Zero-Shot Cross-Lingual NER Using Phonemic Representations for Low-Resource Languages

Anonymous ACL submission

Abstract

Existing zero-shot cross-lingual NER approaches require substantial prior knowledge of the target language, which is impractical for low-resource languages. In this paper, we propose a novel approach to NER using phonemic representation based on the International Phonetic Alphabet (IPA) to bridge the gap between representations of different languages. Our experiments show that our method significantly outperforms baseline models in extremely low-resource languages, with the highest average F-1 score (46.38%) and lowest standard deviation (12.67), particularly demonstrating its robustness with non-Latin scripts.

1 Introduction

Named entity recognition (NER) plays a crucial role in many Natural Language Processing (NLP) tasks. Achieving high performance in NER generally requires extensive resources for both sequence labeling and gazetteer training (Das et al., 2017). However, access to training resources for many low-resource languages (LRLs) is very limited, motivating zero-shot approaches to the task. While various strategies have been explored to enhance zero-shot NER performance across languages, they required either parallel data or unlabeled corpora in the target language, which is difficult and sometimes impossible to obtain.

Our work tackles zero-shot NER under a strict condition that disallows any target language training data. We decided to approach this condition by projecting data into an International Phonetic Alphabet (IPA) space. Since different languages often share similar pronunciations for the same entities, such as geopolitical entities and personal names (e.g., the word for China is /t͡ʃajnə/ in English and /tʃina/ in Sinhala), the model trained on one language can be transferred to others without target-language training in NER. As shown in Figure 1, we first convert orthographic scripts into IPA, and then fine-tune a pre-trained model on the phonemes of the source language, i.e., English. By using a shared notation system—IPA—we can apply the model to target languages directly. Our findings show that fine-tuning phoneme-based models outperforms traditional grapheme-based models(e.g., mBERT (Devlin et al., 2019)) by a large margin for LRLs not seen during pre-training. Furthermore, our approach demonstrates robustness with non-Latin scripts, exhibiting stable performance across languages with different writing systems.

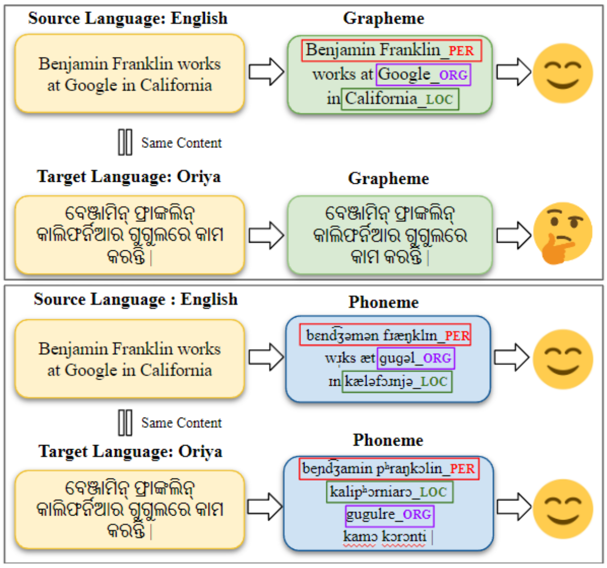

Figure 1: Zero-shot Cross-Lingual NER with IPA phonemes.

2 Related Work

2.1 Zero-shot Cross-lingual NER

Recent approaches for zero-shot cross-lingual NER can be categorized into three groups based on how they use resources from target languages. One line of work involves using translation between source and target languages to transfer NER capability (Yang et al., 2022; Liu et al., 2021; Mo et al., 2024). These methods require parallel data from both languages, which is not always available. Alternatively, some methods use unlabeled target language data and adopt knowledge distillation without needing parallel data (Deb et al., 2023; Li et al., 2022). However, these approaches are still not widely applicable to languages with extremely low-resources, as such languages often lack sufficient resources for training. On the other hand, (Rathore et al., 2023) assumes that no data in target language is available during training. While it provides a practical setting for extremely low-resource languages, it requires language adapters pre-trained on similar languages to the target language, as well as typological information (i.e., language family) of various languages.

