Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark

Stephen Mayhew[α] **Terra Blevins**[β] **Shuheng Liu**[γ] **Marek Šuppa**[δϵ]
Hila Gonen[β] **Joseph Marvin Imperial**[ικ] **Börje F. Karlsson**[λ] **Peiqin Lin**[η]
Nikola Ljubešić[μ] **LJ Miranda**[θ] **Barbara Plank**[ηζ] **Yuval Pinter**[ν]
[α]Duolingo, [β]University of Washington, [γ]Georgia Institute of Technology,
[δ]Comenius University in Bratislava, [ϵ]Cisco, [ι]National University, [κ]University of Bath,
[λ]Beijing Academy of Artificial Intelligence, [η]LMU Munich, [μ]Jožef Stefan Institute,
[θ]Allen Institute for Artificial Intelligence, [ζ]IT University of Copenhagen, [ν]Ben-Gurion University

Abstract

We introduce Universal NER (UNER), an open, community-driven project to develop gold-standard NER benchmarks in many languages. The overarching goal of UNER is to provide high-quality, cross-lingually consistent annotations to facilitate and standardize multilingual NER research. UNER v1 contains 18 datasets annotated with named entities in a cross-lingual consistent schema across 12 diverse languages. In this paper, we detail the dataset creation and composition of UNER; we also provide initial modeling baselines on both in-language and cross-lingual learning settings. We release the data, code, and fitted models to the public.[1]

1 Introduction

High-quality data in many languages is a necessity for broadly multilingual natural language processing. In named entity recognition (NER), the majority of annotation efforts are centered on English, and cross-lingual transfer performance remains brittle (e.g., Chen et al., 2023b; Ma et al., 2023). Amongst non-English human-annotated NER datasets, while there have been multiple separate efforts in this front (e.g., Agić and Ljubešić, 2014; Plank, 2019; Adelani et al., 2022), these either have disjoint annotation schemes and labels, cover a single language or small set of related languages, or are not widely accessible (e.g., Strassel and Tracey, 2016). For most of the world's languages, the only readily available NER data is the automatically annotated WikiANN dataset (Pan et al., 2017), though this annotation paradigm introduces data quality issues and limits its usefulness for evaluation (Lignos et al., 2022).

To address this data gap, we propose Universal NER (UNER), an open community effort to develop gold-standard named entity recognition benchmarks across a large number of languages. Each dataset in Universal NER is annotated by primarily native speakers on the text of an existing Universal Dependencies treebank (UD; Nivre et al., 2020). Inspired by Universal Dependencies, the overarching philosophy of the UNER project is to provide a shared, universal definition, tagset, and annotation schema for NER that is broadly applicable across languages (Figure 1).

EN	The expedition was led by General Joseph Burgoyne who intended to reach Albany.
DE	Die Expedition wurde angeführt von General Joseph Burgoyne, der Albany erreichen wollte.
RU	Поход возглавлял генерал Джон Бергойн, который намеревался достичь Олбани.
ZH	約翰·伯戈因將軍領導了這次遠征，他希望能到達奧爾巴尼。

Figure 1: Parallel sentences annotated with person (PER) and location (LOC) named entities in English (EN), German (DE), Russian (RU), and Chinese (ZH).

The current version of Universal NER, UNER v1, contains 18 datasets spanning 12 languages (Section 4). To establish performance baselines on UNER, we finetune an XLM-R model on various training configurations (Section 5) showing that while NER transfer performance between European languages is relatively strong, there remains a gap when transferring to different scripts or language typologies.

The goal of the UNER project is to facilitate multilingual research on entity recognition by addressing the need in the multilingual NLP community for standardized, cross-lingual, and manually annotated NER data. With the release of UNER v1, we plan to expand UNER to new languages and datasets, and we welcome all new annotators who are interested in developing the project.

[1] https://www.universalner.org, UNER v1 available at https://doi.org/10.7910/DVN/GQ8HDL

2 Dataset Design Principles

Named entity recognition (NER) is the task of identifying spans of text in a given context that uniquely refer to specific *named entities*. The task of NER has a long tradition (Grishman, 2019) and facilitates many downstream NLP applications, such as information retrieval (Khalid et al., 2008) and question answering (Mollá et al., 2006). Furthermore, successful NER tagging requires a model to reason about semantic and pragmatic world knowledge, which makes the task an informative evaluation setting for testing NLP model capabilities.

As with Universal Dependencies, our goal is to develop an annotation schema that can work in any language. Traditionally, the UD (Nivre et al., 2016) and UPOS (Petrov et al., 2012) projects have chosen what amounts to the intersection of tags across all language-specific tagsets, keeping the resultant tagset broad and simple. We follow a similar strategy, picking tags that broadly cover the space of proper nouns.

Universal NER's annotation schema emphasizes three coarse-grained entity types: Person (`PER`), Organization (`ORG`), and Location (`LOC`). We provide a short description and an example for each tag in our schema.

`PER` The `PERSON` tag includes names of people, real or fictional, but not nominals.

"Mr. Robinson$_{\texttt{PER}}$ smiled at the teacher."

`ORG` The `ORGANIZATION` tag is used for named collections of people.

"The FDA$_{\texttt{ORG}}$ announced time travel pills tomorrow."

`LOC` The `LOCATION` tag covers all types of named locations.

"I will arise and go now, and go to Innisfree$_{\texttt{LOC}}$"

Figure 1 demonstrates how named entities and their corresponding annotations surface across languages. In some cases (such as in the English and German sentences), the surface forms of named entities are shared. However, often these forms vary—as in the Russian and Chinese examples—which makes entity identification and tagging more challenging, particularly in cross-lingual settings.

Annotation Guidelines In preparation for annotation, we developed extensive annotation guidelines,[2] using the NorNE project guidelines Jørgensen et al. (2020) as a starting point. Along with tag descriptions, our guidelines include many examples, as well as instructions for dealing with ambiguity and unclear constructions, such as email addresses, pet names, and typographical errors.

