Practical Approaches for Low-Resource Named Entity Recognition of Filipino Telecommunications Domain

Kyle Chan
PLDT Inc
kachan@pldt.com.ph

Kaye Ann Delas Alas
Smart Communications
kcdelasalas@smart.com.ph

Charles Orcena
PLDT Inc
clorcena@pldt.com.ph

Dan John Velasco
Senti AI
dan.velasco@senti.com.ph

Qyle John San Juan
Senti AI
qyle.sanjuan@senti.com.ph

Charibeth Cheng
De La Salle University
charibeth.cheng@dlsu.edu.ph

Abstract

Telecommunication companies understand the value of engaging with customers and analyzing their interactions on a large scale. Extracting entities such as names, organizations, and locations from unstructured data is crucial for improving customer experience, making informed decisions, and supports compliance with local and international data privacy requirements. This study highlights the use of cross-lingual task learning in developing a named entity recognizer for a low-resource language, with a specific focus on entities relevant to telecommunication companies. Additionally, the study demonstrates how data augmentation techniques can enhance the performance of the model. Among the baseline models, XLM-RoBERTa-Large$_{CoNLL}$ performed best with an F1 score of 0.673 for full-span entity recognition and 0.843 for partial recognition. Additionally, data augmentation improved performance by up to 8.6% for full-span recognition and up to 6.94% for partial recognition.

1 Introduction

Telecommunications companies (Telecom) aim to constantly engage and listen to their customers through their customer service platforms. The ability to contextualize and analyze all engagement at scale is an important tool in providing quality and timely service. With the onset of social media as a customer service tool, it has rapidly become an important source of information, however, it is mostly untapped or under utilized due to its unstructured form.

Named Entity Recognition (NER) is a well established sub-task of Information Extraction (IE) dealing with the identification of entities within a text. Extracting relevant entities such as names, organizations and locations from unstructured data is crucial for enhancing customer experience and making informed business decisions. By accurately recognizing customer names and enterprise organizations, telecoms can deliver personalized offers and recommendations to their customers. Meanwhile proper identification of location tags will greatly aid in the ability to identify important network events and outages enabling proactive maintenance and repairs leading to improved network reliability and performance. Accurate identification and masking of these entities is also an important endeavor in support of local and international data privacy requirements.

A unique aspect of entity recognition task within telecoms context is the classification of numerical entities within the corpora. Distinguishing whether a number represents a phone number or an account number plays a pivotal role in customer identification facilitating the connection between unstructured online information and organized telecommunications datasets.

In this paper, we show how the language model, cross-lingual task learning and data augmentation affect the Filipino NER task in low-resource settings. Telecom domain relevant entities were identified and tagged from Filipino-English based social media extracts; minimal subset was used due to limited annotation capacity. Initial experiments using different pretrained language models – multilingual, English, and Filipino language models were conducted for baseline comparison. F1, precision, and recall were used to evaluate all experiments following CoNLL-2003 standards (Tjong Kim Sang and De Meulder, 2003).

2 Related Works

2.1 Named Entity Recognition for low-resource languages

Prior to the emergence of deep learning, the NER task for low-resource languages was addressed using techniques such as manually crafted rules, Conditional Random Fields (Alfonso et al., 2013) and Hidden Markov Models (Ekbal and Bandyopad-

Table 1: Pretrained models used in experiments along with the corresponding language it was trained on, parameter size, and size on disk.

Model	Language	Parameter size	Size on disk
Telecom-RoBERTa-Base	Tagalog-English	125M	501MB
Tagalog-RoBERTa	Tagalog	108M	437MB
BERT-Base-Cased	English	109M	436MB
BERT-Base-Cased$_{CoNLL}$	English	109M	436MB
XLM-RoBERTa-Base	Multilingual	279M	1.12GB
XLM-RoBERTa-Base$_{CoNLL}$	Multilingual	279M	1.11GB
XLM-RoBERTa-Large	Multilingual	561M	2.24GB
XLM-RoBERTa-Large$_{CoNLL}$	Multilingual	561M	2.24GB

Table 2: Summary statistics of entities in the dataset. Sequence length refers to range of token count of each entity.