We assume a very strict problem setting where the target language for zero-shot inference, as well as its typological information, is completely unavailable during training. Unlike previous methods that rely on some of the target language data during training, we use IPA phonemes for NER, making our method entirely data-independent for the target language. It only relies on the availability of an easily constructed grapheme-to-phoneme (G2P) module.

2.2 Phonemic Representation

Phonological traits of languages are useful in understanding different languages, as they often share similar pronunciations for similar entities. It is particularly beneficial for NER, where many items, such as geopolitical entities and personal names, are pronounced similarly across various languages. While phonological information has been shown to be helpful in language understanding for cross-lingual transfer (Chaudhary et al., 2018; Sun et al., 2021; Bharadwaj et al., 2016), it has not been explored as a standalone representation for NER, especially on low-resource languages. Given that creating rule-based transcription module for most low-resource languages takes only a few hours and limited training, we use IPA to enable zero-shot cross-lingual NER on languages with very scarce resources, without requiring any additional corpus for those languages.

3 Our Approach

3.1 NER with Phonemes

In this paper, we conduct NER using phonetic transcriptions (IPA) instead of conventional orthographic text. Leveraging the standard practice of using multilingual pre-trained models for cross-lingual transfer, we employ XPhoneBERT (Nguyen et al., 2023), a model pre-trained on phonemes from 94 different languages. By utilizing pre-trained phonemic representations, the model can fully utilize the phonological knowledge across diverse languages.

To create a phoneme-based version of the dataset originally containing graphemes, we convert the dataset into IPA representations. For G2P conversion of various languages, we use Epitran (Mortensen et al., 2018) along with the CharsiuG2P toolkit (Zhu et al., 2022) which XPhoneBERT originally employed. Epitran supports the transliteration of approximately 100 languages, including numerous low-resource languages. We apply transliteration at the word level, maintaining the pre-tokenized units consistent with the original version.

We adopt the BIO tagging scheme for entity tagging. As the phoneme is the input unit for the model, we assign each phoneme a named entity tag. Only the first phoneme segment of the first word of a named entity is assigned with a ‘B’ tag, indicating the beginning of the entity. For example, the phoneme sequence “bɛnd͡ʒəmən (Benjamin)” comprises nine segments[1], and is labeled as [“B-PER”, “I-PER”, ...,“I-PER”].

3.2 Cross-lingual Transfer to Unseen Languages

We perform zero-shot named entity recognition on low-resource languages, where the model is only trained on a single high-resource language, in this case, English. Although the model is fine-tuned on a single language, its pre-training on approximately 100 languages allows it to retain some knowledge of other languages. We hypothesize that (i) each model will leverage its pre-trained knowledge on the target languages in performing NER, and (ii) phoneme-based models will generally achieve superior performance with unseen languages, benefiting from phonological traits shared across languages.

To investigate the generalizability of phonemic representations in extremely low-resource languages, we do not allow any access to the target language during training and exclude their typological information to keep our method language-agnostic.

[1]Phoneme segmentation is performed using the Python library ‘segments,’ as utilized in XPhoneBERT.

Case	Models			Languages	Num
	M	C	X		
1	-	-	-	sin, som, mri, quy, uig, aii, kin, ilo	8
2	-	-	✓	epo, khm, tuk, amh, mlt, ori, san, ina, grn, bel, kur, snd	12
3	✓	✓	-	tgk, yor, mar, jav, urd, msa, ceb, hrv, mal, tel, uzb, pan, kir	13

Table 1: Languages for each case. M, C, X indicates mBERT, CANINE, and XPhoneBERT, respectively, and ✓ represents the languages pre-trained on the model.

We use mBERT and CANINE as baselines, as these models are compatible with our problem setting, requiring no additional training data for the target languages.

