MELD: Melding Diverse Multilingual and Multi-Domain Datasets for Named Entity Recognition Evaluation

Kevin Glocker, Marco Kuhlmann
Department of Computer and Information Science
Linköping University
kevin.glocker@liu.se, marco.kuhlmann@liu.se

Abstract

Zero-shot Named Entity Recognition (NER) has gained prominence for information extraction across diverse domains without being limited to a single, fixed tag set. However, existing NER resources vary widely in data format, licensing terms, annotation schemes, and availability, making it difficult to systematically evaluate the generalization capabilities of zero-shot NER models. Prior attempts to aggregate datasets with broad coverage across domains have largely focused on a small subset of languages, and it is often not transparent how datasets were processed from their sources. This paper introduces MELD, a comprehensive multilingual and multi-domain data collection designed to address these gaps. MELD integrates 60 NER datasets spanning 194 languages, 14 domains, and 601 normalized entity types. While previously introduced multilingual NER datasets are mainly silver-standard, MELD contains gold-standard annotations for 60 languages. All data processing steps are fully open-source and reproducible, facilitating future extensions and ensuring long-term accessibility. While MELD is primarily designed for zero-shot evaluation, it also provides training and development splits in a single, consistent format to support future research in few-shot and supervised NER settings.

Keywords: named entity recognition, benchmark, multilingual, zero-shot

1. Introduction

Named Entity Recognition (NER) is a foundational NLP task that supports a wide range of applications, including information extraction, search, and knowledge graph construction. Traditionally, NER models have been trained on manually annotated datasets with a fixed inventory of entity categories, including general types such as “person” or “location” (Grishman and Sundheim, 1995) and specialized types tailored to specific domains and tasks. While effective, this reliance on pre-specified labels has limited the adaptability of conventional NER systems.

The rise of large language models (LLMs) has led to increased interest in *zero-shot NER*, where a system is instructed to recognize entity types specified by the user at inference time. This paradigm can be approached by prompting general instruction-tuned LLMs (Wang et al., 2025), or with dedicated models trained on synthetic data (Zaratiana et al., 2024) or heterogeneous collections of multi-domain datasets (Yang et al., 2025). By removing the need to manually annotate new datasets for each task, zero-shot NER enables broader applicability and lowers the cost of deployment in new domains. However, no single dataset or benchmark currently covers the range of domains, languages, and annotation conventions needed to systematically evaluate zero-shot NER performance across languages and domains.

To address this gap, we introduce a new collection of *multilingual, entity-labeled datasets (MELD)*.

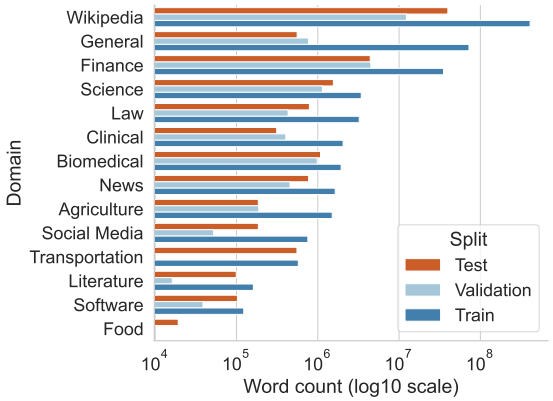

Figure 1: Domain distribution of the gold-standard part of MELD (language distribution in Figure 4).

To our knowledge, MELD is the largest data collection for NER to date, including gold-standard test sets for 60 languages across 14 domains (see Figure 1) alongside widely-used silver-standard datasets for additional multilingual coverage to a total of 194 languages. We consider datasets as “silver-standard” if they were automatically extracted without human verification. Most current multilingual datasets fall into this category, while the availability of gold-standard data is limited to a small number of high-resource languages.[1]

In contrast to most previous large-scale datasets and benchmarks collected from multiple datasets,

[1] A notable exception is MasakhaNER for African languages (Adelani et al., 2022).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1889–1903
11–16 May 2026.

MELD is fully end-to-end reproducible from published sources. Furthermore, data processing is implemented as a set of modular components designed to facilitate future expansions and modifications. MELD is open-source and available from: https://github.com/kgnlp/meld.

2. Related Work

NER is a quintessential NLP task, and a large number of datasets for it have been released in the past decades. Where most of these target single domains and languages, we focus on NER for both highly multilingual and multi-domain applications.

Multilingual Datasets The most widely used dataset for large-scale multilingual evaluation of NER models is WikiANN (Pan et al., 2017), which (in its original version) provides annotations for four general entity types across 282 languages. These annotations have been automatically extracted using Wikipedia anchor links and automatic tagging to identify unlinked entities (Pan et al., 2017). Although WikiANN offers broad language coverage, its use for training and evaluation has been criticized for quality issues such as label noise and the lack of involvement of native speakers for most of the languages represented (Lignos et al., 2022).

More recent multilingual dataset construction efforts include WikiNEuRal (Tedeschi et al., 2021) and the MultiCoNER v2 (Fetahu et al., 2023a,b) benchmark. Like WikiANN, they are silver-standard datasets extracted from Wikipedia. However, they were developed for a smaller set of languages, and utilize improved methods to handle the issues of unlinked entities and entity type assignment. Furthermore, MultiCoNER v2 includes annotations for a broader range of 32 fine-grained entity types, which makes it more suitable for evaluating zero-shot NER models.

