manually, will have a significant number of annotation errors (shown in Figure 6).

8. Conclusion and future work

In this paper, we investigate various statistical properties of the NER datasets and propose a comprehensive dataset evaluation framework with nine statistical metrics. We implement a fine-grained evaluation of ten widely used NER datasets and provide a fair comparison of the

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

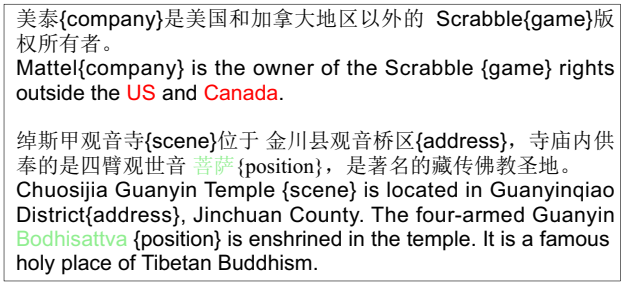

Figure 6. Mislabeled Examples of randomly selected samples from *CLUNER*. Red indicates missing entities not assigned entity labels. Green indicates the entity with the wrong labeled entity type.

existing datasets from three dimensions: reliability, difficulty, and validity. We further explore how the statistical properties of the training dataset influence the model performance and how dataset construction methods affect the dataset quality. In the future, we hope more works dive into dataset quality evaluation from a broader and more general perspective.

Competing interests. The authors declare no other competing interest.

References

Bernier-Colborne G. and Langlais P. (2020). Hardeval: Focusing on challenging tokens to assess robustness of ner. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1704–1711.

Blevins T. and Zettlemoyer L. (2020). Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1006–1017.

Bommasani R., Hudson D. A., Adeli E., Altman R., Arora S., von Arx S. and Liang P. (2021). On the opportunities and risks of foundation models, arXiv preprint arXiv: 2108.07258.

Chiu J. P. and Nichols E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* **4**, 357–370.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46.

Fu J., Liu P. and Neubig G. (2020). Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6058–6069.

Goyal T., Li J. J. and Durrett G. (2022). News summarization and evaluation in the era of gpt-3, arXiv preprint arXiv: 2209.12356.

Gupta S. and Gupta A. (2019). Dealing with noise problem in machine learning data-sets: a systematic review. *Procedia Computer Science* **161**, 466–474.

Gururangan S., Swayamdipta S., Levy O., Schwartz R., Bowman S. and Smith N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, (Short Papers)

He H. and Sun X. (2017). F-Score driven max margin neural network for named entity recognition in Chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 713–718.

Hermann K. M., Kocisky T., Grefenstette E., Espeholt L., Kay W., Suleyman M. and Blunsom P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15)*. MIT Press, Cambridge, MA, USA, pp. 1693–1701.

Jariwala A., Chaudhari A., Bhatt C. and Le D. N. (2022). Data quality for AI tool: exploratory data analysis on IBM API. *International Journal of Intelligent Systems and Applications* **14**(1), 42–56.

Kaushik D. and Lipton Z. C. (2018). How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5010–5015.

Kenton J. D. M. W. C. and Toutanova L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186.

Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270.

Larson S., Lim G. and Leach K. (2023). On evaluation of document classification with rvl-cdip. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2665–2678.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Levow G. A. (2006). The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN workshop on Chinese language processing*, pp. 108–117.

Lewis P., **Stenetorp P. and Riedel S.** (2021). Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1000–1008.

Li X., **Yan H.**, **Qiu X. and Huang X. J.** (2020). FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6836–6842.

Liu Y., **Ott M.**, **Goyal N.**, **Du J.**, **Joshi M.**, **Chen D.**,. . ., **Stoyanov V.** (2019). Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv: 1907.

Loshchilov I. and Hutter F. (2017). Decoupled weight decay regularization, arXiv preprint arXiv: 1711.05101.

Malmasi S., **Fang A.**, **Fetahu B.**, **Kar S. and Rokhlenko O.** (2022). MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3798–3809.

Nadeem M., **Bethke A. and Reddy S.** (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371.

