VBD_NLP at SemEval-2023 Task 2: Named Entity Recognition Systems Enhanced by BabelNet and Wikipedia

Phu Gia Hoang, Thanh Tuan Le, and **Long Hai Trieu**
{v.phuhg1, v.thanhlt41, v.longth12}@vinbigdata.com
VinBigData JSC, Hanoi, Vietnam

Abstract

We describe our systems participated in the SemEval-2023 shared task for Named Entity Recognition (NER) in English and Bangla. In order to address the challenges of the task, where a large number of fine-grained named entity types need to be detected with only a small amount of training data, we use a method to augment the training data based on BabelNet concepts and Wikipedia redirections to automatically annotate named entities from Wikipedia articles. We build our NER systems based on the powerful mDeBERTa pretrained language model and trained on the augmented data. Our approach significantly enhances the performance of the fine-grained NER task in both English and Bangla subtracks, outperforming the baseline models. Specifically, our augmented systems achieve macro-f1 scores of 52.64% and 64.31%, representing improvements of 2.38% and 11.33% over the English and Bangla baselines, respectively.

1 Introduction

The goal of the SemEval-2023 Task 2: MultiCoNER II Multilingual Complex Named Entity Recognition (Fetahu et al., 2023b) is to focus on extracting semantically ambiguous complex named entities based on the MULTICONER v2 (Fetahu et al., 2023a). The first version of MULTICONER (Malmasi et al., 2022b), which used the same taxonomy schema as introduced in WNUT 2017 (Derczynski et al., 2017), was utilized for SemEval-2022 Task 11: MultiCoNER (Malmasi et al., 2022b), that had 13 subtracks in multilingual, code-mixing, and 11 monolingual languages, namely: English, Spanish, Dutch, Russian, Turkish, Korean, Farsi, German, Chinese, Hindi, and Bangla. In the second version, the the MULTICONER v2 taxonomy was changed to challenge the previous systems and evaluate their performance on noisy and fine-grained data. The second version provides 13 subtracks in multilingual and 12 monolingual languages, including English, Spanish, Farsi, German, Chinese, Hindi, Bangla, Swedish, French, Italian, Portuguese, and Ukrainian.

This task is derived from the MultiCoNER shared task last year (Malmasi et al., 2022b), which was a challenging task due to scarcity training instances, with approximately ten thousand training instances and testing data that was ten times larger than the training data. Furthermore, the shared task this year has become more challenging when the number of entity types increases from 6 to 35 fine-grained entity types. This increase in the number of entity types has caused issues for models that train on few training instances and detect entities in a large number of classes. However, this task presents an interesting opportunity to investigate techniques for extending training data to deal with these issues.

As a step towards addressing the gap, we developed and evaluated our systems for the SemEval-2023 Task 2 in English and Bangla by augmenting the training data from a knowledge base. While English is the most commonly used language worldwide, Bangla represents a low-resource language that has not been well-investigated. We exploit a recent and powerful method for building NER data from the well-known BabelNet (Navigli and Ponzetto, 2012), using a combination of knowledge base (KB)-based and model-based techniques.Specifically, we annotated named entities from Wikipedia articles based on BabelNet concepts and Wikipedia redirection links, and further corrected the annotations by checking the agreement between the KB and neural model predictions. For our NER models, we built a strong baseline system based on the mDeBERTa pretrained model. We achieved a significant improvement in performance for both English (2%) and Bangla (12%) by utilizing the augmented data for training our systems, compared to the baseline. Our extensive anal-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1833–1843
July 13-14, 2023 ©2023 Association for Computational Linguistics

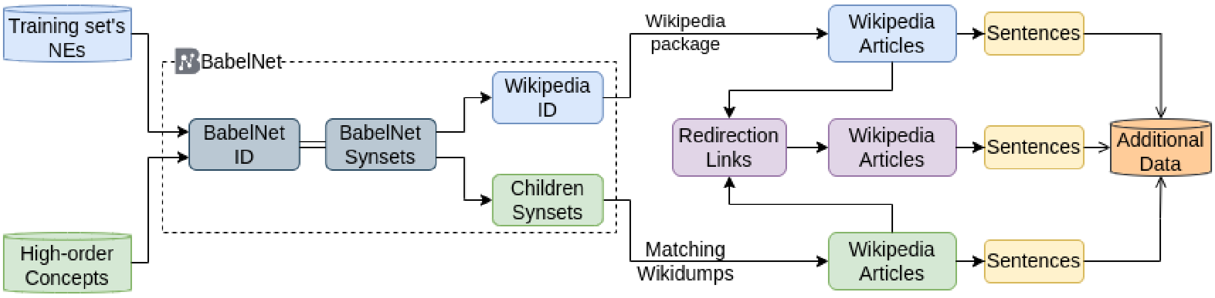

Figure 1: Visualization how we construct knowledge based enhanced data. In which, the blue, green, and purple data pipeline are described in Subsection 4.1, 4.2, and 4.3, respectively.

ysis also revealed the contribution of our method to each fine-grained entity type.

