Bridging the Gap: Transfer Learning from English PLMs to Malaysian English

Mohan Raj Chanthran[1], **Lay-Ki Soon**[1*], **Ong Huey Fang**[1], **and Bhawani Selvaretnam**[2]
[1]School of Information Technology, Monash University Malaysia
{mohan.chanthran, soon.layki, ong.hueyfang}@monash.edu
[2]Valiantlytix
bhawani@valiantlytix.com

Abstract

Malaysian English is a low resource creole language, where it carries the elements of Malay, Chinese, and Tamil languages, in addition to Standard English. Named Entity Recognition (NER) models underperform when capturing entities from Malaysian English text due to its distinctive morphosyntactic adaptations, semantic features and code-switching (mixing English and Malay). Considering these gaps, we introduce MENmBERT and MENBERT, a pre-trained language model with contextual understanding, specifically tailored for Malaysian English. We have fine-tuned MENmBERT and MENBERT using manually annotated entities and relations from the Malaysian English News Article (MEN) Dataset. This fine-tuning process allows the PLM to learn representations that capture the nuances of Malaysian English relevant for NER and RE tasks. MENmBERT achieved a 1.52% and 26.27% improvement on NER and RE tasks respectively compared to the bert-base-multilingual-cased model. Although the overall performance of NER does not have a significant improvement, our further analysis shows that there is a significant improvement when evaluated by the 12 entity labels. These findings suggest that pre-training language models on language-specific and geographically-focused corpora can be a promising approach for improving NER performance in low-resource settings. The dataset and code published in this paper provide valuable resources for NLP research work focusing on Malaysian English.

1 Introduction

With the recent proliferation of Large Language Models (LLMs), the usage of Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), ALBERT (Lan et al., 2020) has been overshadowed. However, PLM has shown some significant improvements when further pre-trained in domain specific (Chalkidis et al., 2020; Lee et al., 2019; Araci, 2019; Huang et al., 2020) or language specific corpus (Martin et al., 2020; Chan et al., 2020; Antoun et al., 2021; Vamvas et al., 2023), and subsequently fine-tuned on NLP tasks. The adaptability of pre-trained language model to diverse languages and dialects has enabled its application to specific linguistic contexts, including Malaysian English, a unique and culturally rich variant of the English language.

Malaysian English is also categorized as a creole language due to its distinct characteristics that include loanwords, compound words, and the derivation of new terms from Malay, Chinese, and Tamil, in addition to the Standard English (Chanthran et al., 2024). Existing state-of-the-art (SOTA) solutions do not produce satisfactory outcomes for downstream tasks performed in Malaysian English (Chanthran et al., 2024). This is mainly due to the morphosyntactic and semantical adaption nature of Malaysian English. Hence, there is a need to improve this SOTA in order to support effective processing of Malaysian English texts.

This work investigates the effectiveness of English pre-trained language models (PLMs) to the low resource language like Malaysian English, in downstream task, particularly Named Entity Recognition (NER), and Relation Extraction (RE) performed on Malaysian English. Our findings contribute to the growing body of research on promoting inclusivity in NLP by exploring the applicability of PLMs to non-standard English varieties.

The contributions of this paper are as follows:

1. Multilingual Pre-trained Model for Malaysian English: We introduce MENmBERT, a BERT-based model pre-trained on a Malaysian English News (MEN) Corpus. MEN Corpus comprises 14,320 articles,

*Corresponding Author.

facilitating research on applying PLMs to Malaysian English (Chanthran et al., 2024). This model will be made public to develop resources and facilitate NLP research in Malaysian English. The code for this experiment, and dataset have been published in https://github.com/mohanraj-nlp/MEN-Dataset/tree/pretrained-lm.

2. Fine-Tuned and Evaluated NER and RE on Malaysian English: We evaluated the effectiveness of fine-tuned MENmBERT for NER and RE tasks on a benchmark MEN-Dataset. MEN-Dataset contains 200 news articles with 6,061 annotated entities and 4,095 relation instances (Chanthran et al., 2024). This analysis demonstrates the applicability of transfer learning from English PLMs to Malaysian English NLP tasks.

