Towards Improved Distantly Supervised Multilingual Named-Entity Recognition for Tweets

Ramy Eskander, Shubhanshu Mishra, Sneha Mehta, Sofía Samaniego, Aria Haghighi
Twitter. Inc.
{ramid,smishra,snehamehta,ssamaniego,ahaghighi}@twitter.com

Abstract

Recent low-resource named-entity recognition (NER) work has shown impressive gains by leveraging a single multilingual model trained using distantly supervised data derived from cross-lingual knowledge bases. In this work, we investigate such approaches by leveraging Wikidata to build large-scale NER datasets of Tweets and propose two orthogonal improvements for low-resource NER in the Twitter social media domain: (1) leveraging domain-specific pre-training on Tweets; and (2) building a model for each language family rather than an all-in-one single multilingual model. For (1), we show that mBERT with Tweet pre-training outperforms the state-of-the-art multilingual transformer-based language model, LaBSE, by a relative increase of 34.6% in F1 when evaluated on Twitter data in a language-agnostic multilingual setting. For (2), we show that learning NER models for language families outperforms a single multilingual model by relative increases of 14.1%, 15.8% and 45.3% in F1 when utilizing mBERT, mBERT with Tweet pre-training and LaBSE, respectively. We conduct analyses and present examples for these observed improvements.

1 Introduction

Named-entity recognition (NER) is the process of detecting named mentions in text, and it is an essential subtask in several NLP applications such as information extraction (Weston et al., 2019), summarization (Aramaki et al., 2009) and question answering (Chen et al., 2019).

While resource-rich languages have received enormous focus over the last two decades, NER for low-resource languages is still under-explored due to the lack of resources — native speakers might not be even accessible — and the cost of labeling data needed to train supervised models for different languages. As a result, there has been emerging interest in multilingual NER, especially to process low-resource languages, in unsupervised and minimally supervised fashions.

One aspect of Multilingual NER is the need to build models that can generalize well across the underlying languages. However, when operating on social media text, multilingual NER becomes even harder (Mishra and Diesner, 2016; Mishra, 2019; Mishra and Haghighi, 2021) because of linguistic diversity, short context and orthographic variation.

Recent research has shown success by leveraging a single multilingual model based on distantly supervised datasets derived from cross-lingual knowledge bases (Nothman et al., 2013; Rahimi et al., 2019). We follow the work on building distantly supervised NER datasets by leveraging Wikidata (Vrandečić and Krötzsch, 2014) for Tweets, where we do not assume access to long contexts nor manually labeled named entities in context. We then propose modeling techniques towards improved multilingual NER models for Tweets, where we investigate how much pre-training language models on domain-specific data (Tweets) and training NER models on the basis of language families improve NER performance. Our contribution is threefold.

1. We build distantly supervised large-scale monolingual and multilingual NER datasets of Tweets [1].
2. We propose a domain-specific pre-trained Tweet language model.
3. We learn different NER models for language families versus a single all-in-one multilingual model.

It is worth noting that while exiting distantly supervised NER datasets have proven efficient, e.g., WikiAnn (Pan et al., 2017), they are either 1) monolingual; 2) based on resources of rich context such as Wikipedia, as opposed to Wikidata, where the named entities are out of context; 3) outside of

[1] The datasets are accessible upon contacting the first author.

Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), pages 115 - 124
December 8, 2022 ©2022 Association for Computational Linguistics

the Twitter domain; or 4) of limited size such as the Tweet datasets by Peng et al. (2019) and Liang et al. (2020). This necessitates the development of our Tweet datasets in order to answer our research questions in a low-resource setting.

We show that mBERT with Tweet pre-training outperforms LaBSE (Feng et al., 2020), a state-of-the-art multilingual language model, when evaluated in a language-agnostic multilingual setting on Twitter data. In addition, we show that learning NER models for language families outperforms a single all-in-one multilingual model. Our interpretation is that languages that belong to one family possess common linguistic features useful to learn an NER model. In contrast, joint learning of too many languages, most of which are unrelated, hinders the ability of the model to well fit any of the underlying languages. Finally, we conduct analyses and present examples in German and Arabic for the observed improvements.

2 Distantly Supervised Multilingual NER

In order to answer our research questions, we construct distantly supervised monolingual and multilingual NER datasets of Tweets (Section 2.1) and train NER models of different characteristics (Section 2.2).

2.1 Building NER datasets of Tweets

We describe below the process for building distantly supervised NER datasets of Tweets using Wikidata.

2.1.1 Initial selection of Tweets

First, we construct an initial corpus of Tweets that lay within a time window of 14 days [2], up to 5,000 Tweets per language on any single day. This results in Tweets in the 65 languages depicted in Figure 1. We then apply white-space tokenization on the selected Tweets.