We expect that, as annotation proceeds, the guidelines will be further refined and updated. To manage this, we track version numbers for the guidelines, as well as changelogs. Each data release will include the corresponding annotation guidelines at time of release.

[2] `http://www.universalner.org/guidelines/`

3 Dataset Annotation Process

Having described the theoretical basis for the tagset, we now discuss the mechanics of annotation.

Sourcing Data We chose to use the Universal Dependency corpora as the default base texts for annotation. This jumpstarts the process: there's high coverage of languages, and the data is already collected, cleaned, tokenized, and permissively licensed. Further, by adding an additional annotation layer onto an already rich set of annotations, we not only support verification in our project (Section 4.3), but we also enable multilingual research on the full pipeline of core NLP. Since UD is annotated at the word level, we follow a BIO annotation schema (specifically IOB2), where words forming the beginning (nonbeginning) part of an `X` entity ($\texttt{X} \in \{\texttt{PER}, \texttt{LOC}, \texttt{ORG}\}$) are annotated `B-X` (`I-X`, respectively), and all other words are given an `O` tag. For the sake of continuity, we preserve all tokenization from UD.

Even though we chose UD as the default data source, we do not limit the project to UD corpora (particularly for languages not currently included in UD). The only criteria for inclusion in the UNER corpus is that the tagging schema matches the UNER guidelines. We are also open to converting existing NER efforts on UD treebanks to UNER: in this initial release, we include three datasets that are transferred from other manual annotation efforts on UD sources (for DA, HR, and SR).

Sourcing Annotators For the initial UNER annotation effort, we recruited annotators from the multilingual NLP community through academic networks on social media. Annotators were organized via channels in a Slack workspace. Annotators of the datasets included in UNER thus far are volunteers and not paid. We expect that annotators are native speakers of their annotation language, or are highly proficient, but we did not issue any language tests.

For the first release of UNER, the choice of the 12 dataset languages is solely dependent on the availability of annotators. As the project continues, we expect that additional languages and datasets will be added as annotators in more languages become available to contribute.

Annotation Tool We collect annotations for the UD treebanks using TALEN (Mayhew and Roth, 2018), a web-based tool for span-level sequence labeling.[3] TALEN includes an optional feature that propagates annotations – if you annotate "McLovin" in one section of the document, every instance of "McLovin" in that document will be annotated. This significantly speeds up annotation but risks certain over-annotation mistakes. For example, consider the token "US", which may show up in contexts such as "The US economy..." or "THEY OFFERED TO BUY US LUNCH!"

Secondary Annotators In addition to collecting a full set of annotations from a primary annotator for each dataset, we also gather secondary annotations from another annotator on (at least) a subset of the data, in order to estimate inter-annotator agreement (Section 4.2). In general, we aim for at least 5% coverage of each data split with these secondary annotations, although most datasets have significantly more. For each document that has multiple annotators, we selected the annotator who made the most annotations in that document and used their annotations as official. This means that a dataset may have multiple annotators, but each document has exactly one annotator. We retain annotator identities in the data files.

Annotation Differences and Resolution When annotators disagreed on annotation decisions, and the inter-annotator agreement scores were low, we encouraged them to discuss the disagreements and decide if they were conflicting interpretations of the guidelines or fundamental disagreements. In the former case, annotators came to an agreement on guideline interpretations and updated annotations accordingly. In the latter, the annotations were kept as-is. Not every dataset had this resolution process.

The multilingual nature of this process also highlighted cross-language differences in named entities that affect NER annotation. For instance, most languages in UNER use capitalization as a marker of proper nouns and therefore named entities. However, Chinese does not include capitalization in its script, which makes identifying named entities more difficult and time-consuming than in other languages, potentially leading to more annotation errors. Differences in annotating NER across languages also stem from divergent definitions of proper nouns (`PROPN`) by language and the effects of translation artifacts; these issues are discussed further in Sections 4.3 and 4.4, respectively.

OTHER Tag As a helpful check for annotators, we allow the option of annotating a fourth entity type, Other (`OTH`), which is not represented in the final dataset. This had several purposes: to store annotations that behaved like mentions, but didn't quite rise to the level of the other tags; to potentially measure annotation disagreement on ambiguous cases; to store an additional layer of annotation. Not all annotators used it, and those that did were not always consistent. In practice, it was applied most often to language and nationality names and brands. The `OTH` tag roughly corresponds to the `MISC` tag used in CoNLL 2003, which has been described as being "ill-defined" (Adelani et al., 2022).

Dataset Transfer Most of the included datasets are annotated from scratch using the annotation process detailed above, but a few (DA `ddt`, HR and SR `set`) are transferred from other sources. The Danish `ddt` annotations are derived from the *News* portion of the DaN+ dataset (Plank et al., 2020); this text corresponds to the Universal Dependencies `ddt` treebank. The Croatian `hr` annotations come from the `hr500k` dataset (Ljubešić et al., 2016), half of which, consisting of newspaper and various web texts, was used for producing the Croatian Universal Dependencies `hr_set` treebank (Agić and Ljubešić, 2015). The Serbian `sr` data come from the `SETimes.SR` dataset (Batanović et al., 2018), which was used in its fullness to produce the Serbian Universal Dependencies `sr_set` treebank (Samardžić et al., 2017). The original Croatian and Serbian NER annotations were annotated and curated in multiple iterations by various native speaker annotations. The annotations added to the UNER dataset were, however, slightly modified, to conform to the UNER annotation guidelines. While nationalities and similar groups were annotated as `PER` in the original dataset, in the UNER dataset such entities were not included.