As shown in Table 1, we define three sets of languages based on whether the language has been seen during pre-training of each model. Let L be the set of all languages in our benchmark dataset that are able to be transliterated, B the set of languages pre-trained on the baseline models, and X the set of languages pre-trained on XPhoneBERT.
Case 1: $(L \setminus (B \cup X))$ includes languages not in the pre-training data for any models.
Case 2: $((L \cap X) \setminus B)$ includes languages in the pre-training data of XPhoneBERT only.
Case 3: $((L \cap B) \setminus X)$ includes languages in the pre-training data of mBERT and CANINE only.

4 Experiments

4.1 Benchmark Dataset

We train and evaluate our method on the WikiANN NER datasets (Pan et al., 2017) which has three different named entity types: person (PER), organization (ORG), and location (LOC). The models are trained only on English data and evaluated on various low-resource languages. We select languages that are (i) supported by either Epitran or CharsiuG2P toolkit for transliteration, and (ii) not included in the pre-training of at least one of the baseline models. This yields 33 languages in total, as listed in Table 1.

4.2 Baseline Models

We use mBERT (Devlin et al., 2019) and CANINE (Clark et al., 2022), both grapheme-based language models, as baselines to compare to XPhoneBERT (Nguyen et al., 2023), a phoneme-based language model. All three models are BERT-like transformer architectures pre-trained on a Wikipedia corpora of multiple languages: mBERT and CANINE are trained on the same 104 languages, while XPhoneBERT is trained on 94 languages and locales. Initializing with pre-trained weights from Huggingface[2], we train the encoders with a fully connected layer added at the end of each encoder for NER prediction.

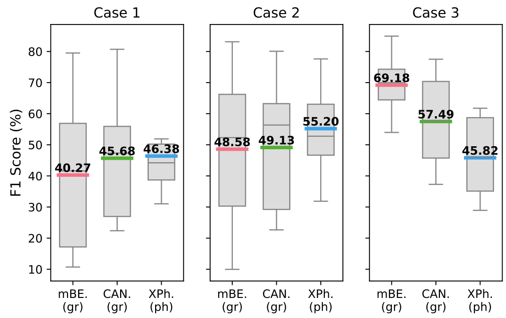

Figure 2: Distribution of F1 scores for each language set. X-axis shows each model using their first three letters, with '(gr)' and '(ph)' indicating their input forms (graphemes and phonemes, respectively). Colored horizontal lines and the numbers above show the average F1 scores for each model.

5 Results

5.1 Zero-Shot NER on Seen Languages

Figure 2 illustrates zero-shot performance of each model for each language set (**Case 1**, **Case 2**, and **Case 3**). Results on **Case 2** and **Case 3** align with our expectation, with languages seen during pre-training achieving better scores with the model. For the 12 languages in **Case 2**, XPhoneBERT, which was pre-trained on these languages, shows an average F1 score of 55.20%, outperforming mBERT and CANINE by 6.62% and 6.07%, respectively. Languages of **Case 3** also performs better with models that were pre-trained on these languages. Specifically, mBERT achieves high scores for pre-trained languages, with average F1 score of 69.18%, indicating its strong ability to generalize across seen languages. F1 scores for all models and languages are shown in Table 3 of Appendix.

5.2 Zero-Shot NER on Unseen Languages

Given the performance bias towards seen languages, we investigate the effect of using phonemes with languages that were not seen by any model—languages from **Case 1**. This ensures a fair comparison for low-resource languages, since extremely low-resource languages are often not included in the pre-training stage of language models. As shown in Table 2, the phoneme-based model

[2]https://huggingface.co/

Input	Model	Languages								AVG	STD
		sin	som	mri	quy	uig	aii	kin	ilo		
grapheme	mBERT	10.71	44.76	38.48	55.07	18.70	12.58	62.37	79.51	40.27	25.00
grapheme	CANINE	26.31	43.35	51.30	59.48	27.19	22.38	54.74	80.70	45.68	19.99
phoneme (ours)	XPhoneBERT	**43.61**	38.91	38.07	51.90	**44.82**	**31.03**	49.67	73.05	**46.38**	**12.67**

Table 2: Zero-shot performance in F1 scores (%) on unseen languages (**Case 1**) using different models and input types.