A key limitation of Wikipedia-derived NER datasets is the lack of diversity in domains and text types. While Wikipedia articles cover a diverse set of topics, evaluation results may not necessarily be representative of performance on, e.g., social media or legal texts (Tatariya et al., 2024). Furthermore, the source articles for low-resource languages can themselves be of low quality as they are often written by non-native speakers (Tatariya et al., 2024).

Addressing the limitations of silver-standard evaluation, Mayhew et al. (2024) proposed Universal NER, a gold-standard dataset consisting primarily of manually annotated portions of Universal Dependency corpora, focusing on a coarse-grained entity tag set with three types. While this allows for consistent multilingual evaluation, it is not sufficient on its own to estimate multi-domain performance, as the Universal Dependencies corpora source text from a limited set of domains. Furthermore, the coarse label inventory cannot approximate the fine-grained distinctions that zero-shot NER systems are expected to make when users specify novel entity types at inference time.

The most recent effort toward constructing multilingual datasets is OpenNER (Palen-Michel et al., 2025), which compiles 36 gold-standard NER datasets covering 52 languages. However, OpenNER does not include nested entity annotations. Furthermore, it focuses on datasets that are available in a pre-tokenized format and center around traditional named entity types, which restricts its usefulness for the evaluation of zero-shot NER.

Multi-Domain Datasets To evaluate multi-domain generalization in English, Liu et al. (2021) introduced CrossNER to measure transfer of NER models to six domains. More recently, Zhou et al. (2024) assemble a large-scale benchmark encompassing 43 datasets across nine domains. However, their data and preprocessing code are not open-source, making it difficult to reproduce. Yang et al. (2025) introduce B2NERd, a collection of 54 datasets designed for training open-domain NER models, covering diverse domains in English and Chinese. Additionally, they propose a shared entity type taxonomy which resolves inconsistencies between different definitions of the same entity type across datasets. They show that this disambiguation leads to substantially improved performance of models trained on the data collection. While B2NERd covers a diverse set of domains in two high-resource languages, additional work is required for multilingual evaluation.

3. MELD

In this section, we describe the data selection process for MELD, our methodology for processing constituent datasets into a standardized format, and our approach for normalizing entity types to facilitate zero-shot NER evaluation.

3.1. Data Selection

Since we aim to cover a broad spectrum of languages and domains, we selected a diverse set of upstream datasets annotated with named entities, terms and concepts. An overview of these sources is given in Table 1.

3.1.1. Exclusion Criteria

To ensure that MELD is reproducible and simple to use, we excluded data sets that cannot

Domain	Datasets	
Agriculture	AgCNER (Yao et al., 2024)	AgriNER (De et al., 2023)
Biomedical	AnatEM (Pyysalo and Ananiadou, 2013) BC4CHEMD (Krallinger et al., 2015) BioRED (Luo et al., 2022) NCBI-Disease (Doğan et al., 2014)	BC2GM (Smith et al., 2008) BC5CDR (Li et al., 2016) JNLPBA (Collier et al., 2004)
Clinical	CANTEMIST (Miranda-Escalada et al., 2020) RaTE-NER[2l] (Zhao et al., 2024)	EBM-NLP (Nye et al., 2018)
Finance	FiNER-139 (Loukas et al., 2022)	
Food	TASTEset (Lawrynowicz et al., 2023)	
General	Arabic-Cross-Dialectal-NER (Elkhbir et al., 2023) Naamapadam[2l] (Mhaske et al., 2023) pioNER[2l] (Ghukasyan et al., 2018) Turku-NER-corpus (Luoma et al., 2020) CrossNER (Liu et al., 2021)	UniversalNER (Plank et al., 2020; Jørgensen et al., 2020; Mayhew et al., 2024) NYTK-NerKor (Simon and Vadász, 2021) Thai-NER (Phatthiyaphaibun, 2024) TurkuONE (Luoma et al., 2021)
Law	E-NER (Au et al., 2022) LegalNERo (Pais et al., 2021)	German-LER (Leitner et al., 2019)
Literature	Herodotos-Project-NER (Erdmann et al., 2019)	
News	CLEANANERCorp (Mashael AlDuwais and Al-Salman, 2024) EverestNER (Niraula and Chapagain, 2022) FoNE (Snæbjarnarson et al., 2023) CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003)	MasakhaNER-X (Ruder et al., 2023; Adelani et al., 2022, 2021) FiNER-ORD (Shah et al., 2024) idner-news-2k (Khairunnisa et al., 2020) PhoNER-COVID19 (Truong et al., 2021)
Science	FabNER (Kumar and Starly, 2021) SCIERC (Luan et al., 2018) SOFC-Exp (Friedrich et al., 2020) WIESP2022 (Grezes et al., 2022)	SciER (Zhang et al., 2024) SciREX (Jain et al., 2020) SoMeSci (Schindler et al., 2024, 2021) WLP (Kulkarni et al., 2018)
Social Media	DanfeNER (Niraula and Chapagain, 2023) MIT-Movie (Liu et al., 2013) Tweebank-NER (Jiang et al., 2022) Weibo-NER (He and Sun, 2017; Peng and Dredze, 2016, 2015)	HarveyNER (Chen et al., 2022) MIT-Restaurant (Liu et al., 2013) TweetNER7 (Ushio et al., 2022) WNUT2017 (Derczynski et al., 2017)
Software	StackOverflowNER (Tabassum et al., 2020)	
Transportation	FindVehicle (Guan et al., 2024)	
Wikipedia	Few-NERD (Ding et al., 2021) MultiCoNER[ℂ] (Fetahu et al., 2023a,b) WikiNEuRal[ℂ] (Tedeschi et al., 2021) Polyglot-NER[ℂ] (Al-Rfou et al., 2015)	Japanese-Wikipedia (近江崇宏, 2021) MultiNERd[ℂ] (Tedeschi and Navigli, 2022) WikiANN[ℂ] (Pan et al., 2017; Rahimi et al., 2019)