2.1.2 Constructing a Wikidata Lookup

Utilizing cross-lingual knowledge bases to build multilingual NER datasets and gazetteers has proven successful (Pan et al., 2017; Al-Rfou et al., 2015). We next build a gazetteer of named entities by leveraging Wikidata (Vrandečić and Krötzsch, 2014), a large-scale cross-lingual knowledge base of nearly 100M entities, where each entity has a unique identifier and a list of categories and is defined as labels and alternate aliases in multiple languages.

For each language in our initial corpus of Tweets, we construct a Wikidata lookup trie (suffix tree) that stores all the labels and aliases of each entity in the underlying language. We apply white-space tokenization on the labels and aliases and store the resulting tokens in the tries, one token per level. We also store entity information, such as the identifier and the list of feasible categories, within the corresponding leaf nodes.

2.1.3 Tagging of Tweets

We apply the maximum matching algorithm used by Peng et al. (2019), with a context size $k = 5$, to tag our corpus of Tweets for NER. In order to speed up the search process, we scan the Wikidata lookup tries in a top-down fashion with early termination.

Marking all the matching Wikidata labels and aliases as named-entity mentions in the Tweets results in over-tagging. For instance, the common English word *be* is an alias for *Belgium* (LOCATION). Accordingly, we ignore unigram mentions, mentions exclusively composed of the most frequent 1,000 tokens in the underlying language[3] and mentions starting with a lower-cased letter (if different from its upper-cased form), which results in empirical improvements in precision.

2.1.4 Curation of Tags

Next, we map the Wikidata categories into NER labels and filter out the Tweets that do not contain mentions belonging to the main NER labels, namely PERSON, LOCATION and ORGANIZATION. Moreover, since the PERSON label is common in Tweets, we only select the Tweets that contain a single PERSON mention with a 20% probability. In addition, since a mention might belong to two or more categories, a Tweet is replicated to reflect all the possible combinations of the underlying labels. For instance, a Tweet that has the mention *Michael Kors* is replicated twice in order to indicate both the PERSON and ORGANIZATION interpretations [4].

2.1.5 Defining the Datasets

We build monolingual NER datasets for each language. In addition, we build multilingual datasets

[2] In order to avoid Tweets of insufficient context, we filter our Tweets that are replies, containing more than five hashtags, five mentions or three URLs, or containing less than five tokens.

[3] We derive the lists based on the initial corpus of Tweets.

[4] The replication results in better empirical performance, where the models learn to detect and overlook unlikely label assignments

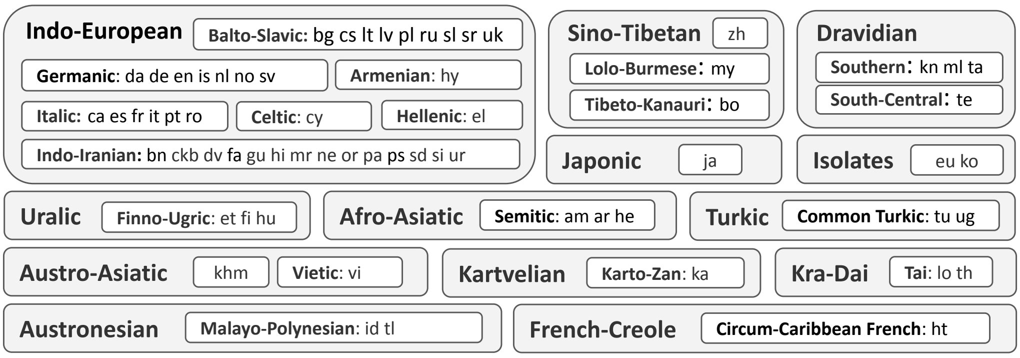

Figure 1: Our training languages, grouped into their families and sub-families

for language families, defined as the first and second language-family levels according to Wikipedia (See Figure 1). We do so for all the language families that include three or more languages and at least one experimental language (the first column in Table 1). This results in four family-based multilingual NER datasets, namely AAS (Afro-Asiatic, Semitic), IEG (Indo-European, Germanic), IEI (Indo-European, Italic) and IEII (Indo European, Indo-Iranian). Finally, we build a single all-in-one multilingual dataset that contains all the training languages.

In addition, we construct additional datasets that are the merge between our datasets and the training sets of WikiAnn (Pan et al., 2017), distantly supervised cross-lingual NER and entity-linking datasets of Wikipedia articles, towards higher coverage. The sizes of the datasets are reported in Table 1.