This paper is structured as follows. Section 2 explores the utilization of pre-trained language models in both English and non-English scenarios. In Section 3, we dive deeper into the pre-training and fine-tune methodologies of MENmBERT and MENBERT. Section 4 we will discuss on the results of fine-tuned NER and RE. Finally, Section 5 concludes the work and shares potential enhancement as the future work.

2 Related Work

2.1 Pre-Trained Language Model for Non-English Context

Language-specific Pre-trained Language Models (PLMs) are essential to handle complex languages and improve performance on downstream NLP task specific to particular language. Considering this, AraBERT has been proposed to address the morphological and syntactic differences in the Arabic language compared to other languages, as Arabic shares very little with Latin-based languages and has unique characteristics (Antoun et al., 2021). Multilingual models to learn representations for multiple languages simultaneously resulted in little data representation and small language-specific vocabulary for Arabic, hindering performance compared to a single-language model. (Antoun et al., 2021) overcome these challenges, the researchers pre-trained AraBERT specifically for the Arabic language to capture the contextualized representations needed for Arabic NLP tasks. By customizing the model for Arabic and optimizing factors such as data size, vocabulary size, and pre-processing techniques, AraBERT was able to achieve state-of-the-art performance on various Arabic NLP tasks. The pre-training has been completed with 70 million sentences and around 24GB of textual Arabic data. AraBERT performs better in downstream tasks like Sentiment Analysis, Named Entity Recognition (NER), and Question Answering (Antoun et al., 2021). Antoun et al. (2021) compared the AraBERT fine-tuned model to SOTA and M-BERT (also know Multilingual BERT), the results shows AraBERT performing better than mBERT or SOTA.

Following the success of AraBERT in Arabic NLP, similar approaches can be applied to other low-resource languages with unique characteristics. One such example is KinyaBERT, a recent model specifically designed to address the challenges of Natural Language Processing (NLP) tasks in Kinyarwanda (Nzeyimana and Niyongabo Rubungo, 2022). KinyaBERT has been implemented with a two-tier BERT architecture that token-level morphology encoder and sentence/document level encoder. By dividing the model's processing into these two tiers, KinyaBERT aims to effectively capture both the fine-grained morphological details of individual tokens and the broader contextual information present in the input text. The pre-training task has been completed with 16 million and 2.4GB of Kinyarwanda language texts. KinyaBERT is evaluated on NLP downstream tasks such as NER, News Categorization Task (NEWS) and Machine-Translated GLUE Benchmark. From the evaluation, KinyaBERT has outperformed the baseline model like BERT Base Pre-trained on Kinyarwanda Corpus (BERT BPE), BERT Tokenized by Morphological Analyzer (BERT MORPHO) and XLM-R.

Similarly to AraBERT and KinyaBERT, SwissBERT has been proposed specifically for national languages of Switzerland (Vamvas et al., 2023). SwissBERT trained using a combination of domain adaptation, language adaptation, and multilingual approaches. SwissBERT has been fine-tuned for NER task and it was able to outperform the baseline model which has been further pre-trained. SwissBERT has undergone several key adaptations and additions to make it impact for processing Switzerland-related text, this includes:

1. Multilingual Adaptation: SwissBERT is trained on a corpus of more than 21 million Swiss news articles in the national languages

of Switzerland, including German, French, Italian, and Romansh Grischun.

2. Custom Language Adapters: SwissBERT utilizes custom language adapters in each layer of the transformer encoder for the four national languages of Switzerland.

3. Switzerland-Specific Subword Vocabulary: To further enhance its performance on Switzerland-related text, SwissBERT is equipped with a Switzerland-specific subword vocabulary.