2 Dataset

We present the data statistics for English and Bangla in Table 1.

Compared to the Bangla subset, the English subset has almost twice as many sentences in the training phase. As a result, the English subset has more distinct entities and a larger vocabulary than the Bangla subset (19,466 versus 6,081 in terms of the number of distinct entities; 34,140 versus 22,274 in terms of vocabulary size, respectively). On average, each sentence consists of approximately 13 to 15 tokens, which is similar for both subsets. In addition, more statistics about the fine-grained taxonomy of the train, dev, and test sets of both subtasks are shown in Table 8 in Appendix A.

On the other hand, the test sets for both languages are extremely large in comparison to the train and dev sets. The reason for this, as stated in MULTICONER (Malmasi et al., 2022a), is that the organizers want to evaluate systems' ability to generalize to unseen and complex entities, as well as their performance on cross-domain adaptation tasks.

3 Systems Description

3.1 Baseline

We use pretrained language model mDeBERTaV3 (He et al., 2021) as the encoder. A sentence tokenized by the byte-pair-encoding algorithm (Sennrich et al., 2016) is feeding into the encoder to extract the representation of tokens. Then they passed through a pooling layer to get the representation of words. The BiLSTM-CRF architecture (Huang et al., 2015) is used to enhance the feature of words before projecting to n-labels-dimension embedding and predicting the label of the sequence.

The AdamW optimizer (Loshchilov and Hutter, 2017) is used to optimize the objective function. In addition, we split the parameters of model into two groups before feeding into optimizer. The first one contains the parameters of pretrained encoder with the use of a small learning rate. The second one contains the rest of parameters in the model which are tuned by using a larger learning rate.

3.2 Knowledge Based Enhanced Systems

To begin with, we use the provided dataset as seed samples to extend the data using the method based on MultiNERD, which we present in detail in Section 4. The enhanced data is then combined with the original dataset for training the model.

4 Knowledge Based Enhanced Data

4.1 Data Augmentation Using Provided Entities

In order to expand the training set of the English subset, we retrieved all the entities directly from the sentences to obtain their corresponding Wikipedia articles. As shown in Figure 7, we used BabelNet to retrieve the corresponding articles. For ambiguous entities with multiple possible articles or no associated article, we either discarded them or kept only the first article, respectively. After obtaining the article names, we used the Wikipedia[1] package to extract the article texts, which were then segmented into sentences, and only the sentences that contained the corresponding entities were kept.

4.2 Data Augmentation Using BabelNet

Following the approach proposed in (Tedeschi and Navigli, 2022), we manually curated 1,254 synsets to encompass a wide range of high-level concepts

[1] Version: 1.4.0, https://pypi.org/project/wikipedia/

	English			Bangla		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Number of instances	16,778	871	249,980	9,708	507	19,859
Number of distinct named entities	19,466	1159	226,609	6,081	544	10,614
Average instance length	15	15	15	13	13	13
Vocabulary size	34,140	4,733	235,431	22,274	3,282	33,979

Table 1: Statistics of training, dev, test sets of English subset and Bangla subset.

based on the task's taxonomy. Examples of some of these synsets are provided below:

- SCIENTIST: bibliotist, linguist, physicist, etc
- DISEASE: horse disease, seafood poisoning, etc
- PRIVATE CORP: public limited company, public corporation, etc

To expand the initial set of concepts, we utilized BabelNet's hyponymy (has-instance) relationships to obtain child synsets. For example, for the 'linguist' concept (`bn:00051385n`), we retrieved its children synsets, such as Noam Chomsky[2] (`bn:00000162n`) and Leonard Bloomfield[3] (`bn:00011432n`), which inherited the SCIENTIST label.

We followed a similar approach for Bangla, but we did not annotate high-order concepts in Bangla. Instead, we used the English-Bangla mechanisms of BabelNet to obtain children synsets. In total, this process yielded 120k children synsets for English and 113k for Bangla.

4.3 Wikipedia Redirections

Apart from using BabelNet synsets, we also leveraged another resource, namely Wikipedia redirection links, to expand our concept list. For example, the term *Apple (corporation)* can be redirected to *Apple Inc.*, even though the terms have different surface text, they refer to the same Wikipedia page.