[3] https://github.com/mayhewsw/talen-react

		Sentences				Entities				Tokens			
Lang.	Dataset	Train	Dev	Test	All	Train	Dev	Test	All	Train	Dev	Test	All
DA	ddt	4,383	564	565	5,512	3,022	379	446	3,847	80,378	10,332	10,023	100,733
EN	ewt	12,543	2,001	2,077	16,621	7,022	966	1,088	9,076	204,579	25,149	25,097	254,825
HR	set	6,914	960	1,136	9,010	8,261	1,218	1,403	10,882	152,857	22,292	24,260	199,409
PT	bosque	7,018	1,172	1,167	9,357	8,101	1,401	1,215	10,717	171,776	28,447	27,604	227,827
SK	snk	8,483	1,060	1,061	10,604	2,707	636	915	4,258	80,628	12,733	12,736	106,097
SR	set	3,328	536	520	4,384	5,020	742	847	6,609	74,259	11,993	11,421	97,673
SV	talbanken	4,303	504	1,219	6,026	967	23	196	1,186	66,646	9,797	20,377	96,820
ZH	gsd	3,997	500	500	4,997	6,136	754	767	7,657	98,616	12,663	12,012	123,291
	gsdsimp	3,997	500	500	4,997	6,118	753	763	7,634	98,616	12,663	12,012	123,291
DE	pud	–	–	1,000	1,000	–	–	1,039	1,039	–	–	21,331	21,331
EN	pud	–	–	1,000	1,000	–	–	1,038	1,038	–	–	21,176	21,176
PT	pud	–	–	1,000	1,000	–	–	1,099	1,099	–	–	23,407	23,407
RU	pud	–	–	1,000	1,000	–	–	1,036	1,036	–	–	19,355	19,355
SV	pud	–	–	1,000	1,000	–	–	1,029	1,029	–	–	19,076	19,076
ZH	pud	–	–	1,000	1,000	–	–	1,137	1,137	–	–	21,415	21,415
CEB	gja	–	–	188	188	–	–	49	49	–	–	1,295	1,295
TL	trg	–	–	128	128	–	–	92	92	–	–	734	734
	ugnayan	–	–	94	94	–	–	61	61	–	–	1,097	1,097

Table 1: Universal NER has broad coverage of named entities in several languages and domains, adding annotations to the development, testing, and training sets from Universal Dependencies (Nivre et al., 2020).

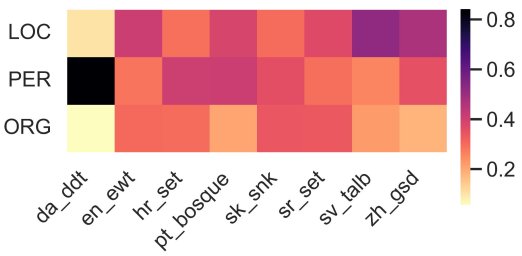

Figure 2: Distribution of tags in different UNER training sets. zh_gsdsimp has the same distribution as zh_gsd.

4 Universal NER: Statistics and Analysis

This section presents an overview of the Universal NER (UNER) dataset. UNER v1 adds an NER annotation layer to 18 datasets (primarily treebanks from UD) and covers 12 geneologically and typologically diverse languages: Cebuano, Danish, German, English, Croatian, Portuguese, Russian, Slovak, Serbian, Swedish, Tagalog, and Chinese[4]. Overall, UNER v1 contains nine full datasets with training, development, and test splits over eight languages, three evaluation sets for lower-resource languages (TL and CEB), and a parallel evaluation benchmark spanning six languages.

[4]Languages sorted by their ISO 639-1/639-2 codes (International Organization for Standardization, 2002, 1998)

4.1 Dataset Statistics

In Table 1 we report the number of sentences, tokens, and annotated entities for each dataset in UNER. The datasets in UNER cover a wide range of data quantities: some provide a limited amount of evaluation data for a commonly low-resourced language, whereas others annotate thousands of training and evaluation sentences.

The datasets in UNER also cover a diverse range of domains, spanning web sources such as social media to more traditional provenances like news text. Table 5 in the appendix presents the full set of sources for the data, and the distribution of NER tags in each dataset, along with references to original treebank papers. The variety in data sources leads to varied distributions of tags across datasets (Figure 2).

4.2 Inter-Annotator Agreement

We calculate inter-annotator agreement (IAA, Table 2) for each dataset in UNER that was annotated with the above process and for which we have secondary annotations. Table 2 reports agreement as per-label F_1 score, using one annotator as "reference," and the other as "prediction."

ORG vs LOC Confusion The agreement on ORG and LOC is generally lower than that on PER. The

		Train				Dev				Test			
Lang.	Dataset	LOC	ORG	PER	% Docs	LOC	ORG	PER	% Docs	LOC	ORG	PER	% Docs
DA	ddt	.875	.778	.959	100%	.917	.765	.934	100%	.882	.805	.975	100%
EN	ewt	.696	.533	.925	20%	.786	.640	.949	20%	.825	.869	.969	20%
PT	bosque	.928	.902	.974	11%	.850	.885	.980	25%	.955	.914	.975	23%
SK	snk	.840	.743	.900	100%	.801	.597	.770	100%	.837	.621	.823	100%
SV	talbanken	.857	.670	.913	100%	.800	.461	.888	100%	.937	.812	.871	100%
ZH	gsd	.800	.724	.917	14%	.795	.661	.956	100%	.860	.711	.944	23%
DE	pud	-	-	-	-	-	-	-	-	.709	.840	.812	6%
EN	pud	-	-	-	-	-	-	-	-	1.00	.936	.966	6%
PT	pud	-	-	-	-	-	-	-	-	.903	.920	.985	14%
RU	pud	-	-	-	-	-	-	-	-	.719	.531	.891	100%
SV	pud	-	-	-	-	-	-	-	-	.865	.735	.944	100%
ZH	pud	-	-	-	-	-	-	-	-	.752	.776	.971	20%
CEB	gja	-	-	-	-	-	-	-	-	.769	1.00	.914	71%
TL	trg	-	-	-	-	-	-	-	-	.833	-	.957	100%
	ugnayan	-	-	-	-	-	-	-	-	.913	-	-	100%