SwissBERT's superior performance in Switzerland-related tasks, such as Named Entity Recognition and Stance Detection, highlights its efficacy in handling diverse linguistic content specific to Switzerland.

Devlin et al. (2019) has further pre-trained multilingual BERT (M-BERT) with 104 languages Wikipedia corpus. Wang et al. (2020) suggests enhancing M-BERT by pre-training with low-resource corpora, as 11 languages are not covered in the current 104 languages. The research found that fine-tuning M-BERT (E-MBERT) on low-resource language corpora enhanced NER task performance. Wu and Dredze (2020) found that multilingual BERT (mBERT in Wu and Dredze (2020)) may not perform well in low-resource languages. This inconsistency may be attributed to the fact that mBERT is trained with a multitude of languages. Pre-training mBERT for a low resource language model can negatively impact the performance compared to training a monolingual BERT model for that language. Findings in related works have inspired us to establish better evaluation and model selection criteria for the development of a pre-trained language model for Malaysian English.

3 MENmBERT and MENBERT

3.1 Overview

Chalkidis et al. (2020) discussed about two possible further pre-training strategies:

1. Continued / Further Pre-training (FP): FP creates domain- or language-specific BERT models. FP utilises pre-trained model parameters, saving time and data (Kalyan et al., 2021).

2. Pre-Training from Scratch (SC): Pre-training from scratch lets you train the model with a lot of data. Pre-training the model from scratch will utilise existing language model architecture and parameters (Kalyan et al., 2021).

The difference between two approach is, FP has been pre-trained with generic corpora like BookCorpus, and English Wikipedia (Devlin et al., 2019) while SC has not been pre-trained with any corpus. With this in mind, we proposed to explore several further pre-training strategies:

1. MENBERT-FP: Further pre-train bert-base-cased model with MEN-Corpus

2. MENmBERT-FP: Further pre-train bert-base-multilingual-cased model with MEN-Corpus

3. MENBERT-SC: We pre-train bert-base-cased from scratch with MEN-Corpus

Malaysian English features loan words, compound blend and derivations of new terms from multiple languages local language, making multilingual BERT an effective model for understanding the contexts of news articles. The need to further pre-train BERT, mBERT and train BERT from scratch stems from our hypothesis that PLM with rich language-based understanding will improve the performance of NER and RE after being fine-tuned. Section 3.2 presents the pre-training setup and hyperparameters used, while Section 3.3 explains the model fine-tuning tasks.

3.2 Further Pre-Training MENmBERT and MENBERT

Python library Transformers (Wolf et al., 2020) has been used to pre-train BERT. For MENmBERT and MENBERT we have used bert-base-multilingual-cased and bert-base-cased respectively. Since MENBERT-SC was trained from scratch, we used bert-base-cased architecture. We generated our own vocabulary using BertWordPieceTokenizer for MENBERT-SC, meanwhile for MENmBERT and MENBERT we have used BertTokenizer. We selected the hyperparameter combination with the lowest training loss. Table 1 lists the important hyperparameters used to train the models. Section 4 details the pre-training results and some analyses.

3.3 Fine-Tuning MENmBERT and MENBERT

Fine-Tune is an adaptation method to train pre-trained model for any NLP downstream task (Kalyan et al., 2021). Three pre-trained model

Hyperparameters	MENBERT-FP	MENmBERT-FP	MENBERT-SC
epoch	30	30	30
batch_size	32	16	32
learning_rate	5e-5	5e-5	5e-5
weight_decay	0.001	0.001	0.001
max_sequence_length	512	512	512

Table 1: Hyperparameters used to train MENBERT-FP, MENmBERT-FP, MENBERT-SC

MENBERT-FP, MENmBERT-FP, and MENBERT-SC were fine-tuned for NER and RE using MEN-Dataset. Additionally, we also fine-tuned pre-trained models bert-base-cased and bert-base-multilingual-cased. Fine-tuning on pre-trained and further pre-trained models helps us to compare the performance and validate our hypothesis (see Section 3.1).