4.4 Annotation Enhancement

Figure 2 illustrates our methodology for improving the quality of our augmented data. This methodology was inspired by the method proposed in (Tedeschi and Navigli, 2022). The augmented data, which is a combination of additional data and the provided data, was divided into n non-overlapping subsets. The first subset (D_0) was used to train a self-annotating model, which is a Transformer-based neural classifier (mBERT + Bi-LSTM + CRF, (Mueller et al., 2020)), to obtain an initial model (M_0). The initial model is used as the first version for the self annotating loop, which proceeds as follows:

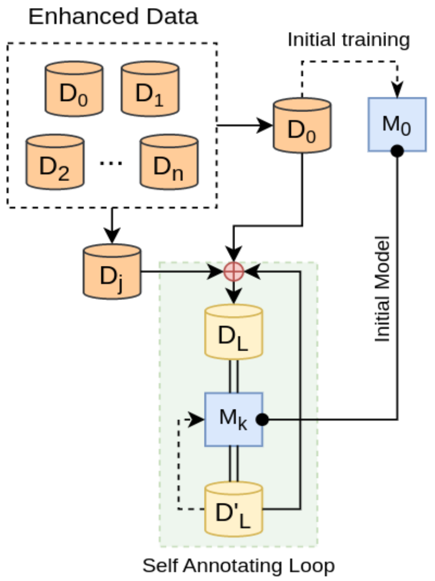

Figure 2: Visualization of our self-annotating model, in which straight arrows represent data paths, dot arrows represent the training process, double lines represent the annotation process performed by the corresponding model, and the connection between M_k and M_0 represents the model inherited the initial model in the first loop.

1. The first subset (D_0) is concatenated with another subset ($D_j, j \in \{1, 2, ..., n\}$) to form a larger set ($D_L$).

2. The self-annotating model (M_k, $k \in \{1, 2, ..., t\}$, in which $t = 0$ denotes the initial model, while t represents the number of loops) annotates the larger set (D_L). During this annotation process, if the NER class assigned by the self-annotating model for a predicted

[2] https://en.wikipedia.org/wiki/Noam_Chomsky

[3] https://en.wikipedia.org/wiki/Leonard_Bloomfield

	English	Bangla
#Instances	419,483	22,216
#Distinct NEs	23,314	7,168
Avg. instance length	22	29
Vocabulary size	226,346	74,425

Table 2: Statistics of augmented training set for both languages. In which, #instances: number of instances, #distinct NEs: number of distinct entities, avg. instance length: average instance length.

entity is different from the one assigned by the data augmentation process (Section 4.4, 4.2, and 4.1), the corresponding sentence is removed to obtain a noise-reduced set (D'_L).

3. D'_L is then used to combine with the next subset in the queue in the first step of the next loop instead of the first subset (D_0).

4.5 Augmented Data statistics

The augmentation method we applied significantly improved the training sets for both English and Bangla, as demonstrated in Table 2. With the aid of additional data, the total number of training instances for English increased from 16,788 to 419,483, and from 9,708 to 22,216 for Bangla. Furthermore, our method increased the vocabulary size by nearly 6.5 times in English and 3 times in Bangla, indicating that it not only enriched the context of existing entities in the training sets, but also provided new contexts for previously unseen entities. These findings highlight the effectiveness of our augmentation method in improving the quality and size of the training sets. Such improvements can have significant implications for the development of better entity recognition models (see Section 5 for more details).

4.6 Experimental Settings

Our experiments were conducted on hardware with a single RTX 2080Ti GPU. The training process was set up with the following hyper-parameters: a batch size of 8, a dropout rate of 0.3, learning rates of 0.001 and 0.00005 for two parameter groups (as mentioned above), and a default of 10 training epochs. The best checkpoint was chosen based on the f1-macro scores on the dev set during training.

4.7 Results

We present our results in Tables 3 and 4. In overall, our augmented systems significantly improve the baseline systems in all the metrics, which are precision, recall, and F1-macro.

Model	Precision	Recall	F1
Baseline	51.64	51.29	50.26
Augmented	**57.89**	**51.36**	**52.64**

Table 3: Results on English test set.

Model	Precision	Recall	F1
Baseline	52.83	59.97	52.98
Augmented	**63.01**	**68.18**	**64.31**

Table 4: Results on Bangla test set.

5 Results and Analyses

In particular, the results on English presented in Table 3 show that the precision, recall, and F1-score of the baseline model were 51.64%, 51.29%, and 50.26%, respectively, while these numbers were 57.89%, 51.36%, and 52.64% for the augmented model. The F1-score saw a moderate improvement as a result of the augmentation technique. The precision ratings of the two models, however, did not differ significantly from one another. As a result of having a slightly higher recall score than the baseline model, the augmented model was able to recognize more pertinent instances. The model's performance may be affected by several factors, including the task's complexity and the quantity and quality of the training data. In this instance, it seems that the augmentation technique was advantageous for enhancing the model's capacity to to identify relevant instances, resulting in an improvement in the overall F1-score.

Meanwhile, the augmented model improve all scores by almost 10±2% in all scores for the Bangla results (Table 4). The augmentation technique had a significantly positive impact on all three metrics, greatly enhancing the performance of the model. The precision of the augmented model increased by 10.18%, demonstrating that it was able to recognize more relevant instances while fewer false positive predictions were made. Additionally, the recall score of the augmented model increased by 8.21%, showing that it was able to extract more relevant instances from the dataset. The augmented model's improved overall performance is evidenced by the F-1 score's increase.