Entity	Count	Sequence Length	Mean Length
PNUM	324	1-5	1
NAME	34	1-4	2
ANUM	37	1	1
ORG	69	1-6	2
LOC	48	1-20	3

hyay, 2007). In Filipino NER, various machine learning (ML) techniques have been used in recent years, including Maximum Entropy (Eboña et al., 2013) , Support Vector Machines (Castillo et al., 2013) , combining ML techniques and hand-crafted rules (Livelo et al., 2017), and hybrid deep learning techniques (Gonzales et al., 2023). These works highlighted the need for a larger training data to improve the performance of their NER systems.

2.2 Transfer Learning and Cross-lingual transfer

Transfer learning in NLP involves using pretrained language models to enhance the performance of a new or similar NLP task (Howard and Ruder, 2018; Devlin et al., 2019). Instead of starting from scratch with a large, labeled dataset, a language model is initially trained on a vast collection of text data, enabling it to learn statistical properties, syntactic structures, and semantic relationships within the text. The acquired knowledge is then transferred to a different task through fine-tuning, where the pretrained model is further trained on downstream tasks.

When dealing with low-resource languages, there are additional challenges for NER. Limited resources, such as annotated data and language models, make it difficult to train NER models effectively in such scenarios. Nonetheless, several techniques can help overcome these challenges. One such technique is cross-lingual transfer learning. This approach involves training a model on a high-resource language and subsequently transferring the acquired knowledge to the low-resource language. Studies demonstrate successful cross-lingual unsupervised transfer learning experiments without using bilingual dictionaries or parallel data, including Uyghur (Xie et al., 2018); Spanish, Dutch, German, Arabic and Finnish (Bari et al., 2020); and Russian and Vietnamese (Le and Burtsev, 2019).

2.3 Data Augmentation for NLP

Data augmentation (DA) is a technique of artificially increasing the amount of training data using existing data. Recent survey on DA techniques for NLP (Feng et al., 2021) categorize the approaches into rule-based, interpolation-based, and model-based, listed in order of increasing complexity. Rule-based approaches are simple manipulations such as synonym replacement, random insertion, random swap, and random deletion (Wei and Zou, 2019) to create new data points. Interpolation-based approaches interpolate or fuse the input and label of a real example to create new data points (Zhang et al., 2017). Lastly, model-based approaches include the use of models for DA such as backtranslation (Sennrich et al., 2016) which generates new data by translating the original input to another language and back to the original language, and the use of generative models such as GPT-2 to generate new data (Anaby-Tavor et al., 2019).

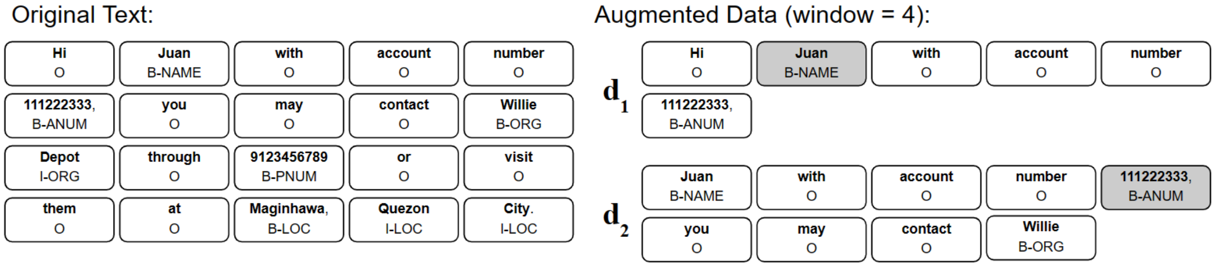

Figure 1: Example of annotated data using IOB scheme (left). Example of data augmentation (right) by extracting the sub-string using a fixed-sized window with the entity as the center (shown as shaded part).

3 Practical Approaches

3.1 Transfer Learning

By employing transfer learning, we utilize pre-trained models to develop task-specific models that perform effectively even with limited training data. Due to the low-resource nature of Filipino, we focus on models that have demonstrated success in similar situations (Nguyen and Chiang, 2017) (Nag et al., 2023). We experimented with different general-purpose pretrained models including cross-lingual language models (Conneau et al., 2020), English language models (Devlin et al., 2018; Liu et al., 2019), Filipino language models (Cruz and Cheng, 2022), and also specialized language models that were fine-tuned on telecommunications domain data and CoNLL-2003 dataset. A summary of pretrained models used can be found in Table 1.