3.3.1 Named Entity Recognition

We used the Python library Transformers Wolf et al. (2020), specifically the BertForTokenClassification module, for fine-tuning. As suggested by Devlin et al. (2019), we went through hyperparameter optimization to find an optimal hyperparameter for fine-tuning. We leveraged on WandB (Biewald, 2020) for hyperparameter optimization and logging. We used [2e-5, 5e-5] for the learning_rate, [10, 20, 30] for num_train_epochs, [0.01, 0.001, 0.0001] for weight_decay, and finally [4, 8, 16] for the per_device_train_batch_size.

We used a grid-based search approach to find an optimal hyperparameter with maximum F1-Score and minimum evaluation loss. Table 2 provides the hyperparameters used to fine-tune for NER. The MEN-Dataset is split into training (75%), test (10%) and validation (15%), with total entities of 5065, 453 and 618 respectively. Models fine-tuned with optimal hyperparameter were evaluated using the validation set, and discussed in the following Section 4.2.1.

Hyperparameters	epoch	batch_size	learning_rate	weight_decay
bert-based-cased	20	4	5e-5	0.0001
MENBERT-FP	30	4	5e-5	0.01
mbert-based-cased	30	4	5e-5	0.01
MENmBERT-FP	30	4	5e-5	0.01
MENBERT-SC	30	4	5e-5	0.01

Table 2: Optimal Hyperparameters used to fine-tune bert-base-cased, bert-base-multilingual-cased, MENBERT-FP, MENmBERT-FP, MENBERT-SC for NER

3.3.2 Relation Extraction

To efficiently fine-tune PLM for RE on the MEN-Dataset, we leveraged existing fine-tuning code for document-level relation extraction[1]. We then carefully modified this code to accommodate the specific characteristics and labeling scheme of the MEN-Dataset. Similarly like NER we went through hyperparameter optimization to find an optimal hyperparameter for fine-tuning. We used [2e-5, 5e-5] for the learning_rate, [10, 20, 30] for num_train_epochs, [0.01, 0.001, 0.0001] for weight_decay, and finally [4, 8, 16] for the per_device_train_batch_size. Table 3 provides the hyperparameters used to fine-tune for RE.

Hyperparameters	epoch	batch_size	learning_rate	weight_decay
bert-based-cased	30	4	5e-5	0.1
MENBERT-FP	30	4	5e-5	0.1
mbert-based-cased	30	4	5e-5	0.1
MENmBERT-FP	30	4	5e-5	0.1
MENBERT-SC	30	4	5e-5	0.1

Table 3: Optimal Hyperparameters used to fine-tune bert-base-cased, bert-base-multilingual-cased, MENBERT-FP, MENmBERT-FP, MENBERT-SC for RE

MEN-Dataset has relation labels adapted from prominent RE dataset like DocRED (Yao et al., 2019) and ACE-2005 (Walker, 2005). There are 84 relation labels that are adapted from DocRED, and 16 relation labels from ACE-2005. For this study, we concentrated on the relation labels originating from the DocRED dataset. One of the reason for our decision are due to Label Distribution. The DocRED labels constitute the majority within the MEN-Dataset. Focusing on these prevalent labels allows the model to learn robust representations for the most frequently occurring relation types, leading to potentially better performance on tasks involving these relations. Apart from that, we have also excluded a special relation label "NO_RELATION", as they are used is to indicate entities that might have a relation but not captured by the predefined relation set (Chanthran et al., 2024). Including "NO_RELATION" could introduce noise or ambiguity during model training.