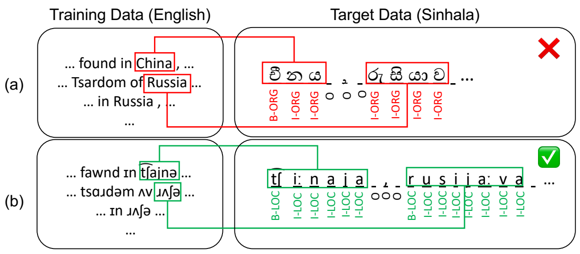

Figure 3: NER results on the target language (Sinhala) produced by each model trained on English data: (a) CANINE (b) XPhoneBERT.

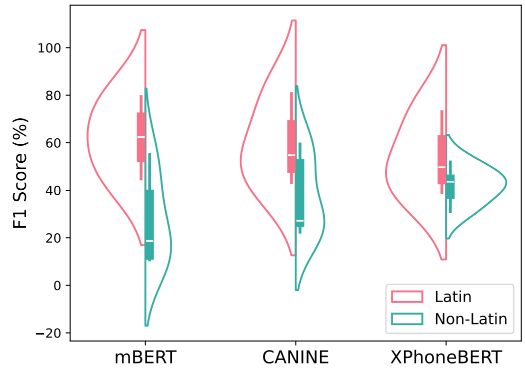

Figure 4: Performance distribution of each model on Latin and non-Latin languages from unseen languages.

demonstrates the best overall performance, achieving the highest scores on 3 out of 8 languages by a significant margin. Furthermore, the phoneme-based model exhibits the most stable performance across unseen languages, with the lowest standard deviation in scores.

Figure 3 shows a qualitative result of zero-shot inference on Sinhala, a language that is not in the pre-training data any model. While the character-based model (a) fails to generalize to the language with different writing system, the phoneme-based model (b) successfully predicts the named entity tags due to the similar pronunciation of "China" and "Russia" across the languages. These results indicate the robustness provided by phonemic representations, validating our hypothesis about the advantages they convey in NER tasks.

5.3 Robustness Across Writing Systems

One of the important advantages of using phonemic representations for named entity recognition is that it allows use of IPA. Using IPA for multilingual tasks provides a unified notation system. Observing the significant performance drop of mBERT on unseen low-resource languages (Figure 2), we consider this gap is largely attributed to the different writing systems of languages. Figure 4 shows the distribution of F1 scores of each model on Latin and non-Latin languages from **Case 1**. mBERT, which performs the strongest on seen languages, exhibits the largest performance discrepancy between Latin and non-Latin based languages when evaluated on unseen languages. This highlights the limitation of the grapheme-based model, as it depends on the specific scripts.

On the other hand, the phoneme-based model—XPhoneBERT—demonstrates the most consistent performance over different unseen languages with little performance gap between Latin-based and non-Latin-based languages. This suggests that taking advantage of phonemes with its unified notation system allows for better generalization on extremely low-resource languages.

6 Conclusion

This paper presents the novel method of employing phonemes for identifying named entities for low-resource languages in zero-shot environments.

Our experiments compared the results of phoneme-based models with grapheme-based models in a strict zero-shot setting, and have shown that phonemes exhibit the best performance over low-resource languages unseen by all models. The results particularly demonstrate robustness towards non-Latin scripts, which is crucial in context of multilingual NER since languages are written in diverse writing systems.

7 Limitations

One limitation is that we examined only the languages included in WikiANN dataset and G2P modules we employed, resulting in a comparison of a small number of completely unseen languages. Additionally, we used a limited number of baselines with models of restricted scales, making it difficult to ensure that the results would remain consistent if the models were more extensively tailored to the task.

Perhaps more concerning, the performance achieved by these approaches is not sufficient for production use. While this is probably to be expected of zero-shot approaches, it demonstrates how much work is left before these approaches have practical utility.