Table 1: Datasets included in MELD grouped by their primary domains. The superscript ℂ indicates silver standard datasets, and 2l indicates datasets with both silver- and gold-standard splits.

be redistributed or automatically downloaded from their source. The primary reasons for exclusion were the requirement for users to acquire paid licenses, personally request access, or complete a registration process. For these reasons, we also excluded some historically relevant datasets that have been commonly used for NER evaluation and included in previous benchmarks, but are subject to restrictive licensing agreements, including OntoNotes (Hovy et al., 2006), ACE 2004 (Mitchell et al., 2005) and ACE 2005 (Walker et al., 2006). However, we include the widely-used CoNLL 2003 dataset (Tjong Kim Sang and De Meulder, 2003), which can be accessed programmatically under the assumption that the user signs and submits an agreement with the copyright holder. Users of our data may opt not to include the CoNLL 2003 subset during data processing if no comparison with

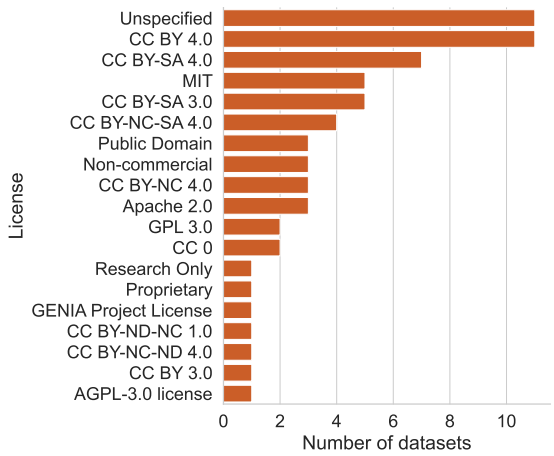

Figure 2: Licenses of datasets included in MELD. Where text and annotations are licensed separately, all licenses are included.

previous work evaluated on the dataset is desired.

Figure 2 shows the distribution of dataset licenses in MELD. Of the 60 included datasets, 45 were released under open licenses and can be redistributed under their respective conditions. 13 datasets were made publicly available for research purposes but not published under specific license terms covering redistribution.

3.1.2. Selection Process

We started by selecting English and multilingual datasets across domains to match previous open-domain NER work by Zhou et al. (2024) and Yang et al. (2025), provided they were not ruled out by our exclusion criteria. As a consequence of these criteria, MELD currently covers a smaller set of Chinese data compared to Yang et al. (2025). After establishing this core set of data, we then focused on expanding coverage by incorporating diverse datasets across additional languages.

Gold-Standard Datasets When selecting gold-standard datasets, we prioritized datasets with diverse domains. This is in contrast to Palen-Michel et al. (2025), who focus on general entity types. Our second priority were languages that do not use the Latin script. Script transfer is notoriously challenging in multilingual models, including state-of-the-art LLMs (Nguyen et al., 2025). For instance, Zaratiana et al. (2024) found that while their zero-shot NER model GliNER generalizes remarkably well cross-lingually after only fine-tuning on English NER data, there is a substantial drop in performance when evaluating on languages using non-Latin scripts.

From Universal NER (Mayhew et al., 2024), we include all datasets from its initial release, as well as a more recent Norwegian subset based on the NorNE dataset (Jørgensen et al., 2020). For Finnish, we include both the Turku NER Corpus (Luoma et al., 2020) and TurkuONE (Luoma et al., 2021). TurkuONE includes the Turku NER Corpus and is annotated with more fine-grained entity types. In contrast, document boundaries are preserved in the Turku NER Corpus for long context evaluation which were discarded in TurkuONE.

Silver-Standard Datasets For the majority of the world's languages, no gold-standard NER data is currently available. Therefore, we additionally include commonly used silver-standard datasets in our collection. We select the more frequently used version of WikiANN with balanced splits for 176 out of the originally 282 languages (Rahimi et al., 2019) due to availablility.

We include Polyglot-NER (Al-Rfou et al., 2015), since it has been used in previous work (Zhou et al., 2024; Yang et al., 2025). However, it was designed for training rather than evaluation without a pre-defined test split. Since we already include multiple multilingual silver standard datasets, we therefore decided to not define our own splits but still provide Polyglot-NER in our standardized format to facilitate future work on supervised training and evaluation.

Hybrid Silver/Gold-Standard Datasets MELD contains three additional datasets that are partially silver-standard but include gold-standard annotations for evaluation: the multilingual Naamapadam dataset (Mhaske et al., 2023), which covers 11 languages from the Indian subcontinent and includes manually annotated test sets for 9 of them; the Armenian pioNER dataset (Ghukasyan et al., 2018), which contains a manually annotated gold-standard test set; and RaTE NER (Zhao et al., 2024), an English NER dataset of radiology reports of which part of the training and validation sets have been annotated by LLMs.