MEN-Dataset has a total of 2,237 relation instances adapted from DocRED relation labels, distributed across training (1,693), testing (267), and validation (277) sets. To ensure a comprehensive representation of relation labels during training, we employed a stratified sampling approach on the MEN-Dataset. While the original split allocates 75%, 10%, and 15% for training, validation, and

[1] DocRED_Bert Github Link

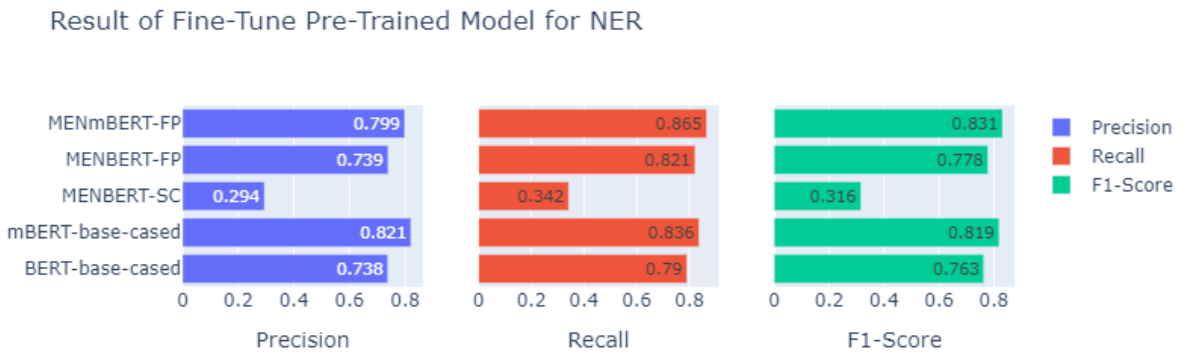

Figure 1: Precision, recall, and F1-score calculated for NER on the MEN-Dataset validation set.

testing, respectively, our stratified sampling guarantees that all relation labels are present in the training data. The result and analysis of fine-tuned PLM for RE have been discussed in Section 4.2.2.

4 Experiment Result and Analysis

4.1 Further Pre-trained Language Model

(Salazar et al., 2020; Kauf and Ivanova, 2023) suggests a "pseudo-log-likelihood" score calculated by masking tokens individually. The score is computed by summing the log-losses at the different masked positions. However, we are more interested in how accurately the models predict the masked token. This will help us to understand PLM models understanding and contextual awareness. We have collected 100 sentences from Malaysian English news article platform and we have done validation to ensure those sentence not part of our pre-training MEN-Corpus. In the first 70 sentences, one token from the local language, such as Bahasa Malaysia, has been randomly masked. In the remaining 30 sentences, one token of Standard English has been masked. We employ accuracy metrics to assess each model's effectiveness, providing a clear differentiation in their performance. The results demonstrate that additional pretraining on language-specific data significantly enhances the models' predictive capabilities, underscoring the importance of tailored training for improved language understanding.

For each pre-trained model, we calculated the accuracy of correctly predicted masked tokens. We used bert-based-cased and bert-base-multilingual-cased as baseline to investigate the improvement made by further pre-trained model. Based on

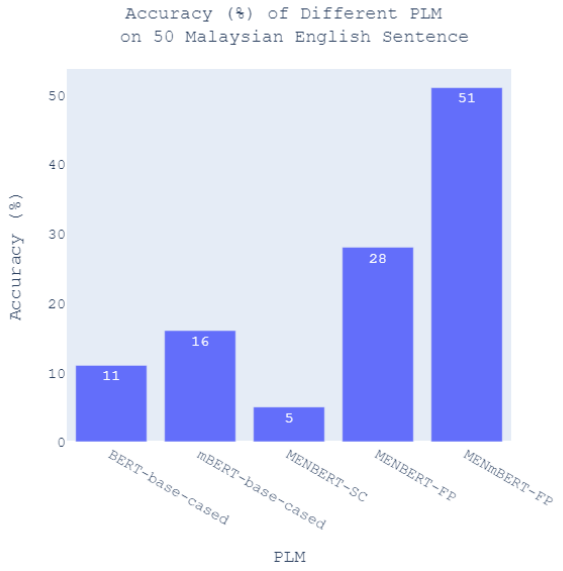

Figure 2: Accuracy of different pre-trained model predicting masked tokens in 50 Malaysian English Sentence.