Family-Based Multilingual NER We hypothesize that a restricted multilingual model that is focused on languages within one family outperforms a multilingual model that spans two or more language families. This is because languages within one family tend to share morphosyntactic and syntactic features useful to learn an NER model, while learning a model across unrelated families limits the ability of the model to learn the latent patterns per language. Previous research highlights the role of family relatedness in different NLP tasks. Pires et al. (2019) show that fine-tuning mBERT on some language and applying zero-shot model transfer onto another only performs well across related languages in the tasks of NER and POS tagging. Cross-lingual POS tagging has also proven most successful across languages that belong to the same family (Eskander et al., 2020; Eskander, 2021). In addition, Fan et al. (2021a) show that selecting a pivot language within the same language family of the language of interest helps improve translation performance.

Lang/Family	Without WikiAnn	With WikiAnn
en	35K	55K
de	24K	44K
nl	30K	50K
es	22K	42K
pt	7K	27K
fr	19K	39K
it	24K	44K
hi	30K	35K
ur	77K	97K
bn	6K	16K
ja	25K	45K
ar	15K	35K
tr	12K	33K
te	6K	7K
AAS	36K	76K
IEG	112K	234K
IEI	106K	226K
IEII	149K	210K
All	609K	1425K

Table 1: The sizes of the monolingual and multilingual NER datasets. AAS = Afro-Asiatic, Semitic. IEG = Indo-European, Germanic. IEI = Indo-European, Italic. IEII = Indo-European, Indo-Iranian.

2.2 Modeling

We build our multilingual NER models by fine-tuning multilingual transformer-based language models, namely (basic) mBERT [5] (Devlin et al., 2019), mBERT pre-trained on Tweets (mBERT+Tweets) and LaBSE (Feng et al., 2020),

[5] While XLM-Roberta (Conneau et al., 2019) is superior to mBERT in the task of multilingual NER (Adelani et al., 2021), the use of mBERT is sufficient to draw conclusions on the use of the different multilingual settings, where our purpose is not to produce an NER system with the state-of-the-art results.

Language-Agnostic BERT Sentence Embedding [6]. We use the same setup proposed by Devlin et al. (2019), where we predict the NER tags only for the first subword of each token in a sequence.

Our choice of mBERT is used as a baseline, while the use of LaBSE is motivated by the fact that mBERT's transfer across languages can be improved by aligning embeddings of translations (Mishra and Haghighi, 2021), which is in line with the pre-training objective of LaBSE. Moreover, both mBERT and LaBSE have achieved success in the task of NER as demonstrated in the work by Pires et al. (2019) and Hakala and Pyysalo (2019), respectively.

The mBERT+Tweets model is basically the basic mBERT model pre-trained on Tweets (plain Tweet texts) for the masked language-modeling (MLM) objective. For pre-training, we use a dataset of 700M Tweets in 65 languages, randomly sampled using mBERT's methodology[7] that is based on exponentially smoothed language probabilities (S=0.7) to slightly increase the representation of low-resource languages. We initialize our model with mBERT weights and further train on the MLM objective. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $5e^{-5}$ and a weight decay of 0.01, along with a batch size of 2K and 800K training steps.

3 Evaluation and Analysis

3.1 Experimental Setup

Languages We perform our experiments on 14 simulated low-resource languages [8] of diverse typologies where we do not assume access to labeled data in the form of texts tagged for named entities. This consists of 10 Indo-European languages, namely English, German and Dutch (Germanic); Spanish, Portuguese, French and Italian (Italic); and Hindi, Bengali and Urdu (Indo-Iranian), in addition to Arabic (Afro-Asiatic, Semitic), Japanese (Japonic), Turkish (Turkic, Common-Turkic) and Telugu (Dravidian, South-Central).

Training We follow Devlin et al. (2019) for the training of our NER models by fine-tuning the multilingual transformer-based language models, namely mBERT, mBERT+Tweets and LaBSE, on our distantly-supervised NER datasets presented in Section 2.1.

We train monolingual NER models for each experimental language; we denote this setting by MONO. In addition, we train multilingual NER models for the language families defined in Section 2.1.5; we denote this family-based learning setting by FB-MULTI. Finally, we train a single multilingual model for the 65 languages in Figure 1; we denote this setting by ALL-MULTI.

We use the AdamW optimizer with a learning rate of $1e^{-5}$ and a weight decay of $1e^{-5}$, along with a batch size of 16 and up to 10 epochs with early stopping. We use 12 NVIDIA A100 GPUs, averaging nearly an hour of training per NER model.

Testing We utilize in-house gold standard test sets for English, Spanish, Portuguese, Arabic and Japanese, containing 3K, 2K, 10K, 10K and 2.3K Tweets, respectively [9]. In addition, we use seven public benchmarks, namely CoNLL'03 (Tjong Kim Sang and De Meulder, 2003) (for English and German), CoNLL'02 (Tjong Kim Sang and De Meulder, 2003) (for Dutch and Spanish), Europeana Newspapers (Neudecker, 2016) (for French), xLiMe [10] (for Italian), SSEA (Singh, 2008) (for Hindi, Urdu, Bengali and Telugu), Code-Switch'18-(validation) (Aguilar et al., 2018) (for Arabic) and JRC (Küçük et al., 2014) (for Turkish).