8 Ethics Statement

In this work, we use WikiANN (Pan et al., 2017) which is publicly available dataset to train various models with different languages. The WikiANN authors already grappled with many of the ethical issues involved in the curation and annotation of this resource. We did not find any outstanding ethical concerns, including violent or offensive content, though there are likely strong biases in the named entities represented in the data. We used the dataset as consistent with the intended use. Nevertheless, we need to emphasize that, considering the characteristic of NER task, the dataset may contain personal information such as a specific person's real name or actual company name. We do not believe that this affects our result and the code and data distributed with our paper do not include any sensitive data of this kind.

References

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Arjun Das, Debasis Ganguly, and Utpal Garain. 2017. Named entity recognition with word embeddings and wikipedia categories for a low-resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(3).

Ujan Deb, Ridayesh Parab, and Preethi Jyothi. 2023. Zero-shot cross-lingual transfer with learned projections using unlabeled target-language data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 449–457, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. 2022. An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 170–179, Dublin, Ireland. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. 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 5834–5846, Online. Association for Computational Linguistics.

Ying Mo, Jian Yang, Jiahao Liu, Qifan Wang, Ruoyu Chen, Jingang Wang, and Zhoujun Li. 2024. mcl-ner: Cross-lingual named entity recognition via multi-view contrastive learning. *Preprint*, arXiv:2308.09073.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. Xphonebert: A pre-trained multilingual model for phoneme representations for text-to-speech. *arXiv preprint arXiv:2305.19709*.

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.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: Zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987, Singapore. Association for Computational Linguistics.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, C. Tran, Philipp Koehn, and Francisco Guzmán. 2021. Alternative input signals ease transfer in multilingual machine translation. *ArXiv*, abs/2110.07804.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. Crop: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. *Preprint*, arXiv:2210.07022.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech*.

A Appendix

A.1 Implementation Details

We ran training on English subset of WikiANN dataset for 10 epochs, with learning rate of 1e-5, weight decay 0.01, batch size 128, and warmup ratio 0.025 on 1 NVIDIA RTX A5000 GPU. We set the maximum sequence length of the input 128 for all the models. We experimented with models of BERT-base scale: mBERT with 177M parameters, CANINE-C with 132M, and XPhoneBERT with 87M.

A.2 Quantitative Results of Case 2 and Case 3

We present the quantitative result of all three cases in Table 3. The method using phoneme representation outperforms in Case 1 and Case 2 in terms of average F1 score and demonstrates more stable results with a lower standard deviation.

A.3 Comparison of Latin and Non-Latin Languages

In Figure 5, we visualize the results of the experiment separately for Latin and non-Latin languages in all cases. Compared to mBERT and CANINE that exhibit significant performance gaps between Latin and non-Latin languages, XPhoneBERT shows little difference in performance distribution.

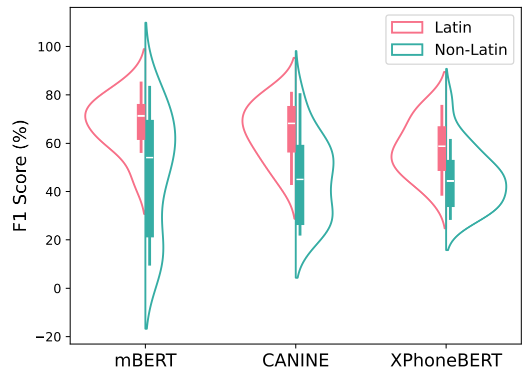

Figure 5: Latin and non-latin comparison

A.4 Language codes

In Table 4, we organized both ISO 639-1 and ISO 639-3 languages codes of all the languages used in the experiments.

A.5 Benchmark and License.

In Table 5, we provide the datasets, their statistics, and license. We also used CharsiuG2P (Zhu et al., 2022) toolkit for transliteration, which is under MIT license.