3.2 Data Augmentation

The challenge of learning from small datasets is worsened by imbalanced distribution of classes. Handling imbalanced data is especially harder on small datasets because of two reasons: (1) under-sampling the majority class further reduces the already small dataset, (2) oversampling the minority class in such a small dataset can easily lead to overfitting. Backtranslation and synthetic data generation is also not an option since the token-label alignment needs to to be preserved for NER tasks. As seen in Table 2, the dataset shows significant class imbalance. To address this issue, a simple yet effective data augmentation technique is introduced, which preserves token-label alignment and enhances the performance of the NER.

The process is similar to oversampling, but it involves extracting a sub-string using a fixed-sized window with the entity at the center. In this study, a window size of 4 is used for both left and right (see Figure 1). In practice, only specific classes are chosen for data augmentation, while others are skipped. Specifically, the data augmentation is applied to the NAME, LOC, ORG, and ANUM entities, whereas PNUM is excluded due to its relatively higher frequency compared to other entities. The pseudocode can be found at Algorithm 1.

Algorithm 1 Data Augmentation Algorithm

```
$target\_classes \leftarrow$ list of string of classes
$D \leftarrow$ NER dataset
$ND \leftarrow$ empty list
$w \leftarrow$ integer window size
for $sequence$ in $D$ do
    # $s$ is a list of tokens
    $s \leftarrow sequence$
    for $token$ in $s$ do
        $t \leftarrow token$
        # data augmentation
        if $t.label$ in $target\_classes$ then
            $t\_pos \leftarrow t.index$
            $x \leftarrow get\_substring(s, t\_pos, w)$
            append_to_list($ND$ , $x$)
        end if
    end for
end for
```

4 Experimental Setup

4.1 Dataset

The dataset used in fine-tuning the NER model is a collection of 200 social media posts containing Telecom-relevant keywords and were collected from January to March 2022. Personally-identifiable information like phone numbers were masked and were programmatically regenerated. To ensure non-contextual posts and spam-like posts are not included in the subset, only those that have at least 20 and at most 39 tokens were included in the training and evaluation set of the NER model.

The collected posts were manually annotated

Table 3: Summary statistics of languages in the dataset. Mixed refers to a mixture between English, Filipino or regional languages.

Language	Count	Percentage
English	138	69%
Filipino	29	15%
Mixed	33	17%

and each token was given a tag using the Inside-Outside-Beginning (IOB) annotation scheme. IOB is an annotation scheme that discriminates whether a token is a beginning tag (B), an inside-to-end (I) tag, or an outside (O) tag of a sequence of words (Sang and Buchholz, 2000). We are focused in the recognition of (1) NAME for names of people, (2) LOC for location, (3) ORG for names of organizations, (4) PNUM for phone numbers, and ANUM for account numbers. Example annotation using IOB can be seen at Figure 1.

To gain better understanding of the dataset and its adequacy for NER, the distribution and summary statistics of the tags from the annotated dataset are shown in Table 2. To focus on the relative frequencies of the named entities or non-O tags, the O tag is not included in the table. The PNUM comprises the majority of the annotated named entity tags, followed by ORG, LOC, ANUM, and NAME.

Table 3 presents the linguistic composition within the acquired dataset. The data distribution distinctly emphasizes the low-resource characteristic of the Filipino language, as it occupies a notably small proportion compared to English. Specifically, English being the dominant linguistic component, representing 69% of the entire dataset, followed by instances of mixed language usage, with Filipino trailing behind. The corpus primarily manifests in the form of business-oriented English. However, instances involving complaints and troubleshooting requests tend to exhibit a greater presence of Filipino or mixed Filipino-English expressions — especially important in this domain. This observation underscores the necessity of the development of systems with the capacity to effectively accommodate linguistic variations across diverse domains.

4.2 Pretrained Language Models

We experimented with pretrained models trained on Tagalog, English, Tagalog-English, and Multilingual data. For Tagalog models, the Tagalog-RoBERTa-Base model (Cruz and Cheng, 2022) was used, which was trained with a Masked Language Modeling (MLM) objective using the TLUnified dataset (Cruz and Cheng, 2020). Two size variants were trained for Tagalog-RoBERTa following the original RoBERTa paper, i.e., a base model with 110M parameters and a large model with 330M parameters. The Tagalog RoBERTa models have been shown to outperform existing baselines specifically on different classification tasks.