Table 2: Inter-annotator agreement scores for the datasets annotated natively for the Universal NER project. We don't report IAA for the datasets adapted from other sources, or from zh_gsdsimp, which has nearly identical annotations to zh_gsd.

annotation guidelines allow that certain named entities may take either the ORG or LOC tag based on context. In certain cases, the context is underspecified, and there is a natural ambiguity. For example, a restaurant is a LOC when you go there to eat, but it is an ORG when it hires a new chef. A city is a LOC when you move there, but it is an ORG when it levies taxes. Officially, it is the *city government* that levies taxes, but common usage allows, for example, "Springfield$_{\text{ORG}}$ charges a brutal income tax." CoNLL 2003 English also has this ambiguity, with many documents where city names, representing sports teams, are annotated as ORG. This ambiguity was especially common in the en_ewt train and validation splits, primarily in documents in the *reviews* domain, which are short and very informal (e.g. "we love pamelas").

4.3 Agreement with the PROPN POS Tag

The proper noun (PROPN) part-of-speech tag used in UD represents the subset of nouns that are used as the name of a specific person, place, or object (Nivre et al., 2020). We hypothesize that named entities as defined in UNER act roughly as a subset of these PROPN words or phrases, although not a strict subset due to divergent definitions. To test this, we calculate the precision of the UNER annotations against the UD PROPN tags (Table 3, F_1 scores reported in Table 4). Overall, precision is relatively high, with a mean precision of 0.745 across datasets. Lower precision is often due to multi-word names containing non-PROPN words (e.g., "Catherine the Great"). The differences in precision can also be due to language-specific PROPN annotation guidelines: for example, while the English PUD treebank tags the United States entity as "United$_{\text{PROPN}}$ States$_{\text{PROPN}}$", Russian PUD tags it as "Соединенных$_{\text{ADJ}}$ Штатов$_{\text{NOUN}}$".

4.4 Cross-lingual Agreement in UNER

UNER contains sentence-aligned evaluation sets for six languages (German, English, Portuguese, Russian, Swedish, and Chinese) that are annotated on top of the Parallel Universal Dependencies treebanks (PUD; Zeman et al., 2017). Figure 3 summarizes the similarity of the NER annotations across these target languages in PUD.

We find that the overall distribution of tags is similar for the Western European languages (left panel): the English, German, and Swedish annotations contain very similar counts of LOC and PER entities, while there is slightly more variance in ORG tags. Portuguese has a similar distribution with slightly more LOC entities. However, the Russian and Chinese annotations contain differing distributions from both these languages and each other.

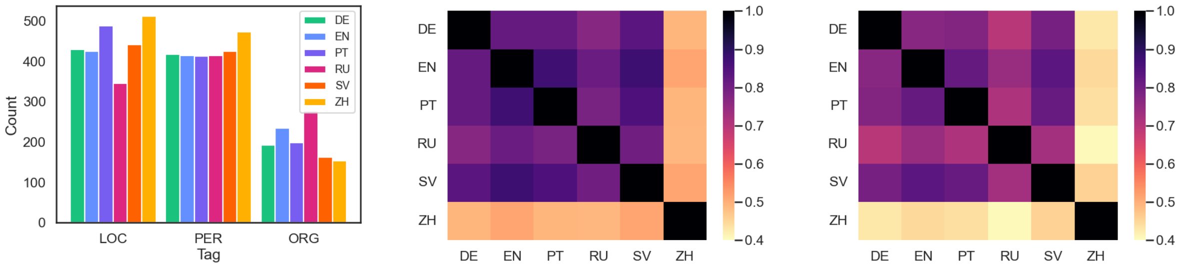

Figure 3: Cross-lingual comparison of NER Annotations on top of PUD treebanks. **Left**: Global distribution of tags for each PUD language. **Center**: Sentence-level agreement between languages for the number of entities. **Right**: Sentence-level agreement between languages for the identity of entities.

Lang.	Dataset	Train	Dev	Test
DA	ddt	.709	.729	.722
EN	ewt	.890	.895	.892
HR	set	.683	.651	.671
PT	bosque	.864	.881	.844
SK	snk	.803	.783	.688
SR	set	.687	.631	.680
SV	talbanken	.766	.756	.842
ZH	gsd	.605	.624	.616
ZH	gsdsimp	.601	.604	.617
DE	pud	-	-	.712
EN	pud	-	-	.872
PT	pud	-	-	.749
RU	pud	-	-	.708
SV	pud	-	-	.810
ZH	pud	-	-	.634
CEB	gja	-	-	.980
TL	trg	-	-	.958
TL	ugnayan	-	-	.654

Table 3: Comparing the overlap (Precision) between UNER annotations and UD PROPN tags.

A similar trend occurs in the sentence-level pairwise agreement on entity counts and identities between languages (center). There is relatively high agreement on the number of entities between European languages, with Russian differing slightly more from English, German, Portuguese, and Swedish. However, the Chinese benchmark agrees less frequently: the Chinese annotations match other languages on the number of entities in 50.4% of sentences; the other languages have an average agreement of 71.7–75.6%. Pairwise agreement on the specific entities in a given sentence shows similar behavior, albeit with lower agreement overall (right).

Many of these annotation differences likely stem from the translation process. While the data is aligned at the sentence level, linguistic variation and translator decisions may cause an entity to be added to or removed from the sentence, or the concept to be expressed in a manner that no longer qualifies as a named entity under the annotation guidelines.[5] While we cannot directly measure inter-annotator agreement across languages because of the above differences, some variation also undoubtedly stems from annotation differences and errors, just as these cause disagreement between annotators on the same benchmark.