3.2 Evaluation

We refer to a combination of a test set and a learning setting as an experimental pair. For instance, {es: CONLL'03, FB-MULTI} means that we apply the family-based multilingual NER model that is trained on the Italic dataset on the Spanish CONLL'03 test set, while {tr: JRC, ALL-MULTI} means that we apply the multilingual NER model that is trained on our 65 languages on the Turkish JRC test set. We report all the results in entity-level micro-averaged F1.

It is worth mentioning that our target is to compare the different multilingual settings towards improved NER for Tweets. However, we do not assess the quality of our Tweet datasets with respect to existing distantly supervised ones. This is because, to

[6] We cannot pre-train LaBSE on Tweets since LaBSE is pre-trained for the translation-pair prediction (TPP) objective, which requires translation pairs that are not available for Tweets.

[7] https://github.com/google-research/bert/blob/master/multilingual.md

[8] While most of our experimental languages are not low-resource, we use them in a low-resource setting.

[9] We plan to make our in-house test sets publicly available upon publication.

[10] https://clarin.si/repository/xmlui/handle/11356/1078

Lang.	Dataset	Monolingual			Mulilingual (Family-Based)			Multilingual (All-in-One)		
		mBERT	mBERT+Tweets	LaBSE	mBERT	mBERT+Tweets	LaBSE	mBERT	mBERT+Tweets	LaBSE
en	CONLL'03	41.8	40.7	**43.1**	40.1	38.9	**42.9**	**37.9**	36.0	33.3
en	INH*	38.0	**43.2**	42.3	34.1	**42.5**	36.8	32.8	**38.6**	27.5
de	CONLL'03	44.9	42.0	**46.4**	42.3	40.9	**44.2**	38.1	**38.8**	29.0
nl	CONLL'02	44.5	43.3	**50.7**	**46.8**	43.6	42.2	**41.2**	35.8	25.2
es	CONLL'02	**31.2**	30.5	27.6	**31.5**	27.5	29.0	**29.0**	27.4	24.8
es	INH*	40.3	**41.8**	39.7	35.9	**39.0**	33.1	32.4	**37.2**	24.8
pt	INH*	33.0	**41.2**	38.1	29.1	**36.2**	26.3	27.6	**33.9**	18.5
fr	EuropeanaNP	**36.4**	35.4	34.4	**33.6**	31.3	29.7	**28.1**	26.8	22.0
it	xLiMe*	14.4	**17.7**	16.3	14.4	**18.9**	16.6	16.3	**19.3**	16.3
hi	SSEA	26.4	30.6	**33.7**	19.0	20.1	**29.4**	**19.1**	17.1	9.1
ur	SSEA	17.9	16.5	**20.5**	14.7	16.6	**19.6**	15.6	12.3	**15.8**
bn	SSEA	25.1	21.2	**45.3**	19.1	18.9	**36.8**	16.5	18.9	**19.3**
ar	Code-Switch'18*	26.8	**28.0**	27.6	23.4	25.5	**28.9**	21.9	**23.0**	23.0
ar	INH*	16.0	**20.4**	16.4	14.1	**20.7**	15.7	11.4	**16.2**	10.8
ja	INH	17.3	**23.9**	18.5	NA	NA	NA	17.2	**20.3**	15.1
tr	JRC*	31.5	**37.6**	31.2	NA	NA	NA	26.9	**32.1**	28.0
te	SSEA	13.0	10.8	**17.6**	NA	NA	NA	12.0	6.6	**18.0**
Average (Tweets)		27.2	**31.7**	28.7	25.2	**30.5**	26.2	23.3	**27.6**	20.5
Average (IEG)		42.3	42.3	**45.6**	40.8	41.5	**41.5**	**37.5**	37.3	28.8
Average (IEI)		31.1	**33.3**	31.2	28.9	**30.6**	26.9	26.7	**28.9**	21.3
Average (IEII)		23.1	22.8	**33.2**	17.6	18.5	**28.6**	**17.1**	16.1	14.7
Average (All)		29.3	30.9	**32.3**	28.4	30.0	**30.8**	24.9	**25.9**	21.2

Table 2: NER Results (entity-level micro-averaged F1) without the addition of the WikiAnn training sets. The best result per experimental pair ({test set, learning setting}) is in **bold**. The best result per test set is underlined. Tweet datasets are denoted by *. IEG = Indo-European, Germanic. IEI = Indo-European, Italic. IEII = Indo-European, Indo-Iranian.

our knowledge, our datasets are the only available large-scale NER Tweet datasets that are based on a non-contextual knowledge base, Wikidata, where we simulate learning in truly low-resource scenarios.