the result in Figure 2, we have identified that MENmBERT-FP has highest accuracy on predicting masked token from Malaysian English sentence. In Table 4, we present some sample of sentences showcasing how different PLMs perform in predicting masked tokens. Here are the findings obtained from the sample result shown in Table 4:

1. MENmBERT-FP: Even though the model did not predict the exact ground-truth token, in some cases it identified semantically similar tokens. For instance, in the Malaysian context, *Bumiputera* often refers to the *Muslim* community. Here, MENmBERT-FP might predict a token related to ethnicity but not strictly syn-

Masked Sentence	Masked Token	Pretrained Model				
		bert-based -cased	mbert-based -cased	menbert-fp	menbert-sc	menmbert-fp
There are three levels of disaster management, the first involves a locality in a district, secondly when more than two districts of a state is involved and the third involves two or three states. So everyone is aware of this," he told a press conference at <MASK> Sri Muda today.	Taman	the	Sri	the	'	Taman
On the Perlindungan Tenang Voucher, he said all eight million recipients of <MASK> Prihatin Rakyat are eligible to receive the voucher worth RM50 announced in Budget 2021 for the benefit of the B40 group.	Bantuan	the	the	the	the	Anugerah
Ismail Sabri also hoped that ties between Umno and PAS in Bera would remain strong and despite the harsh statement issued by the top leaders of the two parties, priority should be given to unite the Malay Muslims and <MASK>.	Bumiputera	Christians	Muslims	Muslims	,	Muslims
The Ministry of <MASK> Territories (KWP) while ensuring the flood management in Kuala Lumpur is proceeding well.	Federal	New	Protected	Federal	##am	Federal
Hamzah said he had discussed the issue with Inspector-General of Police Tan Sri Acryl Sani <MASK> Sani.	Abdullah	-	.	Abdullah	ser	Abdullah
Hamzah also said police had set up a Tactical Command Centre in <MASK> Langat district in Selangor to coordinate flood relief operations of all units.	Hulu	the	the	Kuala	,	Hulu

Table 4: Some sentences from the MEN-Dataset were used to predict masked tokens using various PLMs. Bold tokens indicate correctly predicted tokens when compared to the ground truth.

onymous with *Muslim*. This highlights the model's ability to capture semantic nuances, even when encountering challenging cases.

2. MENBERT-FP: When we compared the performance of MENBERT-FP with bert-based-cased, we can understand that further-pretraining has improved the performance of language model. The success ratio of bert-based-cased is 0, and once further pre-trained, there is an improvement of +33%.

3. bert-based-cased: Since bert-based-cased has only been trained with English corpus, it was not able to unmask any tokens with compound blend correctly.

4. MENBERT-SC: Based on our observation, MENBERT-SC has produced bad results when unmasking the tokens. Once fine-tune, we will be able to understand better on the performance of the model.

In Section 4.2 we have discussed the performance of pre-trained model once fine-tune them for NER and RE.

4.2 Fine-Tuning Pre-Trained Model

4.2.1 Named Entity Recognition

Figure 1 shows the comparison of Precision, Recall, and F1-Score among five different pre-trained models. To evaluate the performance of further pre-trained models, we also fine-tuned pre-trained model (bert-base-cased, and bert-base-multilingual-cased) as the baseline. Meanwhile in Table 5, we detailed the performance of model by entity labels.