In the case of Chinese and English, we manually audited the annotation discrepancies. The differences in the LOC and ORG tags mostly stem from the confusion outlined in Section 4.2. Additionally, we saw more than 30 instances that could be explained by language-specific morphological inflection rules. Specifically, country names are used directly to modify the following nouns in Chinese as opposed to English using the adjectival form.[6] Finally, the increase in PER entities can be best explained by the style of Chinese writing, which tends to transliterate non-Chinese names into Chinese and append the Latin name in parentheses; in these cases, each instance of the name would be tagged as a separate PER entity.[7]

[5] Consider the phrases: “奧巴馬對在北卡羅來納大學運動場上的群衆説道。” and “he told the crowd gathered on a sports field at the University of North Carolina.” In Chinese, *Obama* (奧巴馬) is referred to by name, whereas the English version uses a pronoun.

[6] I.e., “韓國公司” ‘South Korean company’. The Chinese word “韓國” means the country ‘South Korea’, and in this case, directly modifies the noun “公司” ‘company’. This word was consequently labeled as LOC, whereas its English counterpart is O.

[7] An example is “聖羅斯季斯拉夫（St. Rastislav)”, in which the English name is parenthesized and kept in the Chinese sentence, causing both names to be annotated.

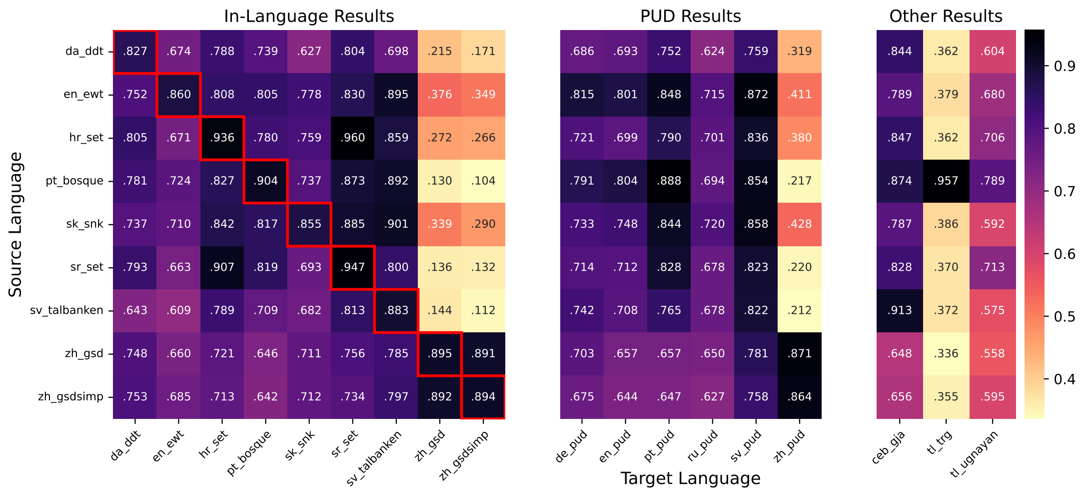

Figure 4: Heatmap of micro F_1 scores on test sets with different fine-tuned models. y-axis indicates the dataset that the model is fine-tuned in, and the x-axis indicates the datasets that the models are evaluated on. **Left**: Model performance on datasets that contains the train, dev and test splits. The highlighted diagonal cells are the in-dataset results. **Center**: Model performance on the PUD datasets. **Right**: Model performance on all other datasets.

5 Baselines for UNER

In this section, we establish initial performance baselines on the datasets in UNER v1. We provide results leveraging XLM-R_{Large}, on both in-language and cross-lingual learning settings.

5.1 Experiment Setup

We fine-tune XLM-R_{Large} (Conneau et al., 2020) on the UNER datasets in which train and dev sets are available, i.e. the set of datasets {ddt, ewt, set, bosque, snk, set, talbanken, gsd, gsdsimp}. We also evaluate the performance of XLM-R_{Large} jointly finetuned on all training sets (all) listed above. We use a learning rate of 3e-5 and batch size of 16, except in the cased of talbanken and all, where we instead use a batch size of 4.

5.2 Results and Discussion

Figure 4 reports the micro F_1 scores on all test sets when XLM-R_{Large} is fine-tuned on different languages. The in-language performance shown on the diagonal on the left of Figure 4 is almost always the highest among all test sets, with a few exceptions such as Simplified Chinese vs Traditional Chinese (ZH) and Croatian (HR) vs Serbian (SR). This is most likely due to the fact that both pairs are closely related languages.

We also observe that in most cases (i.e., between European languages) cross-lingual transfer performs well, achieving over .600 F_1. However, this setting results in strikingly low performance on all three Chinese datasets {gsd, gsdsimp, pud}, as well as on the two Tagalog (TL) datasets {trg, ugnayan}. The results on the Chinese datasets align with observations from previous work (Chen et al., 2023a; Wu et al., 2020a; Bao et al., 2019) that other languages do not transfer well to Chinese. In the case of Tagalog, this is likely due to the linguistic differences from the source languages as well as the small sample size of the evaluation set, as listed in Table 1. We also noticed that many sentences from both Tagalog benchmarks have no entities tagged at all.

The performance of the model fine-tuned on all is included in Figure 5. Most all F_1 scores are similar to the F_1 scores from individual training sets; however, in some cases the joint training improves performance, such as on zh_pud which improved from .410 using a model fine-tuned on en_ewt to .860. Fine-tuning on a diverse multilingual dataset helps preserve and even improve the performance on benchmarks in diverse languages.

In addition, Table 6 also shows the tag-level performance breakdown. For all languages, F_1 on ORG is always the lowest, and LOC is almost always the second lowest. This likely stems from the similarity between ORG and LOC entities discussed in Section 4.2, whereas the names of people are usually less ambiguous, resulting in the highest F_1 on PER for most datasets. Overall, the trained models finetuned on the UNER datasets exhibit promising results for various cross-lingual experiments, and we leave further improvements on multi- and cross-lingual NER with these datasets to future work.