WNUT2017's	**Datasets**	**Baseline**			**Augmented**		
Taxonomy	**Fine-grained Taxonomy**	**P**	**R**	**F1**	**P**	**R**	**F1**
Person	Cleric	77.78	63.64	**70.00**	58.33	63.64	60.87
	Scientist	21.05	36.36	26.67	50.00	45.45	**47.62**
	SportsManager	75.00	21.43	33.33	100.00	71.43	**83.33**
	Politician	46.15	17.14	25.00	80.00	45.71	**58.18**
	OtherPER	28.30	46.88	35.29	46.00	71.88	**56.10**
	Athlete	53.57	60.00	56.60	65.52	76.00	**70.37**
	Artist	66.67	60.61	63.49	56.32	74.24	**64.05**
Product	Clothing	83.33	100.00	90.91	100.00	90.00	**94.74**
	Drink	92.31	100.00	**96.00**	84.62	91.67	88.00
	Food	52.38	73.33	61.11	82.35	93.33	**87.50**
	Vehicle	90.91	90.91	**90.91**	66.67	90.91	76.92
	OtherPROD	55.56	52.63	54.05	78.95	78.95	**78.95**
Medical	Symptom	83.33	100.00	**90.91**	81.82	90.00	85.71
	Medication/Vaccine	48.00	85.71	61.54	80.00	85.71	**82.76**
	Medication/Vaccine	48.00	85.71	61.54	80.00	85.71	**82.76**
	Disease	91.67	73.33	81.48	93.33	93.33	**93.33**
	AnatomicalStructure	66.67	85.71	**75.00**	66.67	57.14	61.54
Location	OtherLOC	55.56	41.67	47.62	100.00	91.67	**95.65**
	Station	80.00	92.31	85.71	80.00	92.31	85.71
	Facility	66.67	72.73	**69.57**	70.00	63.64	66.67
	HumanSettlement	75.27	87.50	80.92	78.72	92.50	**85.06**
CreativeWork	ArtWork	91.67	100.00	**95.65**	100.00	90.91	95.24
	MusicalWork	54.55	50.00	52.17	61.54	66.67	**64.00**
	Software	78.57	84.62	81.48	96.00	92.31	**94.12**
	WrittenWork	77.27	65.38	70.83	78.26	69.23	**73.47**
	VisualWork	57.14	71.43	63.49	90.00	64.29	**75.00**
Groups	PrivateCorp	100.00	58.33	73.68	100.00	83.33	**90.91**
	AerospaceManufacturer	77.78	93.33	84.85	100.00	93.33	**96.55**
	CarManufacturer	68.75	91.67	78.57	91.67	91.67	**91.67**
	PublicCorp	75.00	63.16	68.57	100.00	84.21	**91.43**
	SportsGRP	100.00	94.12	96.97	100.00	94.12	96.97
	MusicalGRP	75.00	60.00	66.67	86.67	86.67	**86.67**
	ORG	81.48	68.75	74.58	88.00	68.75	**77.19**
Average micro		66.71	69.08	67.88	76.70	78.40	**77.54**
Average macro		70.58	71.29	69.37	82.16	79.55	**80.16**
Average weighted		68.60	69.08	67.44	79.08	78.40	**77.82**

Table 5: Performance comparison of baseline and augmented systems on fine-grained Bangla dev set.

5.1 Analyses

Table 5, 6, 9, and 10 compare the baseline system's results to augmented system's results on the fine-grained taxonomy of the dev and test set of English and Bangla languages.

In the Bangla language, the augmented system using additional data (see Section 4) exhibits significantly better performance than the baseline model in most classes. For example, in the dev set (Table 5), the F1-scores of classes such as Scientist are 33.33% and 83.33%, OtherLOC are 47.62% and 95.65%, Food are 61.11% and 87.50%, Software 81.48% and 94.12%, among others. Similarly, in the test set (Table 9), examples include SportManager with scores of 15.93% and 50.64%, Symptom with scores of 62.41% and 81.06%, and PrivateCorp with scores of 50.81% and 78.95%. However, the augmented model has worse performance than the baseline model in some classes, namely Cleric, Drink, Vehicle, AnatomicalStructure, Facility, and