Referring to the results, we observe that MENmBERT-FP achieves the highest F1-Score (0.831), while MENBERT-SC obtains the lowest F1-Score (0.316). Nevertheless, Figure 1 demonstrates an improvement when further pre-training the BERT model, which validated our hypothesis (discussed in Section 3.1). A few other observations from this experiment:

1. MENmBERT-FP has a higher F1-Score (0.831) than mBERT-base-cased (0.819). We observe a +1.52% improvement. Although the improvement is not significant, but when

Entity Label	Total Annotated Entity in MEN-Dataset	Total Annotated Entity in Validation Set	bert-based-cased	mbert-based-cased	menbert-fp	menbert-sc	menmbert-fp
PERSON	1646	108	0.74	0.84	0.79	0.17	**0.86**
LOCATION	1157	150	0.86	0.88	0.87	0.48	**0.91**
ORGANIZATION	1624	262	0.81	**0.89**	0.82	0.29	**0.89**
EVENT	386	30	0.67	0.61	0.66	0.13	**0.77**
PRODUCT	72	6	0.24	**0.33**	0.13	0	0.07
FACILITY	208	27	0.24	0.11	**0.47**	0	0.25
ROLE	485	35	**0.39**	0.4	0.37	0.35	0.6
NORP	114	5	0.88	0.6	**0.89**	0.21	0.57
TITLE	300	4	**0.55**	0	0.43	0.18	0.5
LAW	62	5	0.15	0.12	**0.17**	0.1	0.13
LANGUAGE	0	0	0	0	0	0	0
WORK_OF_ART	7	2	0.1	0.15	0.12	0	**0.16**
Total Entities	6061	634					
Overall Micro F1-Score			**0.763**	**0.819**	**0.778**	**0.316**	**0.831**

Table 5: Fine-Tuned model performance (based on F1-Score) calculated based on validation set for each entity labels.

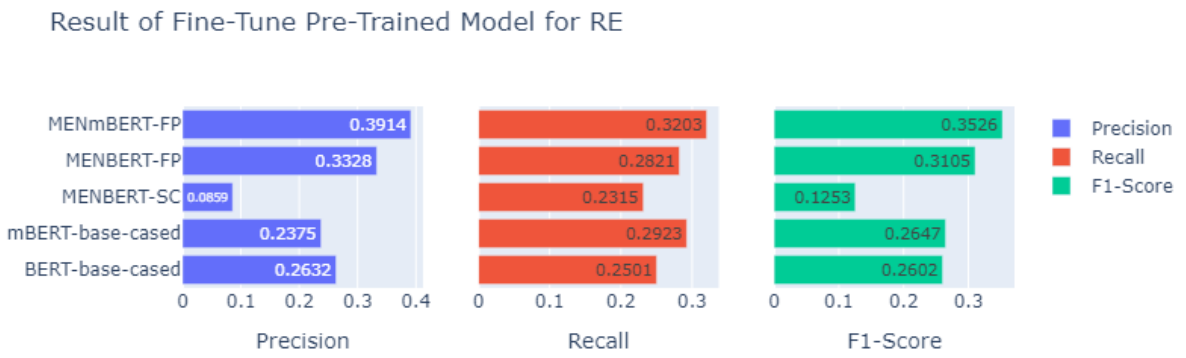

Figure 3: Precision, recall, and F1-score calculated for RE on the MEN-Dataset validation set.

we analyse the F1-Score based on the entity label:

(a) There is an significant improvement for 6 out of 11 entity labels that are evaluated.

(b) On average, the difference in F1-Score between MENmBERT-FP and mBERT-base-cased is +10%.

(c) mBERT-base-cased is only able to achieve on-par in terms of F1-Score with MENmBERT-FP, specifically for entity label ORGANIZATION.

This in-depth observation proves the significant performance of MENmBERT-FP.

2. MENBERT-FP (F1-Score is 0.778) has a higher F1-Score than bert-based-cased (F1-Score is 0.763), with an improvement of +1.94%. However, after further investigated the performance based on entity labels:

(a) MENBERT-FP has only improved the performance of only 7 out of 11 entity labels.

(b) On average, the improvement is only around +2%.

(c) When compared with MENmBERT-FP, MENBERT-FP is able to outperform in terms of F1-Score for 4 out of 11 entity labels. These include entity labels PRODUCT, FACILITY, NORP and LAW.