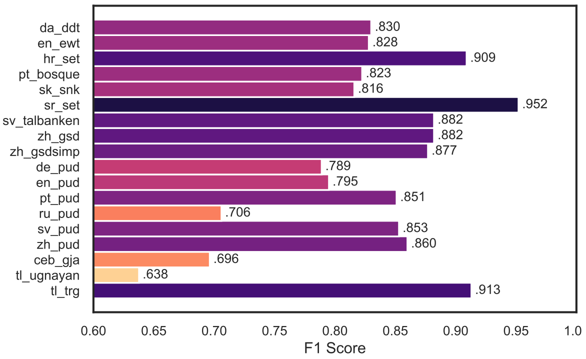

Figure 5: F_1 scores of each UNER test set after finetuning XLM-R_{Large} on `all` training sets.

6 Related Work

Adding an NER layer to UD Some single-language efforts have added a manually annotated NER layer to emerging or existing UD data. Agić and Ljubešić (2014) annotated the `SETimes.HR` dataset with linguistic and NER information, this dataset becoming the `set_hr` UD dataset for Croatian one year later (Agić and Ljubešić, 2015). Plank (2019) added a layer of NER to the dev and test portions of the Danish UD treebank (DDT) for cross-lingual evaluation; following up on this work, Plank et al. (2020) fully annotated the Danish UD dataset with nested NER entities. In a contemporary effort, Hvingelby et al. (2020) annotated the same Danish UD data with a flat annotation scheme for NER. In this work, we compare the UNER Danish DDT with the annotations by Hvingelby et al. (2020) to obtain inter-annotator statistics (see Table 2).

Other languages have seen efforts in a similar spirit. Jørgensen et al. (2020) added a named entity annotation layer on top of the Norwegian Dependency Treebank, Luoma et al. (2020) leveraged texts from the Turku Dependency Treebank (TDT) to build the Turku NER corpus, and Plank (2021) added a layer on top of English EWT. Most recently, Muischnek and Müürisep (2023) introduced the largest publicly available Estonian NER dataset annotated from the Estonian UD (EDT) and Web (EWT) treebanks. UNER builds upon this practice, though we provide much more comprehensive language coverage compared to earlier works.

Multilingual NER resources Several benchmark datasets for named entity recognition offer coverage for a variety of representative languages. Aside from well-known benchmarks such as CoNLL 2002/2003 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), some datasets were specifically tailored to address a unique need, such as focusing on low-resource languages like LORELEI (Strassel and Tracey, 2016) or incorporating particularly challenging annotations, as seen in MultiCoNER (Malmasi et al., 2022a,b). MasakhaNER (Adelani et al., 2022) harnessed the *Masakhane* community to produce gold-standard annotations for ten African languages.

Other datasets were constructed using a silver-standard process, often leveraging English Wikipedia as a primary resource. Expanding annotations to other languages involved entity-linking approaches, as seen in WikiNER (Nothman et al., 2013), or the utilization of knowledge bases such as WikiANN (Pan et al., 2017). In some instances, a combination of knowledge bases and neural methods was employed, as demonstrated by the approach taken in WikiNEuRal (Tedeschi et al., 2021). Nonetheless, CoNLL 2002/2003 remains one of the main benchmarks in multilingual NER. A recent work, also called UNER (Alves et al., 2020), attempts to produce silver-standard corpora by propagating English annotations across parallel corpora; however, there are no baseline evaluations to show the efficacy of this approach. Lastly, another contemporary work called Universal NER (Zhou et al., 2023) bears no relation to our effort, contains no annotation component, and is evaluated on an ensemble of NER benchmarks with different characteristics and quality levels.

Modeling for multilingual NER Several works have explored the task of NER outside of English. The earliest among them revolve around building language-independent methods that capture linguistic information in an unsupervised manner and then applying that knowledge downstream using a small amount of supervised training data (Cucerzan and Yarowsky, 1999; Lample et al., 2016, *inter alia*). Cross-lingual techniques have also emerged to transfer information between languages, especially from high- to low-resource languages (Ruder et al., 2019) or combining model and data transfer across languages (Wu et al., 2020b). Currently, the standard paradigm for multilingual NER involves finetuning or prompting multilingual language models (e.g., Wu and Dredze, 2020; Muennighoff et al., 2023).

The UNER project offers support to these modeling efforts by providing gold-standard annotations across a variety of languages. We also include experiments finetuning XLM-R_{Large} (Conneau et al., 2020) on UNER as a baseline reference point for future modeling work on this data.

Community-driven annotation projects The field of NLP has been shaped by community-driven annotation projects. One prime example is the Universal Dependencies (UD) project (Nivre et al., 2020), precipitated by the earlier introduction of the universal POS tagset (McDonald et al., 2013). The majority of the texts in UD and its predecessors were sourced from the contributions of several linguists and researchers. Extensions and sister projects to UD have emerged (e.g., Savary et al., 2023; Kahane et al., 2021), to which UNER is now added. Another notable endeavor is UniMorph (Kirov et al., 2018; McCarthy et al., 2020), which covers 182 languages (Batsuren et al., 2022) and supports an annual dedicated shared task, along with its offshoot MorphyNet (Batsuren et al., 2021). The Masakhane Project has also produced a number of high-quality community efforts, to name a few: MasakhaNER (Adelani et al., 2021, 2022), MasakhaPOS (Dione et al., 2023b), and MasakhaNEWS (Dione et al., 2023a).

The UNER project follows the same community-driven approach by asking volunteers to provide annotations for their respective languages.