Nallapati R., **Zhou B.**, **dos Santos C.**, **Gulçehre Ç. and Xiang B.** (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290.

Narayan S., **Cohen S. B. and Lapata M.** (2018). Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807.

Ng A., **Laird D. and He L.** (2021). *Data-Centric Ai Competition*. DeepLearning AI. Available at https://https-deeplearning-ai.github.io/data-centric-comp/.

Novick M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* **3**(1), 1–18.

Pan X., **Zhang B.**, **May J.**, **Nothman J.**, **Knight K. and Ji H.** (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, (Volume 1: Long Papers)*, pp. 1946–1958.

Peng N. and Dredze M. (2015). Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 548–554.

Poliak A., **Naradowsky J.**, **Haldar A.**, **Rudinger R. and Van Durme B.** (2018). Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191.

Rashkin H., **Nikolaev V.**, **Lamm M.**, **Aroyo L.**, **Collins M.**, **Das D. and Reitter D.** (2023). Measuring attribution in natural language generation models. *Computational Linguistics*, pp. 1–64.

Rush A. M., **Chopra S. and Weston J.** (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389.

Sang E. T. K. and De Meulder F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147.

Srivastava A., **Rastogi A.**, **Rao A.**, **Shoeb A. A. M.**, **Abid A.**, **Fisch A. and Wang G.** (2022). Beyond the imitation game: quantifying and extrapolating the capabilities of language models, arXiv preprint arXiv: 2206.04615.

Strauss B., **Toma B.**, **Ritter A.**, **De Marneffe M. C. and Xu W.** (2016). Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 138–144.

Sugawara S., **Stenetorp P.**, **Inui K. and Aizawa A.** (2020). Assessing the benchmarking capacity of machine reading comprehension datasets. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(05), 8918–8927.

Tejaswin P., **Naik D. and Liu P.** (2021). How well do you know your summarization datasets?. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3436–3449.

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A. N.**,. . ., **and Polosukhin I.** (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.

Wang A., **Pruksachatkun Y.**, **Nangia N.**, **Singh A.**, **Michael J.**, **Hill F. and Bowman S.** (2019). Superglue: a stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pp. 32.

Wang B., **Wang S.**, **Cheng Y.**, **Gan Z.**, **Jia R.**, **Li B. and Liu J.** (2020). Infobert: improving robustness of language models from an information theoretic perspective, arXiv preprint arXiv: 2010.02329.

Wang Z., **Shang J.**, **Liu L.**, **Lu L.**, **Liu J. and Han J.** (2019). CrossWeigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5154–5163.

Weischedel R., **Palmer M.**, **Mitchell M.**, **Hovy E.**, **Pradhan S.**, **Ramshaw L.**,. . ., **and Houston A.** (2013). *OntoNotes release 5.0 LDC2013T19*. Linguistic Data Consortium.

Weischedel R., **Pradhan S.**, **Ramshaw L.**, **Palmer M.**, **Xue N.**, **Marcus M. and Houston A.** (2011). *Ontonotes release 4.0. LDC2011T03*. Philadelphia, Penn: Linguistic Data Consortium.

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Xu L., Dong Q., Liao Y., Yu C., Tian Y., Liu W. and Zhang X. (2020). CLUENER2020: fine-grained named entity recognition dataset and benchmark for chinese, arXiv preprint arXiv: 2001.04351.

Xu L., Hu H., Zhang X., Li L., Cao C., Li Y. and Lan Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4762–4772.

Xu L., Liu J., Pan X., Lu X. and Hou X. (2021). Dataclue: a benchmark suite for data-centric nlp, arXiv preprint arXiv: 2111.08647.

Yamada I., Asai A., Shindo H., Takeda H. and Matsumoto Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454.

Yang J., Zhang Y., Li L. and Li X. (2018). YEDDA: A lightweight collaborative text span annotation tool. In *Proceedings of ACL 2018, System Demonstrations*, pp. 31–36.

Yin W., Radev D. and Xiong C. (2021). DocNLI: a large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4913–4922.

Zeng Q., Yu M., Yu W., Jiang T. and Jiang M. (2021). Validating label consistency in NER Data annotation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pp. 11–15.