Table 2 reports the NER performance (entity-level micro-averaged F1) for all the experimental pairs without the addition of the WikiAnn training sets. Overall, there is a noticeable variance in the performance of the different models across the learning settings, and even within the same language when evaluated on different test sets. However, the Germanic languages witness the best NER performance, which we attribute due to the bias in the training data of the utilized language models.

LaBSE The use of LaBSE in the MONO setting yields the best performance for seven experimental pairs: three Germanic ones, three Indo-Iranian ones and the telugu one. It also results in the best on-average F1 of 32.3% across all the experimental pairs in the MONO setting, which is relative increases of 10.2% and 4.5% over the corresponding performance of mBERT and mBERT+Tweets, respectively. However, the performance of LaBSE dramatically drops in the ALL-MULTI setting with average relative decreases of 34.4% and 31.2% compared to the performance in the MONO and FB-MULTI settings, respectively.

mBERT+Tweets The use of mBERT+Tweets in the MONO setting results in the best performance for eight experimental pairs, mostly with the use of our gold standards of Tweets (INH). In addition, when averaging across the Tweet datasets, mBERT+Tweets outperforms both mBERT and LaBSE, where it achieves relative increases of 10.5%, 16.4% and 34.6% compared to LaBSE in the MONO, FB-MULTI and ALL-MULTI settings, respectively. Moreover, mBERT+Tweets yields the best on-average performance in the ALL-MULTI setting, outperforming mBERT and LaBSE by average relative increases of 4.0% and 22.2%, respectively.

mBERT The results illustrate the effectiveness of pre-training the basic mBERT model, where mBERT+Tweets outperforms mBERT by average relative increases of 5.5%, 5.6% and 4.0% in the MONO, FB-MULTI and ALL-MULTI settings, respectively, while LaBSE outperforms mBERT by relative increases of 10.2% and 8.5% in the MONO and FB-MULTI settings, respectively. However, pre-training does not yield improvements in the cases of{fr, EuropeanaNP} and {es, CoNLL02}.

Monolingual vs. Multilingual NER Models The MONO setting yields the best performance for all the experimental pairs except five, two of

Lang.	Dataset	Monolingual			Mulilingual (Family-Based)			Multilingual (All-in-One)		
		mBERT	mBERT+Tweets	LaBSE	mBERT	mBERT+Tweets	LaBSE	mBERT	mBERT+Tweets	LaBSE
en	CoNLL'03	60.1	61.3	**62.9**	56.4	60.4	**62.6**	55.8	56.8	**57.6**
en	INH*	40.8	**48.2**	45.5	27.5	**43.3**	40.8	31.8	**37.3**	34.6
de	CoNLL'03	49.9	**54.8**	53.4	54.9	53.0	**55.2**	49.8	**54.9**	52.2
nl	CoNLL'02	**57.8**	51.8	53.3	47.9	46.2	**49.5**	45.3	**46.2**	45.0
es	CoNLL'02	**51.9**	46.1	48.5	**53.8**	53.0	46.3	**50.4**	49.9	45.3
es	INH*	40.2	40.9	**42.0**	32.7	**39.0**	31.4	29.5	30.8	**32.6**
pt	INH*	33.8	**41.4**	39.7	26.5	**35.4**	24.0	21.4	**27.2**	26.1
fr	EuropeanaNP	**45.1**	38.7	38.4	35.1	**35.2**	32.5	32.2	**35.9**	34.2
it	xLiMe*	13.7	17.3	**19.1**	16.2	**17.5**	15.5	14.5	**15.9**	15.2
hi	SSEA	22.9	23.8	**36.4**	19.5	**28.5**	28.0	22.3	24.2	**27.3**
bn	SSEA	25.6	20.2	**38.3**	20.4	18.4	**39.3**	21.1	20.3	**35.3**
ur	SSEA	22.9	20.7	**28.7**	28.3	27.2	**40.2**	30.7	29.1	**41.8**
ar	Code-Switch'18*	29.8	31.1	**33.1**	24.9	27.6	**29.3**	25.7	28.1	**29.8**
ar	INH*	16.0	**22.3**	21.9	12.1	**20.8**	16.9	12.8	14.4	**14.7**
ja	INH*	22.1	**24.9**	22.4	NA	NA	NA	18.8	**22.2**	22.0
tr	JRC*	38.5	**52.5**	46.2	NA	NA	NA	30.3	**42.9**	40.6
te	SSEA	**17.8**	6.4	16.8	NA	NA	NA	10.6	8.8	**16.3**
Average (Tweets)		29.4	**34.8**	33.8	23.3	**30.6**	26.3	23.1	**27.3**	26.9
Average (IEG)		52.1	**54.0**	53.8	46.7	50.7	**52.1**	45.7	**48.8**	47.4
Average (IEI)		37.0	36.9	**37.6**	32.8	**36.0**	29.9	29.6	**31.9**	30.7
Average (IEII)		23.8	21.5	**34.5**	22.7	24.7	**35.9**	24.7	24.5	**34.8**
Average (All)		34.7	35.4	**38.0**	32.6	36.1	**36.6**	29.6	32.1	**33.6**