WNUT2017's	**Datasets**	**Baseline**			**Augmented**		
Taxonomy	**Fine-grained Taxonomy**	**P**	**R**	**F1**	**P**	**R**	**F1**
Person	Cleric	41.67	33.33	37.04	83.33	33.33	**47.62**
	Scientist	40.00	26.67	**32.00**	20.00	6.67	10.00
	SportsManager	68.75	68.75	68.75	73.33	68.75	**70.97**
	Politician	54.29	35.85	**43.18**	50.00	30.19	37.65
	OtherPER	44.34	51.65	**47.72**	38.03	59.34	46.35
	Athlete	76.32	73.42	**74.84**	76.39	69.62	72.85
	Artist	73.78	78.30	**75.97**	75.85	74.06	74.94
Product	Clothing	41.18	70.00	**51.85**	50.00	40.00	44.44
	Drink	40.00	36.36	38.10	80.00	72.73	**76.19**
	Food	57.14	42.11	48.48	70.59	63.16	**66.67**
	Vehicle	64.71	55.00	**59.46**	44.44	40.00	42.11
	OtherPROD	43.55	55.10	48.65	48.15	53.06	**50.49**
Medical	Symptom	40.00	80.00	53.33	71.43	100.00	**83.33**
	MedicalProcedure	38.89	53.85	45.16	57.14	61.54	**59.26**
	Medication/Vaccine	78.95	83.33	**81.08**	82.35	77.78	80.00
	Disease	50.00	44.44	47.06	66.67	44.44	**53.33**
	AnatomicalStructure	76.92	58.82	66.67	59.09	76.47	66.67
Location	OtherLOC	85.71	37.50	52.17	72.73	50.00	**59.26**
	Station	57.81	71.15	**63.79**	61.11	63.46	62.26
	Facility	86.67	65.00	**74.29**	66.67	70.00	68.29
	HumanSettlement	83.81	80.73	**82.24**	81.73	77.98	79.81
CreativeWork	ArtWork	75.00	23.08	35.29	75.00	46.15	**57.14**
	MusicalWork	68.09	52.46	59.26	70.21	54.10	**61.11**
	Software	53.85	53.85	53.85	70.00	53.85	**60.87**
	WrittenWork	46.58	62.96	53.54	71.43	64.81	**67.96**
	VisualWork	57.89	54.10	**55.93**	51.47	57.38	54.26
Groups	PrivateCorp	20.00	18.18	19.05	33.33	36.36	**34.78**
	AerospaceManufacturer	20.00	18.18	19.05	33.33	36.36	**34.78**
	CarManufacturer	60.00	69.23	**64.29**	56.25	69.23	62.07
	PublicCorp	31.43	39.29	34.92	83.33	35.71	**50.00**
	SportsGRP	81.82	87.80	**84.71**	85.71	73.17	78.95
	MusicalGRP	79.31	62.16	**69.70**	68.75	59.46	63.77
	ORG	50.57	56.41	53.33	64.62	53.85	**58.74**
Average micro		61.29	62.19	61.74	64.80	61.65	**63.19**
Average macro		58.29	55.78	55.54	64.35	58.38	**59.91**
Average weighted		62.72	62.19	61.70	66.27	61.65	**63.01**

Table 6: Performance comparison of baseline and augmented systems on fine-grained English dev set

ArtWork.

In contrast to the Bangla language, the Baseline model in English outperforms the Augmented model in a greater number of classes (Table 6 and 10). Specifically, out of a total of 33 classes, the Baseline model performs better than the Augmented model in 14 classes in the dev set and 15 classes in the test set.

There are a number of possible explanations for why the Baseline model outperformed the augmented model in English while the augmented model performed better in Bangla. The quantity and caliber of additional data used for augmentation could be one factor in the explanation. The augmentation method in Bangla was successful in obtaining a larger and more varied set of data, which might have added more context and enhanced the model's capacity for entity recognition and classification. In contrast, it is possible that the complexity and nuance of the English language prevented