3.1.3. Domain Coverage

While we cover English NER in 14 domains, annotated data for specialized domains in other languages is relatively sparse. Due to our inclusion of Wikipedia-derived synthetic data, MELD includes Wikipedia data for 171 of 194 languages. For 31 languages, annotated text is available for the "general" domain from a mixture of sources such as web crawls, Wikipedia, and news websites. Annotated data exclusively in the news domain is available for 30 languages. For the legal domain, we include dedicated datasets in three lan-

guages (German, English, and Romanian)[2]. Furthermore, we include dedicated datasets for three languages in the social media domain (English, Chinese, and Nepali). For the literature domain, we include the Tagalog subset of UniversalNER (Mayhew et al., 2024), and the literature subset of the Hungarian NYTK-KorNER dataset (Simon and Vadász, 2021). We additionally include the Latin Herodotus Project NER dataset (Erdmann et al., 2019) to cover applications in the digital humanities. Finally, we include agricultural datasets in English and Chinese, and the Spanish CANTEMIST clinical dataset for extracting tumor morphologies (Miranda-Escalada et al., 2020). Specialized datasets for the remaining six domains are only available in English.

Our final collection consists of 60 datasets, including 60 languages with gold-standard test sets, 68 languages with silver-standard test sets (in addition to WikiANN), and 194 languages overall.

3.2. Data Processing

MELD uses a modular processing framework that converts all included datasets into a single, consistent format. This design facilitates incorporating new datasets. If a dataset follows an already supported format, adding it requires only a single configuration file.

Format Standardization The creation of MELD required processing 14 distinct formats of varying complexity, with 9 formats being unique to individual datasets. While 45% of our datasets are distributed in a CoNLL-style columnar format (Tjong Kim Sang and De Meulder, 2003), they differ greatly in their number and type of columns, use of comments, "DOCSTART" tokens, or addition of non-standard metadata. In total, we identified 14 distinct CoNLL "dialects" across 27 datasets. In cases where datasets are distributed in multiple formats, we selected that which included all annotations and the original source text, or sufficient information to reconstruct it. For instance, some nested NER datasets are distributed in a tokenized and flattened format that does not preserve nested annotations or the original source text formatting. We preserve all tag sets in cases where datasets support more than one, such as separate sets of fine- and coarse-grained types.

Data Splitting We preserve the original data splits into training, validation and test sets wherever possible. For datasets that were originally evaluated in a cross-validation setup and lack pre-defined splits, such as FoNE (Snæbjarnarson et al., 2023) and TasteSET (Lawrynowicz et al., 2023), we include the entire dataset as a test set.

Validation and Cleaning Where possible, we implement strict structural validation for each format to identify internal inconsistencies. Through this approach, we identify and automatically resolve misaligned text offsets of annotations and invalid or inconsistent IOB tags, such as both "I-LOC" and "I-LOCATION" being used in the same dataset. Furthermore, we conduct minimal data cleaning, such as removing spurious XML tags from pioNER tokens (Ghukasyan et al., 2018).

In contrast, we generally preserve original noise in the text since we consider it representative of user written text in real-world settings. Fetahu et al. (2023a) deliberately insert artificial noise into MultiCoNER v2 test sets to simulate typographic or OCR errors for this reason. We do not identify incorrect labels unless in clear cases that can be discovered through internal consistency checks and automatically fixed in our implementation. Manual re-labeling is out of scope of this work.

Source-Text Alignment In contrast to Palen-Michel et al. (2025), who use a pre-tokenized format, we align all annotations to the original, untokenized text. This lets us preserve the original document formatting, which is suitable for evaluating LLM-based (Wang et al., 2025) or span-based (Zaratiana et al., 2024) NER, and ensures that our data remain tokenizer-agnostic. In cases where the original text cannot be reproduced, such as in datasets published in a pre-tokenized CoNLL format, we follow the common practice of detokenizing by joining tokens with whitespace (Yang et al., 2025; Zhou et al., 2024).

Document Structure NER has historically been applied at the sentence or paragraph level, and widely-used cross-lingual taggers such as those based on XLM-RoBERTa (Conneau et al., 2020) are limited to a maximum context length of 512 tokens. In contrast, more recent LLMs support much longer contexts, ranging from 128,000 tokens to over 2 million (Ding et al., 2024), facilitating NER at the document level. To support future evaluation in such long-context settings, we preserve the full document structure where possible. For models with shorter context lengths, we also preserve original sentence boundaries in each dataset, where available. For datasets that are annotated on a document level, we provide sentence boundaries via automatic sentence tokenization using Segment Any Text (Frohmann et al., 2024).

[2]While the Hungarian NYTK-KorNer dataset (Simon and Vadász, 2021) includes a subset of legal text, it is not annotated with domain specific tags.