Table 3: NER Results (entity-level micro-averaged F1) with the addition of the WikiAnn training sets. The best result per experimental pair ({test set, learning setting}) is in **bold**. The best result per test set is underlined. Tweet datasets are denoted by *. IEG = Indo-European, Germanic. IEI = Indo-European, Italic. IEII = Indo-European, Indo-Iranian.

which belong to Arabic and one of which belong to Telugu, the language with the least number of instances in our training sets. We hypothesize that for low-resource languages, adding training examples from other languages compensates for the lack of data in the language of interest.

Family-Based vs. All-in-One Multilingual Models Learning NER models for language families (FB-MULTI) outperforms the use of a single all-in-one multilingual model (ALL-MULTI) except on four occasions (7.8% of the time). FB-MULTI also outperforms ALL-MULTI when averaging across all the experimental pairs, yielding relative increases of 14.1%, 15.8% and 45.3% with the use of mBERT, mBERT+Tweets and LaBSE, respectively. FB-MULTI is also superior when averaging across the individual language families. The results suggest that combining too many languages in the training data makes it difficult for the NER model to learn the morphosyntactic and syntactic properties of the individual languages; empirically, the ALL-MULTI setting only yields the best performance for two experimental pairs by a small margin of 0.4% compared to the performance in the other learning settings. In contrast, languages within a language family tend to share linguistic properties, which helps the NER model better fit to the individual languages within the family.

WikiAnn Table 3 reports the NER performance (entity-level micro-averaged F1) for all the experimental pairs with the addition of the WikiAnn training sets. Comparing the results in Table 3 to those in Table 2 shows that the addition of WikiAnn helps derive more efficient NER models.

Grouping by individual languages, WikiAnn improves the performance for all languages excpet German, Portuguese and Italian, where Urdu benefits the most from the addition of WikiAnn, an average relative increase of 83.6%, while the biggest drop in performance occurs in the case of Italian, an average relative decrease of only 3.0%.

WikiAnn also improves the performance on average across the Germanic and Italic languages and when averaging across all the experimental languages. However, the addition of WikiAnn results in noticeable performance drop when considering the Tweet datasets in the case of fine-tuing mBERT in the FB-MULTI setting, where neither mBERT nor WikiAnn leverages Twitter data.

3.3 Analysis

Table 4 lists NER-tagging examples that show cases in which 1) mBERT with Tweet pre-training outperforms LaBSE; and 2) training for distinct language families outperforms the single all-in-one multilingual model. In addition, we show common errors in our best setting. We conduct our manual