7 Conclusion

We introduce Universal NER (UNER), a gold-standard data initiative covering 12 languages for named entity recognition (NER). The datasets included in UNER v1 cover a wide variety of domains and language families, and we establish initial performance metrics for these benchmarks. UNER opens a number of opportunities for research in NER outside of English and for cross-lingual transfer; in particular, this project provides human-annotated and standardized evaluations for multilingual NER.

After releasing the current version of the UNER project, we plan to expand language coverage and diversity of this effort by both recruiting additional annotators and integrating existing NER datasets when possible. This will also allow us to obtain more robust agreement measures and verify the quality of existing annotations in UNER. In the longer term, our aims for Universal NER include rigorous quality checking of annotation results for robustness and further integration of finetuned models and data analysis tools into the project.

Limitations

Dataset Domains and Languages The data included in UNER v1 covers a range of domains and languages, depending on the available annotators and datasets in UD (Appendix Table 5). The variance in domains and languages will generally affect the efficacy of in cross-lingual learning and evaluation. However, we also provide a standardized, parallel evaluation set for a subset of the languages in UNER. Furthermore, we invite researchers who would like to see additional languages in UNER to join the annotation effort.

Springboarding from Universal Dependencies Our preliminary criterion for languages and data to be included in the current version of UNER is that it should be already in the Universal Dependencies (UD) (de Marneffe et al., 2021). This is to ensure the quality of the underlying data and to facilitate research in conjunction with existing UD treebanks, which include part-of-speech tags, tokenization, lemmas, and glosses. However, the next iteration of the UNER initiative is open to all languages, especially low-resource ones, regardless of whether they are present in UD or not.

Number of Annotators The UNER project relies on crowd-sourcing and community participation for annotation efforts. Thus, the languages included have varying numbers of annotators who have accepted the invitation to contribute. Nonetheless, as reported in Table 2, each language has at least **two** annotators for a subset of its documents and thus a corresponding measure of inter-annotator agreement.

Ethics Statement

Our annotated data is built on top of an already-established resource which is Universal Dependencies. Thus, we do not foresee any serious or harmful issues that may arise from its content. Interested volunteer annotators who were invited to the project have also been informed of the guidelines as discussed in Section 3 for annotating NER-ready datasets before starting with the process.

Contributions

Stephen Mayhew conception, kickoff, all initial organization, recruitment, and annotation, development of annotation tool, manuscript writing.

Terra Blevins annotation, organization, PUD analysis scripts, core manuscript writing.

Shuheng Liu annotation, all baseline experiments and analysis.

Marek Šuppa annotation, PROPN analysis, paper writing, GPU resources.

Hila Gonen advising, organization of and feedback on manuscript.

Joseph Marvin Imperial facilitated annotations for Tagalog and Cebuano, additions to manuscript for TL/CEB results, limitations, ethics, and conclusion sections.

Börje F. Karlsson annotation, manuscript writing and editing, advising.

Peiqin Lin annotation, manuscript comments.

Nikola Ljubešić preparation and transfer of the HR SET and SR SET datasets, manuscript comments and edits.

LJ Miranda annotation, related work section, comments, edits.

Barbara Plank preparation and transfer of the DA DDT dataset, manuscript writing, comments and edits.

Yuval Pinter advising, organization of and writing of manuscript.

Acknowledgments

This project could not have happened without the enthusiastic response and hard work of many annotators in the NLP community, and for that we are extremely grateful.

We would like to thank Joakim Nivre, Dan Zeman, Matthew Honnibal, Željko Agić, Constantine Lignos, and Amir Zeldes for early discussion and helpful ideas at the very beginning of this project.

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.

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1724–1727, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 1–8, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Diego Alves, Tin Kuculo, Gabriel Amaral, Gaurish Thakkar, and Marko Tadic. 2020. UNER: Universal Named-Entity Recognition Framework. *arXiv preprint arXiv:2010.12406*.

Angelina Aquino, Franz de Leon, and Mary Ann Bacolod. 2020. UD_Tagalog-Ugnayan. https://github.com/UniversalDependencies/UD_Tagalog-Ugnayan.

Glyd Aranes. 2022. The GJA Cebuano Treebank: Creating a Cebuano Universal Dependencies Treebank. Master's thesis, Itä-Suomen yliopisto.

Zuyi Bao, Rui Huang, Chen Li, and Kenny Zhu. 2019. Low-resource sequence labeling via unsupervised multilingual contextualized representations. 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)*, pages 1028–1039, Hong Kong, China. Association for Computational Linguistics.

Vuk Batanović, Nikola Ljubešić, and Tanja Samardžić. 2018. Setimes.SR—A Reference Training Corpus of Serbian. In *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, pages 11–17.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023a. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Yang Chen, Vedaant Shah, and Alan Ritter. 2023b. Better Low-Resource Entity Recognition Through Translation and Annotation Fusion. *arXiv preprint arXiv:2305.13582*.

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.

Silviu Cucerzan and David Yarowsky. 1999. Language independent named entity recognition combining morphological and contextual evidence. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Cheikh M. Bamba Dione, David Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023a.

Masakhapos: Part-of-speech tagging for typologically diverse african languages.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023b. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A named entity resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

International Organization for Standardization. 1998. Codes for the representation of names of languages—Part 2: alpha-3 code. Standard, International Organization for Standardization, Geneva, CH.

International Organization for Standardization. 2002. Codes for the representation of names of languages—Part 1: Alpha-2 code. Standard, International Organization for Standardization, Geneva, CH.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal Dependencies for Danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

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.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 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*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

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 Twelfth Language Resources and Evaluation Conference*, pages 4615–4624, Marseille, France. European Language Resources Association.