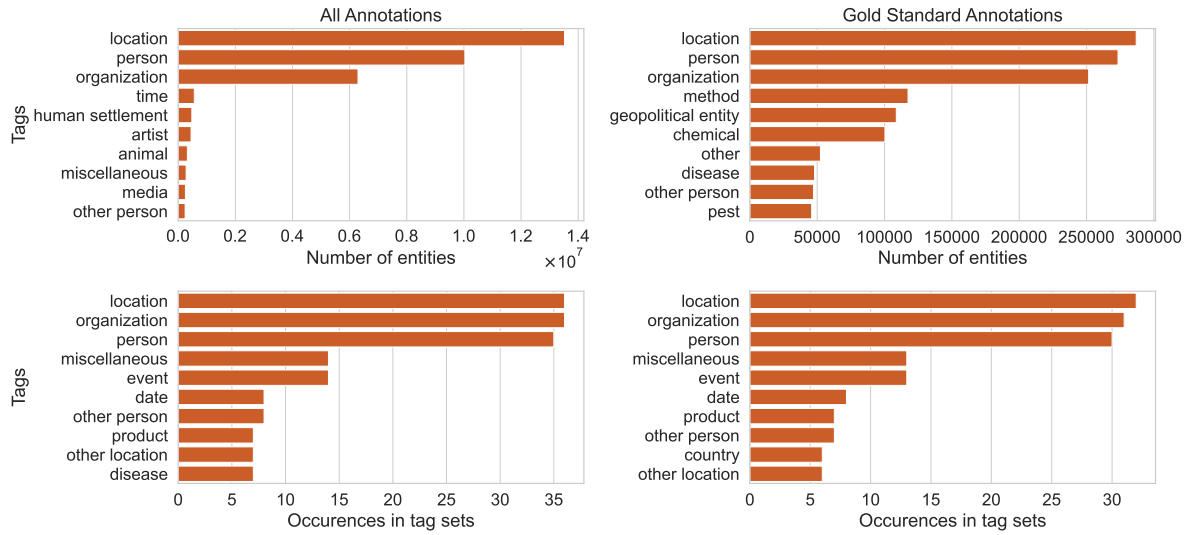

Figure 3: Distribution of annotated entity mentions (above) and entity types (below) across tag sets from all datasets (left) and only gold-standard annotations (right).

Data Format To store our annotations, we opt for a span-based format to allow for nested entities, similar to previous work (Yang et al., 2025). We choose the columnar Apache Parquet format[3] for efficiency and interoperability with widely used libraries such as Hugging Face Transformers (Wolf et al., 2020). Additionally, to support discontinuous NER datasets, we allow a sequence of multiple spans for each annotation.[4]

While we keep the original tag set for most datasets proposed by their authors, the StackOverflow NER (Tabassum et al., 2020) dataset includes annotations for eight rare entity types in addition to the set of 20 types reported, which we preserve in our processed version.

3.3. Entity Type Normalization

After collecting datasets and processing them into a single, consistent format, we additionally normalize all entity types. In contrast to supervised NER models which are trained on a fixed set of entity types, zero-shot NER models such as GliNER (Zaratiana et al., 2024) or UniversalNER (Zhou et al., 2024) encode the tag set at inference time. To ensure consistency, following Zhou et al. (2024) and Palen-Michel et al. (2025), we normalize each individual tag set into a common, more human-readable form for zero-shot methods. While we ensure a degree of consistency across datasets by mapping, e.g., "PER", "PERS" and "Person" to "person", we do not aim to construct a taxonomy as in B2NERd (Yang et al., 2025). Therefore, we do not account for differences in annotation guidelines across different dataset, such as whether "person" should only be used to annotate proper nouns or pronouns as well. Instead, we focus on dataset and tag-set-internal consistency.

We conduct some limited disambiguation for zero-shot evaluation inspired by B2NERd. For instance, the fine-grained tag set in German-LER (Leitner et al., 2019) contains "lawyer" and "judge" types in addition to a generic "person" type. In such cases, zero-shot models for NER may annotate a mention of a judge with both "judge" and "person" types, or only the more general "person" type since both types could apply. We disambiguate such cases by adding an "other" prefix, such as "other person" in this instance to clarify that entities should only be tagged as "person" if no more specific type applies. Furthermore, we replace an entity type with a more explicit type where ambiguity could not be reasonably resolved in the context of the other tags in a tag set. For instance, in TasteSET (Lawrynowicz et al., 2023), we replace the ambiguous "PART" type with "ingredient part".

FiNER-139 is a financial report dataset annotated with a set of 139 standardized US-GAAP XBRL[5] tags (Loukas et al., 2022). Since a practical application would require these standardized entity types to be used exactly, we exclude them from our mapping and keep the original tag set. Finally, since most entity types in our datasets are in English, we translated the entity types of the Japanese Wikipedia dataset (近江崇宏, 2021)

[3] https://parquet.apache.org/

[4] Currently, the only dataset with discontinuous annotations we include is the biomedical BC5CDR (Li et al., 2016) dataset.

[5] http://www.xbrl.us/xbrl-taxonomy/2020-us-gaap/

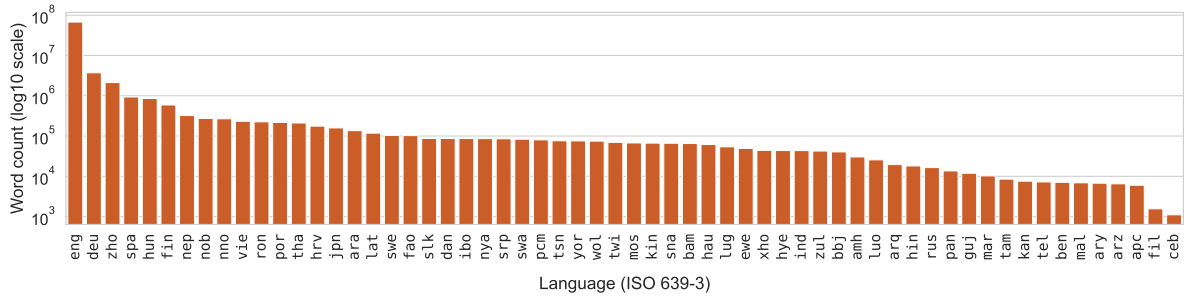

Figure 4: Distribution of languages across the gold-standard subset

from Japanese for consistency. After applying our entity type normalization approach, we reduce the number of unique entity types from 697 to 601.