German Examples																
Tokens		Andersens	starrer	Blick	sagt	viele	Worte	|	Das	ist	modernes	Marketing	für	den	Messia	
Glosses		Andersens	staring	look	says	many	words	|	this	is	modern	marketing	for	the	Messiah	
True Labels	01	B-PER	O	O	O	O	O	|	O	O	O	O	O	O	B-PER	02
mBERT+Tweets (ALL)		B-PER	O	O	O	O	O	|	O	O	O	O	O	O	B-PER	
LaBSE (ALL)		B-PER	O	B-PER	O	O	O	|	O	B-ORG	I-ORG	O	O	O	B-PER	
Tokens		Stepanovic	prophezeit	Wolf	eine	große	Zukunft	|	Der	Stadt	Königstein	geht	es	finanziell	glänzend	
Glosses		Stepanovic	prophesized	Wolf	a	great	future	|	the	city	Konigstein	goes	it	financial	brilliantly	
True Labels	03	B-PER	O	B-PER	O	O	O	|	O	B-LOC	I-LOC	O	O	O	O	04
mBERT+Tweets (FB)		B-PER	O	B-PER	O	O	O	|	O	B-LOC	I-LOC	O	O	O	O	
mBERT+Tweets (ALL)		O	O	B-PER	O	O	O	|	O	O	B-ORG	O	O	O	O	
Tokens		Eine	lange	Schlange	steht	vor	der	Bühne	|	Eröffnung	ist	um	11	Uhr		
Glosses		a	long	queue	stands	in front of	the	stage	|	opening	is	at	11	O’clock		
True Labels	05	O	O	O	O	O	O	O	|	O	O	O	O	O		06
mBERT+Tweets (FB)		B-PER	O	O	O	O	O	O	|	B-ORG	I-ORG	O	O	O		
Arabic Examples (Arabic reads from right to left)																
Tokens			الوسط	خط	في	النهاردة	زيزو	العزيز	عبد	محمود	|	الاسيويه	والامه	كوريا	فخر	
Glosses			the middle	line	in	today	Zizo	Al-Aziz	Abd	Mahmoud	|	the Asian	and the nation	Korea	pride	
True Labels	07		O	O	O	O	I-PER	I-PER	I-PER	B-PER	|	O	O	B-LOC	O	08
mBERT+Tweets (ALL)			O	O	O	O	I-PER	I-PER	I-PER	B-PER	|	O	O	B-LOC	O	
LaBSE(ALL)			O	O	O	O	O	I-PER	I-PER	B-PER	|	O	O	B-PER	O	
Tokens		اليوم	الأمن	مجلس	أمام	السوداني	الملف	|	دينار	١١٠	ب	يتبرع	الشمري	العوام	حمد	
Glosses		today	the Security	Council	before	the Sudanese	the case	|	Dinar	110	for	donates	Al-Shamry	Al-Awam	Hamad	
True Labels	09	O	B-ORG	I-ORG	O	O	O	|	O	O	O	O	I-PER	I-PER	B-PER	10
mBERT+Tweets (FB)		O	B-ORG	I-ORG	O	O	O	|	O	O	O	O	I-PER	I-PER	B-PER	
mBERT+Tweets (ALL)		O	O	O	O	O	O	|	O	O	O	O	O	I-PER	B-PER	
Tokens					الرياض	على	خفيفه	أمطار	|	جديد	عبدالنور	سيرين	مكتشفين	شكلهم	الشباب	
Glosses	11				Riyadh	on	light	rains	|	recently	Abd-Al-Nour	Cyrine	discovering	seem	the youth	12
True Labels					B-LOC	O	O	O	|	O	I-PER	B-PER	O	O	O	
mBERT+Tweets (FB)					O	I-PER	O	O	|	O	B-PER	O	O	O	O	

Table 4: NER Examples in German and Arabic. Errors are circled.

analysis on both German and Arabic [11] using the CONLL’03 and INH test sets, respectively.

German The use of mBERT+Tweets in the ALL-MULTI setting results in 1,335 (out of 3K) correctly tagged Tweets, as opposed to 495 when leveraging LaBSE, where the use of LaBSE results in over-tagging PERSON (ex. 01) and ORGANIZATION (ex. 02). On the other hand, the number of correctly tagged Tweets increases to 1,418 when fine-tuning mBERT+Tweets for the IEG family, where the system improves at detecting PERSON (ex. 03) and LOCATION (ex. 04). However, one common error is the false tagging of PERSON (ex. 05) and ORGANIZATION (ex. 06) at the beginning of Tweets.

Arabic The use of mBERT+Tweets in the ALL-MULTI setting results in 4,805 (out of 10K) correctly tagged Tweets, as opposed to 1,216 when leveraging LaBSE, where the use of LaBSE weakens the detection of non-PERSON mentions (ex. 07) and long mentions of three or more tokens (ex. 08). On the other hand, the number of correctly tagged Tweets increases to 6,229 when fine-tuning mBERT+Tweets for the AAS family as the system further improves at tagging non-PERSON mentions (ex. 09) and long mentions (ex. 10). However, two common issues are the low recall of LOCATION (ex. 11) and the inability to recognize non-Arabic and infrequent Arabic names (ex. 12).

4 Related Work

Leveraging cross-lingual knowledge bases for the construction of multilingual NER datasets and gazetteers has proved successful. Two large-scale efforts are WikiAnn (Pan et al., 2017), Wikipedia-based cross-lingual NER and entity-linking datasets in 282 languages, and Polyglot-NER (Al-Rfou et al., 2015), NER datasets in 40 languages derived from Wikipedia and Freebase (Bollacker et al., 2008). On another hand, there have been a few efforts to construct distantly supervised NER datasets of Tweets such as the work by Peng et al. (2019) and Liang et al. (2020), which presented datasets of only 7,257 and 2,400 Tweets, respectively. We follow similar approaches by leveraging Wikidata (Vrandečić and Krötzsch, 2014) to construct large-scale monolingual and multilingual NER datasets of Tweets.

Fine-tuning transformer-based language models for NER has shown success. Several works have utilized mBERT (Devlin et al., 2019) to construct generic and domain-specific multilingual NER models (Pires et al., 2019; Arkhipov et al., 2019; Baumann, 2019). Another example is LaBSE (Feng et al., 2020). While mostly utilized for sentence-level NLP tasks such as hate-speech identification (Mandl et al., 2021) and claim matching (Kazemi et al., 2021), LaBSE has also proven efficient for NER (Hakala and Pyysalo, 2019). In

[11] We have access to linguists who understand German and Arabic. Moreover, the two languages represent two different families and scripts.

this work, we fine-tune both mBERT and LaBSE for NER in the Twitter domain, where we learn and compare monolingual and multilingual models of different characteristics.