For English models, the BERT-Base-Cased model (Devlin et al., 2018) and RoBERTa-Base (Liu et al., 2019) was used. BERT was pretrained on a massive dataset of unlabeled text, consisting of 3.3 billion words. The pretraining process consists of MLM and Next Sentence Prediction (NSP) objectives. BERT was evaluated on a variety of natural language processing tasks, including question answering, natural language inference, and sentiment analysis. RoBERTa is an improvement over BERT which was only trained on MLM objective and on significantly bigger data than what was used in BERT. We also experimented with BERT[1] fine-tuned on the CoNLL-2003 dataset.

For multilingual models, the XLM-RoBERTa model was used (Conneau et al., 2020). XLM-RoBERTa is a multilingual transformer model pretrained with MLM objective. The model was trained on 100 languages, including Filipino and English, using cleaned CommonCrawls dataset (Conneau et al., 2020). Both the base and the large variants of the model, with 270M and 550M parameters respectively, were used in this study. CoNLL-2003 dataset fine-tuned versions of the base[2] and large[3] models were also included in the experiment.

Lastly, for the telecom domain-specific model, we trained Telecom-RoBERTa-Base, an in-house developed language model with parameter size of 125M. It was trained on Philippine telecom domain corpora composed of social media extracts from January to September 2022 with an MLM objective using the Tagalog-English-RoBERTa model as the base model (Velasco et al., 2022). The Tagalog-English-RoBERTa model was trained on the CO-HFIE dataset (Velasco et al., 2022) consisting of

[1] https://huggingface.co/kamalkraj/bert-base-cased-ner-conll2003

[2] https://huggingface.co/Yaxin/xlm-roberta-base-conll2003-ner

[3] https://huggingface.co/xlm-roberta-large-finetuned-conll03-english

297 million tokens containing both Tagalog and English texts from various domains such as social media, online forums, news sites, and Wikipedia.

4.3 Evaluation setup and hyperparameter search settings

We follow the evaluation metrics of CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) which uses F1, precision, and recall. Precision is the percentage of correct answers found by the model. Recall is the percentage of entities correctly found by the model. F1 is the harmonic mean of precision and recall which assesses the class-wise performance of the model. A prediction is only correct if it is an exact match with the gold standard. The reported F1 refers to the Macro F1 and will be referred to as F1 for the rest of the paper.

The conventional approach found in academic research for evaluating named entities in benchmarks, as shown in CoNLL-2003, requires a correct prediction to be the exact span of the entity. However, in applied scenarios, the exact span is not always necessary for NER systems to be useful. Thus, this study seeks to explore both the conventional full-span prediction metrics and the partial prediction metrics, considering the specific requirements and utility within the telecom domain. To properly assess the performance of the model in terms of partial correctness, we relaxed the evaluation criteria by removing the IOB scheme and just considering if the correct entity type is found. The computation for F1, precision, and recall is still the same but without consideration for IOB schema.

K-fold cross validation with $k = 3$ was used across all experiments to complement the small dataset size. K-fold cross validation is a type of cross validation wherein all samples are seen in both training and evaluation after k iterations (Raschka, 2018). In each iteration, the dataset is split into k parts where one part is used in evaluation and the rest in training. The average F1 score across all folds were reported in Table 4 and Table 5.

To set a baseline for comparison, we follow the standard procedure for fine-tuning pretrained models on downstream tasks where only the classification head of the pretrained model was changed and trained. To ensure fair comparison of models, we performed a hyperparameter search to find the optimal set of hyperparameters that maximizes the F1 score for our dataset. We decided to perform a grid search for the learning rate, batch size and epochs. For the rest of the hyperparameters, we set them to the default values of Adam optimizer[4]. The search space are as follows: $learning\_rate$ =[1E-06, 8E-06, 9E-06, 1E-05, 8E-05, 9E-05, 1E-04, 8E-04, 9E-04, 1E-03], $batch\_size$ =[8, 16] and $epochs$ =[10, 15].

For experiments on data augmentation, we used the same hyperparameter search settings for fair comparison. The same settings was also used for K-fold where data augmentation is applied on each training fold, ensuring that no data leakage will occur. The optimal hyperparameters together with the full-span and partial recognition F1 score of the baseline models and with data augmentation are summarized in Table 4.