Languages	Models	Instances
English	Gold annotation	**eli lilly [OtherPER]** founder president of pharmaceutical company **eli lilly and company [PublicCorp]**.
	Baseline model	**eli lilly [Scientist]** founder president of pharmaceutical company **eli lilly and company [PublicCorp]**.
	Augmented model	**eli lilly [OtherPER]** founder president of pharmaceutical company **eli lilly and company [Station]**.
	Gold annotation	his first film was **on the waterfront [VisualWork]** in 1954.
	Baseline model	his first film was **on the waterfront [WrittenWork]** in 1954.
	Augmented model	his first film was **on the waterfront [VisualWork]** in 1954.
	Gold annotation	these paintings among them the **feast of the gods [ArtWork]** and the **bacchus and ariadne [ArtWork]** were executed by **giovanni bellini [Artist]** and **titian [Artist]**.
	Baseline model	these paintings among them the **feast of the gods [ArtWork]** and the **bacchus [OtherPER]** **and ariadne** were executed by **giovanni bellini [Artist]** and **titian [Artist]**.
	Augmented model	these paintings among them the **feast of the gods [ArtWork]** and the **bacchus and ariadne [ArtWork]** were executed by **giovanni bellini [Artist]** and **titian [Artist]**.
Bangla	Gold annotation	**বাংলাদেশ জাতীয়তাবাদী দল [ORG]** **জামায়াত-এ-ইসলামী বাংলাদেশ [ORG]** জোট ক্ষমতায় আসার পরে তাকে ২০০২ সালে তার অফিস থেকে সরিয়ে দেওয়া হয়েছিল। *[Eng: He was removed from his office in 2002 after the **Bangladesh nationalist party [ORG]** **Jamaat-e-Islami Bangladesh [ORG]** coalition came to power.]*
	Baseline model	**বাংলাদেশ জাতীয়তাবাদী দল** **জামায়াত-এ-ইসলামী বাংলাদেশ [ORG]** জোট ক্ষমতায় আসার পরে তাকে ২০০২ সালে তার অফিস থেকে সরিয়ে দেওয়া হয়েছিল। *[Eng: He was removed from his office in 2002 after the **Bangladesh nationalist party** **Jamaat-e-Islami Bangladesh [ORG]** coalition came to power.]*
	Augmented model	**বাংলাদেশ জাতীয়তাবাদী দল [ORG]** **জামায়াত-এ-ইসলামী বাংলাদেশ [ORG]** জোট ক্ষমতায় আসার পরে তাকে ২০০২ সালে তার অফিস থেকে সরিয়ে দেওয়া হয়েছিল। *[Eng: He was removed from his office in 2002 after the **Bangladesh nationalist party [ORG]** **Jamaat-e-Islami Bangladesh [ORG]** coalition came to power.]*
	Gold annotation	আনাকু (একিউএসইউ আকসু) একটি **স্কার্ট [Clothing]** **ইনকা সাম্রাজ্য [HummanSett.]** এ আদিবাসী মহিলাদের টাইপ পোশাক ছিল। *[Eng: Anaku (AQSU Aksu) was a **skirt [Clothing]** worn by indigenous women in the **Inca Empire [HummanSett.]**.]*
	Baseline model	আনাকু (একিউএসইউ আকসু) একটি **স্কার্ট** **ইনকা সাম্রাজ্য [HummanSett.]** এ আদিবাসী মহিলাদের টাইপ পোশাক ছিল। *[Eng: Anaku (AQSU Aksu) was a **skirt** worn by indigenous women in the **Inca Empire [HummanSett.]**.]*
	Augmented model	**আনাকু [HummanSett.]** (একিউএসইউ আকসু) একটি **স্কার্ট [Clothing]** **ইনকা সাম্রাজ্য [HummanSett.]** এ আদিবাসী মহিলাদের টাইপ পোশাক ছিল। *[Eng: **Anaku [HummanSett.]** (AQSU Aksu) was a **skirt [Clothing]** worn by indigenous women in the **Inca Empire [HummanSett.]**.]*

Table 7: Case studies in the dev set that are complicated for the Baseline model and Augmented model. In which, bold green entities represent gold annotations that were correctly predicted by the models. Conversely, bold red entities correspond to incorrect predictions, while bold black entities indicate entities that were not predicted by the corresponding model. (the [HummanSett.] is a short form of [HummanSettlement])

the augmentation of English language data from having as much of an impact on the model's performance. The structure and characteristics of the languages themselves might also have been important. For example, Bangla language has a more rigid grammatical structure and less variation in word order, which could make it easier for the model to identify and classify entities. In contrast, the English language has a more complex grammatical structure and greater variation in word order, which could make it more difficult for the model to accurately identify and classify entities.

Prediction samples Table 7 shows the varying performance of the baseline and augmented models. In the English subtask, the baseline model performs worse in all cases, but still able to predict some entities. In contrast,the augmented model shows a consistent performance in predicting the correct entity boundaries, albeit with some missed entities, indicating that the augmentation did not significantly improve the performance in this task. In particular, all samples shows a clear weakness of the baseline model in identifying the correct boundaries and classes of the entities. Whereas the augmented model shows an improvement in recognizing all the entities except the second entity in the first sample.In the Bangla subtask, the baseline model's performance is significantly worse than in the English subtask, failing to recognize the first entity in both samples. The augmented model shows better performance, correctly identifying both entities in Sample 1. However, in Sample 2, the augmented model shows a concerning behavior, predicting an additional wrong entity.

Overall, the results suggest that the augmentation has a limited impact on improving the baseline model's performance in the both subtasks, as the two models struggle to identify the correct boundaries of the entities.

Limitations and Future work. For this work, we only evaluated on English and Bangla, a low resource language. We intend to extend our experiments on other languages. Additionally, different NER models can be applied to extract more complex entities such as nested entities using exhaustive methods (Sohrab et al., 2019a,b; Wang et al., 2022). Filtering noisy annotations is also an important task in future work.