Some of the included datasets such as SciREX (Jain et al., 2020) and BioRED (Luo et al., 2022) include annotations for related information extraction (IE) tasks, such as co-reference resolution and relation extraction. While we do not consider additional tasks in this work, our implementation could be extended to additionally support evaluation across multiple IE tasks in future work.

4. Data Statistics

To collect consistent word-level statistics, we retokenize each dataset using the tokenizers suggested for each language per the mapping of Penedo et al. (2025) for processing FineWeb2. Their mapping does not include four of the languages in our dataset, for which we used the original tokenization of the corresponding subsets instead. Our statistics exclude tokens that only consist of Unicode punctuation.

MELD contains a total of 605.3 million words, of which 81.4 million (13.4%) are from gold-standard datasets. All constituent datasets combined contain 37.5 million entity annotations, of which 3.4 million (9%) are gold-standard.

4.1. Entity Types and Mentions

We analyze the distribution of entity mentions by type and the frequency of each tag type across the distinct label sets present in MELD (Figure 3).

Entity statistics across all datasets are primarily dominated by Polyglot-NER, which makes up approximately 54.8% of the data. Due to its limited label set, most mentions are “locations” by a substantial margin, followed by “person” and “organization”. In contrast, while generic entity types still make up the majority of annotated mentions in the gold-standard subset, the distribution is slightly less skewed, with domain-specific tags such as “chemical”, “disease”, and “pest” being among the ten most common types.

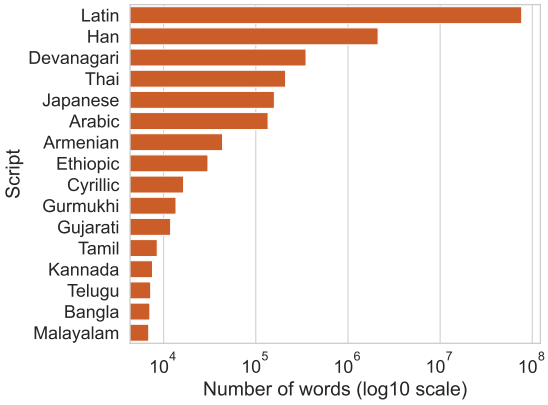

Figure 5: Distribution of scripts across the gold-standard subset

A similar pattern can be observed in the number of occurences of entity types across different tag sets, with “location”, “organization” and “person” being the most frequent types. While “event” and “date” types are frequently included in tag sets, the number of mentions annotated with these types is comparatively low. Otherwise, the distribution of tags across tag sets is mostly identical between the full data collection and gold-standard datasets.

Both in the overall data and the gold-standard subset, generic “miscellaneous” and “other” tags are prominent, occurring in 25% of all unique tag sets. Since these entity types are highly dataset-specific and therefore cannot be understood without the underlying annotation guidelines or samples from the training data, they pose a particular challenge in a zero-shot evaluation setting. For this reason, such generic entity types have been removed in previous work (Zhou et al., 2024).

4.2. Multilingual Coverage

When considering the largest tag set of each language, a tag set with at least four entity types is available for 56 of all languages included in MELD, with more diverse tag sets of at least seven entity types being available for 24 languages.

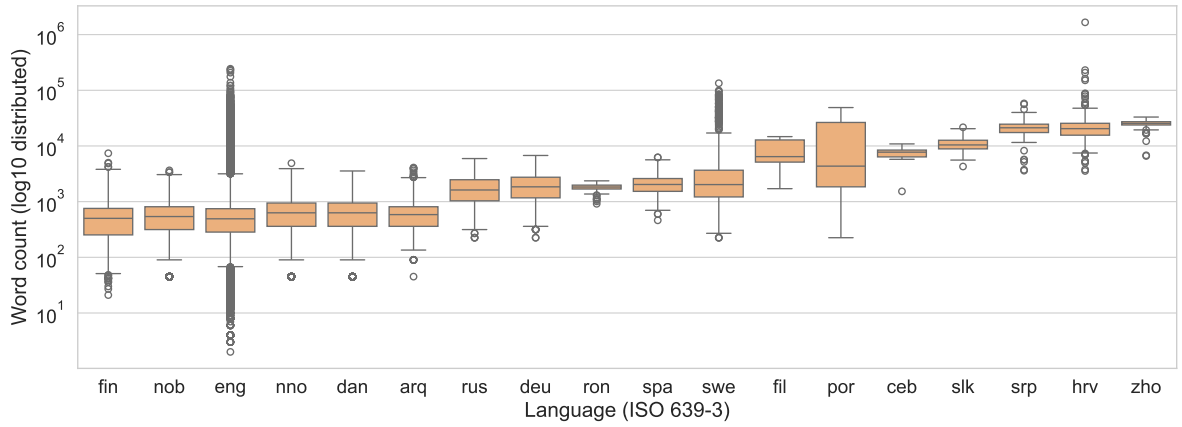

Figure 6: Distributions of document lengths per languages

In the remainder, we focus on the gold-standard subset of our data, which contains the most reliable annotations for both training and evaluation.

Out of the 81.4 million words in the gold-standard subset of MELD, approximately 84.6% are from English datasets. The distribution across all languages in this subset can be seen in Figure 4. The languages with the most gold-standard data other than English are linguistically and geographically diverse and contain both high-resource languages such as German, Chinese and Spanish, but also lower-resource languages such as Vietnamese, Arabic, Nepali and Finnish.