Case	Input	Model	Languages													AVG	STD
			sin	som	mri	quy	uig	aii	kin	ilo							
CASE 1	grapheme	mBERT	10.71	**44.76**	38.48	55.07	18.7	12.58	**62.37**	79.51						40.27	25
	grapheme	CANINE	26.31	43.35	**51.3**	**59.48**	27.19	22.38	54.74	**80.7**						45.68	19.99
	phoneme (ours)	XPhoneBERT	**43.61**	38.91	38.07	51.9	**44.82**	**31.03**	49.67	73.05						**46.38**	**12.67**
			epo	khm	tuk	amh	mlt	ori	san	ina	grn	bel	kur	snd			
CASE 2	grapheme	mBERT	71.31	16.12	**64.52**	11.9	**63.83**	9.96	48.73	**73.89**	50.44	**83.12**	54.16	35.02		48.58	25.13
	grapheme	CANINE	68.19	27.33	58.07	22.65	61.58	33.53	26.79	68.78	**55.37**	80.07	**57.33**	29.87		49.13	19.86
	phoneme (ours)	XPhoneBERT	**75.26**	**31.86**	61.17	**44.85**	52.58	**40.73**	**59.42**	68.68	49.95	77.61	52.95	**47.28**		**55.20**	**13.83**
			tgk	yor	mar	jav	urd	msa	ceb	hrv	mal	tel	uzb	pan	kir		
CASE 3	grapheme	mBERT	**74.1**	**56.6**	**74.3**	**73.59**	**57.09**	74.98	64.44	**84.93**	**69.94**	**67.24**	**80.04**	**53.98**	**68.14**	**69.18**	**9.28**
	grapheme	CANINE	62.12	51.15	44.28	61.11	42.41	**76.82**	**70.36**	77.51	48.29	37.29	72.54	45.74	57.73	57.49	13.77
	phoneme (ours)	XPhoneBERT	48.93	50.87	35.12	45.98	33.37	61.76	58.72	58.76	32.52	28.93	60.92	43.85	35.95	45.82	11.85

Table 3: Zero-shot F1 score (%) result in **Case 1**, **2**, and **3**.

Lang	Code	
	ISO 639-1	ISO 639-3
Amharic	am	amh
Assyrian Neo-Aramaic	aii	aii
Ayacucho quechua	qu	quy
Cebuano	ceb	ceb
Croatian	hr	hrv
English	en	eng
Esperanto	eo	epo
Ilocano	ilo	ilo
Javanese	jv	jav
Khmer	km	khm
Kinyarwanda	rw	kin
Korean	ko	kor
Kyrgyz	ky	kir
Malay	ms	msa
Malayalam	ml	mal
Maltese	mt	mlt
Maori	mi	mri
Marathi	mr	mar
Punjabi	pa	pan
Sinhala	si	sin
Somali	so	som
Spanish	es	spa
Tajik	tg	tgk
Telugu	te	tel
Turkmen	tk	tuk
Urdu	ur	urd
Uyghur	ug	uig
Uzbek	uz	uzb
Yoruba	yo	yor

Table 4: Language codes for all the languages used in the experiments.

Dataset	Lang.	Train	Dev	Test	License
WikiANN	eng	20k	10k	10k	ODC-BY
	sin	100	100	100	
	som	100	100	100	
	mri	100	100	100	
	quy	100	100	100	
	uig	100	100	100	
	aii	100	100	100	
	kin	100	100	100	
	ilo	100	100	100	
	epo	15k	10k	10k	
	khm	100	100	100	
	tuk	100	100	100	
	amh	100	100	100	
	mlt	100	100	100	
	ori	100	100	100	
	san	100	100	100	
	ina	100	100	100	
	grn	100	100	100	
	bel	15k	1k	1k	
	kur	100	100	100	
	snd	100	100	100	
	tgk	100	100	100	
	yor	100	100	100	
	mar	5k	1k	1k	
	jav	100	100	100	
	urd	20k	1k	1k	
	msa	20k	1k	1k	
	ceb	100	100	100	
	hrv	20k	10k	10k	
	mal	10k	1k	1k	
	tel	1k	1k	1k	
	uzb	1k	1k	1k	
	pan	100	100	100	
	kir	100	100	100	

Table 5: Statistics and license types for the dataset. The table lists the number of examples in the training, development, and testing sets for languages in the WikiANN dataset. The dataset is strictly used within the bounds of these licenses.