6 Conclusion

In this paper, we have introduced our NER systems for the SemEval-2023 Task 2: MultiCoNER II Multilingual Complex Named Entity Recognition. Our NER systems are finetuned on the mDeBERTa pretrained language model and enhanced by augmented data from BabelNet and Wikipedia. Specifically, BabelNet concepts are utilized to annotate named entities from Wikipedia articles. Experimental results show that the augmented data significantly improve baseline systems trained on the limited training data, especially on the low-resource language such as Bangla. For future work, we plan to improve the quality of augmented data by filter noisy annotations as well as extend our systems on other languages.

References

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

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

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.

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)*. Association for Computational Linguistics.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Mohammad Golam Sohrab, Thang M. Pham, and Makoto Miwa. 2019a. A generic neural exhaustive approach for entity recognition and sensitive span detection. In *IberLEF@SEPLN*.

Mohammad Golam Sohrab, Thang M. Pham, Makoto Miwa, and Hiroya Takamura. 2019b. A neural pipeline approach for the pharmaconer shared task using contextual exhaustive models. In *Conference on Empirical Methods in Natural Language Processing*.

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

Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested named entity recognition: A survey. 16(6).

A Taxonomy statistics

WNUT2017's Taxonomy	Dataset's Fine-grained Taxonomy	English			Bangla		
		Train	Dev	Test	Train	Dev	Test
Person	Cleric	299	15	4,732	239	11	240
	Scientist	318	15	4,928	254	11	255
	SportsManager	344	16	5,333	239	14	198
	Politician	1,050	53	15,990	635	35	1,294
	OtherPER	1,777	79	22,028	677	32	1,117
	Athlete	1,793	91	27,636	562	25	1,087
	Artist	3,713	212	57,034	1,172	66	2,744
Product	Clothing	198	10	2,244	199	10	17
	Drink	212	11	2,246	218	12	120
	Food	362	19	5,317	345	15	453
	Vehicle	377	20	5,935	217	11	199
	OtherPROD	786	49	11,838	405	19	704
Medical	Symptom	202	10	1,759	206	19	105
	MedicalProcedure	242	13	3,850	229	19	266
	Medication/Vaccine	355	17	5,421	310	14	462
	Disease	372	18	5,623	351	15	554
	AnatomicalStructure	388	18	5,838	300	14	532
Location	OtherLOC	291	16	4,635	213	12	172
	Station	392	20	5,978	242	13	298
	Facility	1,053	52	16,185	423	22	894
	HumanSettlement	2,617	109	41,103	1,579	80	6,011
CreativeWork	ArtWork	199	13	1,270	194	11	455
	MusicalWork	953	61	15,304	261	12	226
	Software	593	26	8,962	462	26	812
	WrittenWork	1,073	54	16,912	566	26	1,224
	VisualWork	1,266	61	19,678	498	28	923
Groups	PrivateCorp	201	11	810	58	12	127
	AerospaceManufacturer	216	10	1,015	230	15	97
	CarManufacturer	249	13	2,984	220	12	84
	PublicCorp	437	28	6,825	372	19	460
	SportsGRP	816	41	13,009	377	17	595
	MusicalGRP	825	37	12,969	304	15	300
	ORG	1,480	78	22,414	666	32	1,988

Table 8: Taxonomy statistics of provided training, dev, test sets of English subset and Bangla subset.

B Test Results Analysis

WNUT2017's Taxonomy	**Datasets Fine-grained Taxonomy**	**Baseline**			**Augmented**		
		P	**R**	**F1**	**P**	**R**	**F1**
Person	Cleric	57.43	59.58	**58.49**	45.80	65.83	54.02
	Scientist	16.34	45.10	23.98	29.51	40.39	**34.11**
	SportsManager	64.29	9.09	15.93	44.03	59.60	**50.64**
	Politician	66.20	21.95	32.97	61.21	46.21	**52.66**
	OtherPER	24.86	42.44	31.35	33.96	48.88	**40.07**
	Athlete	64.70	56.49	**60.31**	66.51	53.36	59.21
	Artist	62.58	54.74	58.40	56.62	64.40	**60.26**
Product	Clothing	17.57	76.47	28.57	28.00	82.35	**41.79**
	Drink	60.11	89.17	71.81	65.00	86.67	**74.29**
	Food	39.68	59.82	47.71	53.06	66.89	**59.18**
	Vehicle	57.81	68.84	62.84	74.77	81.91	**78.18**
	OtherPROD	48.01	49.57	48.78	54.81	61.51	**57.97**
Medical	Symptom	49.72	83.81	62.41	75.41	87.62	**81.06**
	MedicalProcedure	60.06	76.32	67.22	79.51	84.59	**81.97**
	Medication/Vaccine	42.44	84.42	56.48	78.59	81.82	**80.17**
	Disease	61.84	77.80	68.90	80.17	85.38	**82.69**
	AnatomicalStructure	61.48	62.41	61.94	83.47	75.00	**79.01**
Location	OtherLOC	43.98	68.02	53.42	63.06	81.40	**71.07**
	Station	75.65	87.58	81.18	82.01	90.27	**85.94**
	Facility	73.33	52.91	61.47	68.79	65.32	**67.01**
	HumanSettlement	82.27	85.28	83.74	84.09	88.57	**86.27**
CreativeWork	ArtWork	10.71	2.64	4.23	18.03	2.42	**4.26**
	MusicalWork	36.14	53.10	43.01	47.40	64.60	**54.68**
	Software	67.32	75.62	71.23	85.00	73.28	**78.70**
	WrittenWork	66.94	59.07	62.76	74.98	62.91	**68.41**
	VisualWork	46.58	59.80	52.37	54.43	59.91	**57.04**
Groups	PrivateCorp	81.03	37.01	50.81	75.54	82.68	**78.95**
	AerospaceManufacturer	8.74	9.28	**9.00**	15.38	2.06	3.64
	CarManufacturer	34.23	90.48	49.67	68.10	94.05	**79.00**
	PublicCorp	46.42	61.96	53.07	75.44	74.78	**75.11**
	SportsGRP	83.60	88.24	85.85	91.91	91.60	**91.75**
	MusicalGRP	42.09	57.67	48.66	72.26	70.33	**71.28**
	ORG	89.25	72.28	79.88	92.47	73.49	**81.89**
Average micro		61.64	64.05	62.82	69.50	70.34	**69.91**
Average macro		52.83	59.97	52.98	63.01	68.18	**64.31**
Average weighted		64.87	64.05	62.88	70.20	70.34	**69.71**