Zha D., Bhat Z. P., Lai K., Yang F., Jiang Z., Zhong S. and Hu X. (2023). Data-centric artificial intelligence: a survey, arXiv preprint arXiv: 2303.10158.

Zhang Y., Kang B., Hooi B., Yan S. and Feng J. (2023). Deep long-tailed learning: a survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Zhang Y. and Yang J. (2018). Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1554–1564.

Zhu X., Wu X. and Chen Q. (2003). Eliminating class noise in large datasets. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 920–927.

Appendix A. Validation details of the metrics under our statistical evaluation framework

We justify and clarify those metrics under our evaluation framework that we have not discussed further in the main text.

- **Redundancy:** International data standards[o] demand that data be unique. NLP dataset is a particular data type that must adhere to the same criteria as other data types.

- **Accuracy:** Numerous research have demonstrated that flaws in datasets will negatively impact the model's performance (Zhu *et al.* 2003; Tejaswin *et al.* 2021; Gupta and Gupta 2019). The model's performance will increase to some extent after these mistakes are fixed (Zeng *et al.* 2021).

- **Text Complexity:** Several experiments of Fu *et al.* (2020) on English NER datasets supported our use of Entity Density as a valid metric of the difficulty of the dataset. Their experiments showed that NER models are negatively correlated with Entity Density.

- **Model Differentiation:** This extrinsic metric aims to assess the dispersion of model scores on a unified benchmark dataset. As long as enough models are evaluated on the dataset, we can measure the differentiation of a dataset by calculating the dispersion of the scores of different models.

- **Entity Imbalance Degree:** There are category imbalances in many NLP tasks that can seriously affect the model's performance on the long-tail instances (Blevins and Zettlemoyer 2020; Zhang *et al.* 2023; Wang *et al.* 2020). Therefore, the Entity Imbalance Degree of the NER dataset is necessary and practical.

[o]https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

https://doi.org/10.1017/nlp.2024.37 Published online by Cambridge University Press

Appendix B. Details of manually checking data accuracy

We recommend selecting 100 instances from each dataset split and inviting at least three professional linguists who are volunteers to annotate the accuracy. Before the formal work, we conducted face-to-face training, such as introducing the standards of data proofreading.

Appendix C. Specific hyperparameters for our selected evaluation models

Table 10. Hyperparameter settings for various NER models.

W2NER	LSTM CRF	Flat-Lattice
dist_emb_size: 20	gradient clipping: 5.0	decay: −0.05
type_emb_size: 20	layer dimension: 100	momentum: −0.9
bert_hid_size: [768, 1024]	LSTM layer: 1	FFN_size: 480
conv_hid_size: [96, 64]	dropout: 0.5	head: [8, 4, 12]
lstm_hid_size: [768, 512]	char_dim: 25	d_head: [16, 20]
dropout: 0.5	char_lstm_dim: 25	d_model: head × d_head
learning rate (BERT): [1e-5, 5e-6]	word_dim: 300	embed dropout: 0.5
learning rate (others): 1e-3	word_lstm_dim: 100	output dropout: 0.3
batch size: [2, 4, 8]	learning rate: 0.01	learning rate: [1e-3, 8e-4]
-	-	warmup: [10, 1, 5] epoch
-	-	batch size: [10, 8]
Lattice-LSTM	**Roberta**	**LUKE**
embedding size: 50	batch size: 32	batch size: [4, 8]
LSTM hidden: 200	max sentence length: 300	adam β_1: 0.9
batch size: 1	weight decay rate: 0.1	adam β_2: 0.98
learning rate: 0.015	warmup: 100(step)	adam ϵ: 1e-6
dropout: 0.5	-	dropout: 0.1
learning rate: 5e-5	-	warmup ratio: 0.06
-	-	weight decay: 0.01
-	-	maximum word length: 512
-	-	learning rate: 1e-5
-	-	gradient clipping: none

Cite this article: Wang C, Dong Q, Wang X and Sui Z. Statistical dataset evaluation: A case study on named entity recognition. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.37