5 Results

5.1 Baseline results

From the baseline results on Table 4, it can be observed that both XLM-RoBERTa-Base$_{CoNLL}$ and XLM-RoBERTa-Large$_{CoNLL}$ scored the highest in terms of overall F1 scores for full and partial benchmarks respectively. The multilingual nature of the XLM models combined with fine-tuning on English CoNLL allows these models to leverage a robust baseline of knowledge prior to fine-tuning to a low resource target domain of Filipino social media.

Meanwhile language-specific models BERT-Base-Cased and Tagalog-RoBERTa exhibit similar levels of performance, with Tagalog-RoBERTa demonstrating a slight performance advantage, potentially attributed to its increased ability to understand the domain-specific corpus.

It can also be observed that the models BERT-Base-Cased, XLM-RoBERTa-Base and XLM-RoBERTa-Large models are notably improved in both the full and partial benchmarks when they are subject to fine-tuning on English CoNLL-2003 as a source domain prior to fine-tuning on the target domain. Supporting the works of Jia et al. (2019), these results suggest that there is a level of transferable cross-domain knowledge resulting in enhanced performance across tasks.

Despite undergoing fine-tuning on the domain specific corpus, the Telecom-RoBERTa-Base ranked last in both full and partial benchmarks. The

[4] https://pytorch.org/docs/stable/generated/torch.optim.Adam.html; Default parameters for betas (0.9, 0.999), for epsilon (1E-08), weight decay (0)

Table 4: Results of baseline and augmented NER models. Values inside the parentheses are the absolute % change in full-span recognition (F1 Full) and partial recognition (F1 Partial) performance from baseline after applying data augmentation.

Model	**Learning rate**	**Epochs**	**Batch size**	**F1 Full**	**F1 Partial**
Telecom-RoBERTa-Base	1E-04	15	8	0.413	0.626
Telecom-RoBERTa-Base$_{AUG}$	9E-05	15	8	0.499 (+8.6%)	0.695 (+6.94%)
Tagalog-RoBERTa	1E-04	15	16	0.537	0.701
Tagalog-RoBERTa$_{AUG}$	8E-05	15	16	0.547 (+1.0%)	0.707 (+0.57%)
BERT-Base-Cased	8E-05	15	16	0.532	0.703
BERT-Base-Cased$_{AUG}$	1E-04	15	16	0.588 (+5.6%)	0.716 (+1.31%)
BERT-Base-Cased$_{CoNLL}$	1E-04	15	8	0.572	0.717
BERT-Base-Cased$_{CoNLL-AUG}$	8E-05	10	8	0.609 (+3.7%)	0.726 (+0.9%)
XLM-RoBERTa-Base	9E-05	15	8	0.618	0.735
XLM-RoBERTa-Base$_{AUG}$	8E-05	15	16	0.664 (+4.6%)	0.769 (+3.45%)
XLM-RoBERTa-Base$_{CoNLL}$	9E-05	15	8	0.634	0.765
XLM-RoBERTa-Base$_{CoNLL-AUG}$	9E-05	15	16	0.66 (+2.6%)	0.765 (0%)
XLM-RoBERTa-Large	9E-05	15	16	0.645	0.777
XLM-RoBERTa-Large$_{AUG}$	8E-05	15	8	0.683 (+3.8%)	0.787 (+0.97%)
XLM-RoBERTa-Large$_{CoNLL}$	9E-06	15	8	0.673	0.843
XLM-RoBERTa-Large$_{CoNLL-AUG}$	1E-05	10	8	0.708 (+3.5%)	0.839 (-0.42%)

AUG is shorthand for Data Augmentation
$CoNLL$ means the model is tuned on CoNLL 2003 dataset

Table 5: Full-span F1 score for each named entity type. Values inside the parentheses are the absolute % change from baseline after applying data augmentation.