Figure 5 shows the distribution of scripts in our data. MELD includes text in 16 scripts in the gold-standard subset, with the majority (96.2%) of text written in the Latin script. Out of the remaining 3.13 million words, the most common scripts are used primarily throughout South and South-East Asia, such as Hanzi, Devanagari and Thai. The data for the languages written in the seven least common scripts come mainly from the test sets of the Naamapadam dataset (Mhaske et al., 2023), which contains substantially larger silver-standard training splits for these languages.

4.3. Context Lengths

Long context sizes are becoming more common in both LLMs and recently proposed encoder models, such as mmBERT (Marone et al., 2025). Full annotated documents are valuable for evaluating the ability of NER models to leverage contextual information or as part of a benchmark to evaluate long-context LLM performance. In the LLM case, this task is particularly challenging, as it requires not only understanding long inputs but also long generation of all entity mentions in the text, which is an overlooked part of LLM research (Wu et al., 2025). Furthermore, long-context understanding is particularly important in related IE tasks, such as relation extraction (Jain et al., 2020), for which mention detection or NER is usually a prerequisite.

Of the 60 datasets in our collection, 22 preserve document boundaries, covering 11 of the 14 domains in MELD. Across languages, document-level NER annotations are uncommon, leading to fully annotated documents being available for 18 languages. Figure 6 shows the distribution of document lengths across languages. The highest variation of document lengths is available for English, containing both short recipes below 100 words to upwards of 100,000 word-long scientific articles. Across the remaining languages, five languages contain documents with a geometric mean below 1000 words, for nine languages, documents are primarily between 1000 and approximately 10,000 words long, and three languages contain particularly long documents with geometric means between 20,000 and 25,000 words.

5. Conclusion

We introduced MELD, the largest unified collection of multilingual and multi-domain datasets to date. In addition to common silver-standard datasets, we include gold-standard annotations for reliable evaluation across 60 languages and 14 domains. Our work is fully open-source and reproducible to facilitate the development and evaluation of both LLMs for multilingual NER and dedicated multilingual zero-shot NER models. Furthermore, by preserving document boundaries, our collection supports research on long-context NER across languages and domains.

For future work, we aim to further expand MELD with more datasets to support additional languages and domains. Furthermore, we plan to define evaluation tracks based on MELD for, e.g., LLM-based methods using more complex

prompting techniques with annotation guidelines or retrieval-augmented generation.

6. Limitations

While MELD covers a broad range of languages and domains, only silver-standard data is available for 69% of them. This can be partly remedied by adding further existing datasets for currently unsupported languages, but additional manual annotation efforts are required for low-resource languages. Moreover, even in cases where gold-standard datasets exist, they are most commonly annotated with only limited, general entity types. Therefore, further work is required to better support zero-shot NER evaluation on more diverse tag sets across languages. Finally, a common problem in evaluation of LLM generalization is data contamination, where a model has already seen parts of the evaluation data. Further work is required to measure the impact of data contamination when evaluating on MELD.

7. Acknowledgments

We thank Kendija Salome Kasendu for her valuable assistance during the dataset curation and entity type normalization process. This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

8. Bibliographical References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30- May 2, 2015*.

Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pei Chen, Haotian Xu, Cheng Zhang, and Ruihong Huang. 2022. Crossroads, buildings and neighborhoods: A dataset for fine-grained location recognition. In *Proceedings of the 2022 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies, pages 3329–3339, Seattle, United States. Association for Computational Linguistics.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sayan De, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Agriner: An ner dataset of agricultural entities for the semantic web. In *The Semantic Web: ESWC 2023 Satellite Events*, pages 59–63, Cham. Springer Nature Switzerland.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. 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)*, pages 3198–3213, Online. Association for Computational Linguistics.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongRoPE: Extending LLM context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11091–11104. PMLR.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Niama Elkhbir, Urchade Zaratiana, Nadi Tomeh, and Thierry Charnois. 2023. Cross-dialectal named entity recognition in Arabic. In *Proceedings of ArabicNLP 2023*, pages 140–149, Singapore (Hybrid). Association for Computational Linguistics.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Bé atrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2027–2051, Singapore. Association for Computational Linguistics.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Tsolak Ghukasyan, Garnik Davtyan, K. Avetisyan, and Ivan Andrianov. 2018. pioner: Datasets and baselines for armenian named entity recognition. *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 56–61.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Runwei Guan, Ka Lok Man, Feifan Chen, Shanliang Yao, Rongsheng Hu, Xiaohui Zhu, Jeremy Smith, Eng Gee Lim, and Yutao Yue. 2024. Findvehicle and vehiclefinder: a ner dataset for natural language-based vehicle retrieval and a keyword-based cross-modal vehicle retrieval system. *Multimedia Tools and Applications*, 83(8):24841–24874.

Hangfeng He and Xu Sun. 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*, pages 713–718, Valencia, Spain. Association for Computational Linguistics.

Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90% solution. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX - NAACL '06*.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating named entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4547–4556, Marseille, France. European Language Resources Association.

Siti Oryza Khairunnisa, Aizhan Imankulova, and Mamoru Komachi. 2020. Towards a standardized dataset on indonesian named entity recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel V ázquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Dong-Hong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, P. Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber Ahmad Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin M. Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, K. E. Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, C Ata, Tolga Can, Anabel Usie, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzábal, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2 – S2.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An annotated corpus for machine reading of instructions in wet lab protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 97–106, New Orleans, Louisiana. Association for Computational Linguistics.