Tingting Ma, Qianhui Wu, Huiqiang Jiang, Börje F. Karlsson, Tiejun Zhao, and Chin-Yew Lin. 2023. CoLaDa: A collaborative label denoising framework for cross-lingual named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Stephen Mayhew and Dan Roth. 2018. TALEN: Tool for annotation of low-resource ENtities. In *Proceedings of ACL 2018, System Demonstrations*, pages 80–86, Melbourne, Australia. Association for Computational Linguistics.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 51–58, Sydney, Australia.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Kadri Muischnek and Kaili Müürisep. 2023. Named entity layer in Estonian UD treebanks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 179–184, Tórshavn, Faroe Islands. University of Tartu Library.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

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.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Plank. 2019. Neural cross-lingual transfer and limited annotated data for named entity recognition in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 370–375, Turku, Finland. Linköping University Electronic Press.

Barbara Plank. 2021. Cross-lingual cross-domain nested named entity evaluation on English web texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1808–1815, Online. Association for Computational Linguistics.

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.

Peng Qi and Koichi Yasuoka. 2019. UD_Chinese-GSDSimp. https://github.com/UniversalDependencies/UD_Chinese-GSDSimp.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal Dependencies for Serbian in comparison with Croatian and other Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 39–44, Valencia, Spain. Association for Computational Linguistics.

Stephanie Samson and Cagrı Cöltekin. 2020. UD_Tagalog-TRG. https://github.com/UniversalDependencies/UD_Tagalog-TRG.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, and Abigail Walsh. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Mo Shen, Ryan McDonald, Daniel Zeman, and Peng Qi. 2016. UD_Chinese-GSD. https://github.com/UniversalDependencies/UD_Chinese-GSD.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

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.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020a. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Daniel Zeman. 2017. Slovak dependency treebank in universal dependencies. *Journal of Linguistics/Jazykovedný casopis*, 68(2):385–395.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv preprint arXiv:2308.03279*.

A Additional Dataset Details

In this section, we provide additional statistics and analysis of the datasets included in UNER v1. Table 5 documents the domains included in each dataset along with their distributions of NER tags, and Table 4 presents the F1 overlap score between named entities in UNER and PROPN tags in the underlying UD treebanks. We also report the full numerical results of our baseline experiments in Table 6.

Lang.	Dataset	Train	Dev	Test
DA	ddt	.824	.835	.836
EN	ewt	.813	.815	.817
HR	set	.810	.785	.799
PT	bosque	.844	.859	.856
SK	snk	.848	.783	.771
SR	set	.811	.769	.807
SV	talbanken	.830	.805	.839
ZH	gsd	.700	.696	.720
ZH	gsdsimp	.695	.695	.719
DE	pud	-	-	.785
EN	pud	-	-	.826
PT	pud	-	-	.805
RU	pud	-	-	.779
SV	pud	-	-	.877
ZH	pud	-	-	.708
CEB	gja	-	-	.926
TL	trg	-	-	.696
TL	ugnayan	-	-	.723

Table 4: Comparing PROPN overlap (F1 scores).

Data Source	Lang.	Dataset	Domains	Entity Dist. (%) LOC	ORG	PER
Johannsen et al. (2015)	DA	ddt	fiction, news, nonfiction, spoken	28.0	30.8	41.2
Silveira et al. (2014)	EN	ewt	blog, email, reviews, social, web	37.8	21.8	40.4
Agić and Ljubešić (2015)	HR	set	news, web, wiki	37.4	33.0	29.6
Rademaker et al. (2017)	PT	bosque	news	29.5	33.9	36.6
Zeman (2017)	SK	snk	fiction, news, nonfiction	21.2	6.2	72.6
Samardžić et al. (2017)	SR	set	news	41.4	30.2	28.4
McDonald et al. (2013)	SV	talbanken	news, nonfiction	54.0	20.0	25.0
Shen et al. (2016)	ZH	gsd	wiki	48.1	17.9	34.0
Qi and Yasuoka (2019)		gsdsimp	wiki	48.0	18.0	34.0
Zeman et al. (2017)	DE	pud	news, wiki	41.3	18.5	40.2
	EN	pud	news, wiki	39.5	21.9	38.6
	PT	pud	news, wiki	44.4	18.0	37.6
	RU	pud	news, wiki	33.4	26.6	40.0
	SV	pud	news, wiki	43.0	15.7	41.3
	ZH	pud	news, wiki	44.9	13.5	41.6
Aranes (2022)	CEB	gja	grammar examples	12.3	2.0	85.7
Samson and Cöltekin (2020)	TL	trg	grammar examples	10.9	0.0	89.1
Aquino et al. (2020)		ugnayan	fiction, nonfiction	47.5	0.0	52.5

Table 5: Domains and distribution of entity types for datasets in UNER. Domains are categorized for the underlying UD datasets at https://universaldependencies.org/.

Source		Target		F_1			
Lang.	Dataset	Lang.	Dataset	LOC	ORG	PER	Overall
DA	ddt	DA	ddt	.879	.828	.914	.826
EN	ewt	EN	ewt	.868	.721	.949	.859
HR	set	HR	set	.969	.898	.969	.936
SK	snk	SK	snk	.846	.635	.881	.854
PT	bosque	PT	bosque	.882	.861	.966	.904
SR	set	SR	set	.978	.905	.983	.946
SV	talbanken	SV	talbanken	.904	.741	.927	.882
ZH	gsd	ZH	gsd	.905	.818	.922	.895
ZH	gsdsimp	ZH	gsdsimp	.906	.811	.924	.893
EN	ewt	DE	pud	.816	.603	.893	.814
		EN	pud	.770	.583	.931	.801
		PT	pud	.844	.697	.913	.848
		RU	pud	.681	.451	.874	.714
		SV	pud	.887	.655	.927	.872
		ZH	pud	.465	.309	.388	.410
EN	ewt	CEB	gja	.556	.000	.842	.789
		TL	trg	.571	-	.353	.379
		TL	ugnayan	.951	-	.222	.680

Table 6: The full results of our baseline experiments from finetuning XLM-R_{Large} on UNER. All scores are reported in micro-F_1. ORG F_1 scores are not reported for the two TL datasets since there are no ORG entities labeled.