Model	**NAME**	**ORG**	**LOC**	**ANUM**	**PNUM**
Telecom-RoBERTa-Base	0.224	0.457	0.378	0.212	0.794
Telecom-RoBERTa-Base$_{AUG}$	0.474 (+25.06%)	0.418 (-3.96%)	0.408 (+3.05%)	0.362 (+15.03%)	0.832 (+3.82%)
Tagalog-RoBERTa	0.421	0.518	0.417	0.462	0.865
Tagalog-RoBERTa$_{AUG}$	0.413 (-0.84%)	0.523 (+0.46%)	0.515 (+9.78%)	0.422 (-4.05%)	0.865 (-0.05%)
BERT-Base-Cased	0.432	0.549	0.432	0.392	0.857
BERT-Base-Cased$_{AUG}$	0.591 (+15.87%)	0.572 (+2.24%)	0.461 (+2.96%)	0.457 (+6.45%)	0.860 (+0.34%)
BERT-Base-Cased$_{CoNLL}$	0.396	0.616	0.396	0.499	0.881
BERT-Base-Cased$_{CoNLL AUG}$	0.572 (+17.69%)	0.572 (-4.39%)	0.509 (+11.30%)	0.524 (+2.49%)	0.867 (-1.39%)
XLM- RoBERTa-Base	0.563	0.590	0.492	0.545	0.901
XLM- RoBERTa-Base$_{AUG}$	0.688 (+12.52%)	0.570 (-1.93%)	0.610 (+11.79%)	0.543 (-0.196%)	0.909 (+0.83%)
XLM- RoBERTa-Base$_{CoNLL}$	0.526	0.519	0.598	0.620	0.908
XLM- RoBERTa-Base$_{CoNLL-AUG}$	0.585 (+5.89%)	0.603 (+8.42%)	0.608 (+1.03%)	0.584 (-3.57%)	0.920 (+1.20%)
XLM- RoBERTa-Large	0.641	0.550	0.581	0.603	0.848
XLM- RoBERTa-Large$_{AUG}$	0.595 (-4.66%)	0.612 (+6.27%)	0.574 (-0.76%)	0.720 (+11.63%)	0.913 (+6.54%)
XLM-RoBERTa-Large$_{CoNLL}$	0.679	0.559	0.607	0.630	0.890
XLM-RoBERTa-Large$_{CoNLL-AUG}$	0.662 (-1.64%)	0.639 (+7.95%)	0.627 (+1.97%)	0.699 (+6.91%)	0.912 (+2.21%)

AUG is shorthand for Data Augmentation
$CoNLL$ means the model is tuned on CoNLL 2003 dataset

Table 6: Range of per-entity full-span F1 performance increase and decrease after training on augmented data.

	NAME	**ORG**	**LOC**	**ANUM**	**PNUM**
Performance increase	5.83% to 25.06%	0.46% to 8.43%	1.03% to 11.79%	2.493% to 15.03%	0.341% to 6.54%
Performance decrease	-0.84% to -4.67%	-1.93% to -4.39%	-0.756%	-0.196% to -4.05%	-0.051% to -1.39%

decline in performance could be attributed to the decreased ability to generalize when trained on relatively poor data from social media sources. As demonstrated by Raffel et al. (2020), corpora that have undergone a filtering process to extract clean data exhibit improved performance compared to their unfiltered counterparts, suggesting that the quality of the corpus significantly influences the performance of models.

5.2 Data augmentation results

From the data augmentation results on Table 4, it can be observed that training with augmented dataset shows consistent performance improvement across all models over their baseline counterparts. Across all experiments with data augmentation, XLM-RoBERTa-Large$_{CoNLL}$ scored the highest on overall F1 score for full benchmark but not in partial benchmark where it is outperformed by its baseline counterpart.

In terms of F1 Full improvement, Telecom-RoBERTa-Base$_{AUG}$ has the highest improvement of 8.6% followed by BERT-Base-Cased$_{AUG}$ (5.6%) and XLM-RoBERTa-Base (4.6%). For absolute F1 Partial improvement, Telecom-RoBERTa-Base$_{AUG}$ has the highest improvement of 6.94% followed by XLM-RoBERTa-Base$_{AUG}$ (3.45%) and BERT-Base-Cased$_{AUG}$ (1.31%).

The comparison between XLM-RoBERTa-Base$_{AUG}$ and XLM-RoBERTa-Large models yields a notable observation. The former, trained with augmented data, achieves a F1 Full score of 0.664, surpassing the 0.645 attained by the latter, without augmentation. Remarkably, the augmented model demonstrates a 1.9% improvement in F1 score relative to its larger counterpart, despite having a significantly smaller parameter size. This finding highlights the data augmentation technique's efficacy in enhancing model performance, regardless of its scale, while also closing the performance gap between smaller and larger models. Notably, this result carries practical implications as smaller models are generally preferred as deployment targets in industry applications due to lower costs and faster inference times (Menghani, 2021).