Table 9: Performance comparison of baseline and augmented systems on fine-grained Bangla test set.

WNUT2017's Taxonomy	**Datasets Fine-grained Taxonomy**	**Baseline**			**Augmented**		
		P	**R**	**F1**	**P**	**R**	**F1**
Person	Cleric	49.38	34.59	40.69	73.70	36.67	**48.97**
	Scientist	46.62	26.58	**33.86**	63.99	12.30	20.63
	SportsManager	51.77	49.05	50.38	49.79	54.49	**52.03**
	Politician	65.81	37.27	**47.59**	68.02	30.63	42.24
	OtherPER	36.30	49.71	**41.96**	29.72	61.06	39.98
	Athlete	77.51	65.65	**71.09**	82.29	58.67	68.50
	Artist	69.19	78.07	73.36	71.52	76.08	**73.73**
Product	Clothing	39.09	57.22	46.45	48.64	46.12	**47.35**
	Drink	44.25	48.13	46.11	59.54	47.77	**53.01**
	Food	48.66	40.70	44.33	49.15	40.57	**44.45**
	Vehicle	46.78	38.42	**42.19**	36.00	43.57	39.42
	OtherPROD	34.56	45.51	39.28	41.32	44.40	**42.81**
Medical	Symptom	14.69	48.15	22.51	44.13	54.75	**48.87**
	MedicalProcedure	52.42	53.45	52.93	60.29	49.32	**54.26**
	Medication/Vaccine	60.57	69.16	**64.58**	70.50	58.99	64.24
	Disease	44.43	43.45	43.93	72.19	49.67	**58.85**
	AnatomicalStructure	65.15	55.98	**60.22**	52.98	59.22	55.92
Location	OtherLOC	58.19	30.12	39.69	53.63	47.34	**50.29**
	Station	78.64	66.14	**71.85**	69.55	66.61	68.05
	Facility	49.97	63.53	**55.94**	53.27	58.16	55.61
	HumanSettlement	87.06	77.77	**82.15**	81.74	78.18	79.92
CreativeWork	ArtWork	42.46	37.01	**39.55**	32.81	39.37	35.79
	MusicalWork	67.66	60.51	**63.88**	59.92	58.87	59.39
	Software	55.27	63.26	**58.99**	68.21	49.45	57.34
	WrittenWork	44.99	61.75	52.06	61.32	52.02	**56.28**
	VisualWork	63.79	55.13	**59.14**	48.28	59.04	53.12
Groups	PrivateCorp	8.96	19.38	12.25	18.97	43.46	**26.41**
	AerospaceManufacturer	21.71	47.59	29.81	37.70	67.49	**48.38**
	CarManufacturer	58.38	50.90	54.39	65.91	54.36	**59.58**
	PublicCorp	37.47	48.45	42.26	69.52	40.34	**51.05**
	SportsGRP	74.56	72.87	**73.71**	85.68	64.24	73.43
	MusicalGRP	60.01	46.51	52.40	70.16	43.30	**53.55**
	ORG	47.78	50.77	49.23	59.79	48.64	**53.64**
Average micro		58.46	59.91	59.18	60.30	58.23	**59.25**
Average macro		51.64	51.30	50.27	57.89	51.37	**52.64**
Average weighted		60.89	59.91	**59.67**	63.98	58.23	59.57

Table 10: Performance comparison of baseline and augmented systems on fine-grained English test set.