3. MENBERT-SC has not shown any improvement in the performance of NER. Our evaluation in Section 4.1 also shows it was not able to unmask the tokens correctly.

The experimental results and findings conclude that our fine-tuned MENmBERT-FP has achieved highest F1-Score compared to other pre-trained models. MENmBERT-FP could be used to fine-tune for more NLP downstream tasks, involving Malaysian English, for improved performance.

4.2.2 Relation Extraction

Figure 3 shows F1-Scores compared across five PLMs. Like with NER, we used fine-tuned bert-base-cased and bert-base-multilingual-cased as baselines.

MENmBERT-FP achieved the highest F1-score (0.353), indicating a slight improvement over our baseline PLMs. This suggests that further pre-training on MEN-Dataset has been beneficial for RE on Malaysian English context. Here are some of ur observations from the experiment:

1. MENmBERT-FP has made +33.21% improvement in F1-Score compared to baseline approach mBERT-base-cased. This has been proven significant. Our cross-analysis of NER and RE predictions reveals that the TP relation instances have entity pairs correctly classified by the fine-tuned NER model (from previous analysis). This suggests a significant improvement in RE performance due to the model's enhanced entity prediction capabilities.

2. MENBERT-FP (F1-Score is 0.3105) has a higher F1-Score than bert-based-cased (F1-Score is 0.2602), with an improvement of +19.33%. The analysis of the fine-tuned model's predictions did not reveal any surprising or unexpected patterns. This aligns with the observations from the previous point. Meanwhile, for MENBERT-SC was performed badly when fine-tuned for RE task.

Apart from that, it's important to note that our fine-tuned RE models achieved lower performance compared to reported results on other document-level relation extraction datasets like DocRED. For instance, prior work using a fine-tuned BERT model on DocRED (38,269 relation instances) achieved an F1-score of 54.16 (Dev) and 53.20 (Test) (Wang et al., 2019). Compared with our finding, the overall F1-Score could be not significant due to the nature of MEN-Dataset with a small set of annotation instance.

5 Conclusion

This work introduced MENmBERT, a contextualized language model pre-trained on a Malaysian English corpus. Our experiments demonstrated that fine-tuning MENmBERT on language-specific data significantly improves performance on NER tasks with average of +1.74%. For RE, we have achieved average improvement of +26.27% compare our MENmBERT and MENBERT with baseline PLM's. However, the fine-tuned RE models achieved lower performance compared to reported results on other document-level relation extraction datasets. This suggests that while MENmBERT's entity prediction capabilities benefit RE tasks, further exploration is needed to optimize RE performance in the context of our dataset. This has gaps has suggested us for future work, explore several avenues to improve RE performance. One of the approach involve investigating data augmentation techniques to expand our dataset and improve model training. Beyond that, we will extend our experiment do several other downstream NLP task.

6 Acknowledgements

Part of this project was funded by the Malaysian Fundamental Research Grant Scheme (FRGS) FRGS/1/2022/ICT02/MUSM/02/2.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding. *Preprint*, arXiv:2003.00104.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *Preprint*, arXiv:1908.10063.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *Preprint*, arXiv:2010.02559.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *Preprint*, arXiv:2010.10906.

Mohan Raj Chanthran, Lay-Ki Soon, Huey Fang Ong, and Bhawani Selvaretnam. 2024. Malaysian english news decoded: A linguistic resource for named entity and relation extraction. *Preprint*, arXiv:2402.14521.

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. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Preprint*, arXiv:1904.05342.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing. *Preprint*, arXiv:2108.05542.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. *Preprint*, arXiv:2305.10588.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. *Preprint*, arXiv:1907.11692.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suá rez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. Swissbert: The multilingual language model for switzerland. *Preprint*, arXiv:2303.13310.

Christopher Walker. 2005. *Multilingual Training Corpus LDC2006T06. Web Download. Philadelphia: Linguistic Data Consortium*.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *Preprint*, arXiv:1909.11898.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Workshop on Representation Learning for NLP*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.