5.3 Per-entity performance

In the telecoms domain, the NAME, LOC, and ORG were identified to be the most important named entities. This means that if several models have similar F1 scores, there will be preference towards the model with higher F1 score on NAME, LOC, and ORG.

Based on Table 5, it can be observed that data augmentation did have an effect to the per-entity performance across all models since 5 of 8 models have seen improved F1 for NAME entity, 5 of 8 for ORG, 7 of 8 for LOC, 5 of 8 for ANUM, and lastly 6 of 8 models for PNUM. The range of performance increase and decrease per-entity is summarized at Table 6. While there's no consistent pattern on which entity the performance decreases, it can be observed that the performance decrease ranges from -0.051% to -4.67% while the improvement in per-entity performance ranges from 0.341% to 25.06%. From this observation, we can conclude that even though training on augmented data leads to performance decrease on some entities, it is relatively small compared to the performance increase that it gains for other entities.

6 Conclusion

Our research has demonstrated the effectiveness of utilizing cross-lingual task learning to develop a named entity recognizer for a low-resourced language. Although existing crosslingual language models can be employed, the resulting NER models often require significant storage capacity. Alternatively, leveraging pretrained language models in related languages or utilizing fine-tuned language models specifically trained for the same task can be explored. Additionally, addressing data imbalance through data augmentation techniques can enhance performance.

References

Ana Patricia T. Alfonso, Illuminada Vivien R Domingo, Mary Joy F Galope, Ria A. Sagum, Jobert T Villegas, and Rachelle B Villar. 2013. Named entity recognizer for filipino text using conditional random field. *International Journal of Future Computer and Communication*, pages 376–379.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2019. Not enough data? deep learning to the rescue! *CoRR*, abs/1911.03118.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7415–7423.

Jonalyn Castillo, Marck Augustus L. Mateo, Antonio D. C. Paras, Ria A. Sagum, and Vina Danica F. Santos. 2013. Named entity recognition using support vector machine for filipino text documents. *International Journal of Future Computer and Communication*, pages 530–532.

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

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *CoRR*, abs/2005.02068.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Karen Mae L. Eboña, Orlando S. Llorca Jr., Genrev P. Perez, Jhustine M. Roldan, Iluminda Vivien R. Domingo, and Ria A. Sagum. 2013. Named-entity recognizer (NER) for filipino novel excerpts using maximum entropy approach. *Journal of Industrial and Intelligent Information*, 1(1):63–67.

Asif Ekbal and Sivaji Bandyopadhyay. 2007. A hidden markov model based named entity recognition system: Bengali and hindi as case studies. In *Pattern Recognition and Machine Intelligence*, pages 545–552, Berlin, Heidelberg. Springer Berlin Heidelberg.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. *CoRR*, abs/2105.03075.

Joshua Andre Huertas Gonzales, J-Adrielle Enriquez Gustilo, Glenn Michael Vequilla Nituda, and Kristine Mae Monteza Adlaon. 2023. Developing a hybrid neural network for part-of-speech tagging and named entity recognition. In *Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference*, AICCC '22, page 7–13, New York, NY, USA. Association for Computing Machinery.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

The Anh Le and Mikhail S. Burtsev. 2019. A deep neural network model for the task of named entity recognition. *International Journal of Machine Learning and Computing*, 9(1):8–13.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Evan Dennison S. Livelo, Andrea Nicole O. Ver, Jedrick L. Chua, John Paul S. Yao, and Charibeth K. Cheng. 2017. A hybrid agent for automatically determining and extracting the 5ws of filipino news articles. In *Proceedings of the IJCAI Workshop on Semantic Machine Learning (SML 2017) co-located with 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), Melbourne, Australia, August 20, 2017*, volume 1986 of *CEUR Workshop Proceedings*, pages 50–56. CEUR-WS.org.

Gaurav Menghani. 2021. Efficient deep learning: A survey on making deep learning models smaller, faster, and better.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Transfer learning for low-resource multilingual relation classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Sebastian Raschka. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *CoRR*, abs/1811.12808.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *CoRR*, cs.CL/0009008.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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 - Volume 4*, CONLL '03, page 142–147, USA. Association for Computational Linguistics.

Dan John Velasco, Axel Alba, Trisha Gail Pelagio, Bryce Anthony Ramirez, Jan Christian Blaise Cruz, and Charibeth Cheng. 2022. Automatic wordnet construction using word sense induction through sentence embeddings.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412.