Gururangan et al. (2020) shows that pre-training transformers towards a specific task or domains can provide significant benefits. Mishra and Haghighi (2021) show that pre-training mBERT for the translation-pair prediction (TPP) objective improves NER. Pre-training mBERT on Tweets has also been successful for a number of individual languages, such as English (Nguyen et al., 2020) and Arabic (Ahmed Abdelali et al., 2021). In this work, we pre-train mBERT on Tweets in 65 languages.

Several recent works utilize language classification towards improved multilingual models. The clustering can be based on either 1) language embeddings (Kudugunta et al., 2019; Tan et al., 2019; Yu et al., 2021; Fan et al., 2021b); 2) language family with/without the use of hand-crafted rules such as geographical proximity (Tan et al., 2019; Fan et al., 2021a); and 3) token overlap (Chung et al., 2020). We perform family-based clustering for NER, similar to the first approach proposed by Tan et al. (2019) in the task of machine translation. However, we do not assume access to rich embeddings or linguistic knowledge for the language(s) of interest.

5 Conclusion and Future Work

We proposed improvements to distantly supervised multilingual NER for Tweets, where we leveraged Wikidata to build large-scale monolingual and multilingual NER datasets of Tweets. We showed that pre-training mBERT on Tweets outperforms LaBSE by a relative F1 increase of 34.6% when evaluated on Twitter data in a language-agnostic multilingual setting. We also showed that learning NER models for language families outperforms a single all-in-one multilingual model by relative F1 increases of at least 14.1%. In the future, we plan to produce larger Tweet pre-trained language models, study more language families and leverage the work for multilingual entity linking for Tweets in low-resource languages.

6 Limitations

The limitations of the work lay within the Twitter social media domain for the listed training languages and given the reported performance. Also, the datasets are not labeled for named entities that are not included in Wikidata. The models however can generalize well to discover unseen named entities. Another limitation is the lack of a gold standard to intrinsically assess the quality of the labels in our NER datasets. There should be no other potential risks given the stated limitations.

7 Ethical Considerations

We exploit Twitter API [12] for the extraction of Tweets, along with language detection. The datasets are accessible upon contacting the first author. We however replace the text of the Tweets by Tweet IDs in order to prevent sensitive information and negative content, in accordance with Twitter's policy for sharing data. In addition, we are committed to keep the datasets current, making sure that deleted Tweets are removed from the datasets when they become publicly available.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg. 2018. Proceedings of the third workshop on computational approaches to linguistic code-switching. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*.

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *CoRR*.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, and Kazuhiko Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*, pages 185–192.

Mikhail Arkhipov, Maria Trofimova, Yurii Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recogni-

[12] https://developer.twitter.com/en/docs/twitter-api

tion. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93.

Antonia Baumann. 2019. Multilingual language models for named entity recognition in german and english. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 21–27.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Human Language Technologies: The 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ramy Eskander. 2021. *Unsupervised Morphological Segmentation and Part-of-Speech Tagging for Low-Resource Scenarios*. Ph.D. thesis, Columbia University.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021a. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Yimin Fan, Yaobo Liang, Alexandre Muzio, Hany Hassan, Houqiang Li, Ming Zhou, and Nan Duan. 2021b. Discovering representation sprachbund for multilingual pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 881–894, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual bert. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott A Hale. 2021. Claim matching beyond english to scale global fact-checking. *arXiv preprint arXiv:2106.00853*.

Dilek Küçük, Guillaume Jacquet, and Ralf Steinberger. 2014. Named entity recognition on turkish tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 450–454.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schaefer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Shubhanshu Mishra. 2019. Multi-dataset-multi-task Neural Sequence Tagging for Information Extraction from Tweets. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media - HT '19*, pages 283–284, New York, New York, USA. ACM Press.

Shubhanshu Mishra and Jana Diesner. 2016. Semi-supervised Named Entity Recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 203–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Shubhanshu Mishra and Aria Haghighi. 2021. Improved multilingual language model pretraining for social media text via translation pair prediction. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 381–388, Online. Association for Computational Linguistics.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

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.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. *arXiv preprint arXiv:1906.01378*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

Anil Kumar Singh. 2008. Named entity recognition for south and south east asian languages: taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. 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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702.

Dian Yu, Taiqi He, and Kenji Sagae. 2021. Language embeddings for typology and cross-lingual transfer learning. 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 7210–7225, Online. Association for Computational Linguistics.