Aman Kumar and Binil Starly. 2021. "FabNER": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*, 33:2393 – 2407.

Agnieszka Lawrynowicz, Anna Wróblewska, Agnieszka Kaliska, Maciej Pawlowski, Dawid Wiśniewski, Witold Sosnowski, and Jakub Dutkiewicz. 2023. Fine-grained and complex food entity recognition benchmark for ingredient substitution. In *Proceedings of the 12th*

Knowledge Capture Conference 2023, K-CAP '23, page 25–29, New York, NY, USA. Association for Computing Machinery.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany. Springer. 10/11 September 2019.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 523–532, Dublin, Ireland. Association for Computational Linguistics.

Jingjing Liu, Panupong Pasupat, D. Scott Cyphers, and James R. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreferencefor scientific knowledge graph construction. In *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: A rich biomedical relation extraction dataset. *Briefing in Bioinformatics*.

Jouni Luoma, Li-Hsin Chang, Filip Ginter, and Sampo Pyysalo. 2021. Fine-grained named entity annotation for Finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 135–144, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning.

Hend Al-Khalifa Mashael AlDuwais and Abdulmalik AlSalman. 2024. Cleananercorp: Identifying and correcting incorrect labels in the anercorp dataset. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools*.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Šuppa, Hila Gonen, Joseph Marvin Imperial, Börje F. Karlsson, Peiqin Lin, Nikola Ljubešić, LJ Miranda, Barbara Plank, Arij Riab, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

A Miranda-Escalada, E Farré, and M Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. Linguistic Data Consortium, Philadelphia.

Hoang H Nguyen, Khyati Mahajan, Vikas Yadav, Julian Salazar, Philip S. Yu, Masoud Hashemi, and Rishabh Maheshwary. 2025. Prompting with phonemes: Enhancing LLMs' multilinguality for non-Latin script languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11975–11994, Albuquerque, New Mexico. Association for Computational Linguistics.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: Data sets and algorithms. *The International FLAIRS Conference Proceedings*, 35.

Nobal Niraula and Jeevan Chapagain. 2023. DanfeNER - named entity recognition in nepali tweets. *The International FLAIRS Conference Proceedings*, 36(1).

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chester Palen-Michel, Maxwell Pickering, Maya Kruse, Jonne Sälevä, and Constantine Lignos. 2025. OpenNER 1.0: Standardized open-access named entity recognition datasets in 50+ languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33649–33674, Suzhou, China. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 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)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Processings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 548––554.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 149–155.

Wannaphong Phatthiyaphaibun. 2024. Thai NER 2.2. https://doi.org/10.5281/zenodo.10795907.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented lan-

guages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krü ger. 2021. Somesci- a 5 star open data gold standard knowledge graph of software mentions in scientific articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4574–4583, New York, NY, USA. Association for Computing Machinery.

David Schindler, Tazin Hossain, Sascha Spors, and Frank Krüger. 2024. A multilevel analysis of data quality for formal software citation. *Quantitative Science Studies*, 5(3):637–667.

Agam Shah, Abhinav Gullapalli, Ruchit Vithani, Michael Galarnyk, and Sudheer Chava. 2024. Finer-ord: Financial named entity recognition open research dataset. *arXiv preprint arXiv:2302.11157*.

Eszter Simon and Noémi Vadász. 2021. Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer.

Larry L. Smith, Lorraine K. Tanabe, Rie Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, C. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence E. Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter W. Adriaans, Christian Blaschke, Rafael Torres, Mariana L. Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9:S2 – S2.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands. Linköping University Electronic Press, Sweden.

Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter. 2020. Code and named entity recognition in StackOverflow. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4913–4926, Online. Association for Computational Linguistics.

Kushal Tatariya, Artur Kulmizev, Wessel Poelman, Esther Ploeger, Marcel Bollmann, Johannes Bjerva, Jiaming Luo, Heather Lent, and Miryam de Lhoneux. 2024. How good is your wikipedia? auditing data quality for low-resource and multilingual nlp. *arXiv preprint arXiv:2411.05527*.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 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*, pages 142–147.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Asahi Ushio, Leonardo Neves, Vitor Silva, Francesco. Barbieri, and Jose Camacho-Collados. 2022. Named Entity Recognition in Twitter: A Dataset and A nalysis on Short-Term Temporal Shifts. In *The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuhao Wu, Yushi Bai, Zhiqing Hu, Shangqing Tu, Ming Shan Hee, Juanzi Li, and Roy Ka-Wei Lee. 2025. Shifting long-context llms research from input to output.

Yuming Yang, Wantong Zhao, Caishuang Huang, Junjie Ye, Xiao Wang, Huiyuan Zheng, Yang Nan, Yuran Wang, Xueying Xu, Kaixin Huang, Yunke Zhang, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Beyond boundaries: Learning a universal entity taxonomy across datasets and languages for open named entity recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10902–10923, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaochuang Yao, Xia Hao, Ruilin Liu, Lin Li, and Xuchao Guo. 2024. AgCNER, the first large-scale chinese named entity recognition dataset for agricultural diseases and pests. *Scientific Data*, 11(1):769.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13083–13100, Miami, Florida, USA. Association for Computational Linguistics.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted distillation from large language models for open named entity recognition. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

近江崇宏. 2021. Wikipedia を用いた日本語の固有表現抽出のデータセットの構築. 言語処理学会第 *27* 回年次大会発表論文集, pages 350–352.