Named Entity Recognition for the Irish Language

Jane Adkins[1], Hugo Collins[2], Joachim Wagner[1], Abigail Walsh[1], Brian Davis[1]

[1]ADAPT Centre, Dublin City University, Dublin 9, Co. Dublin, Ireland
[2]School of Computing, Dublin City University, Dublin 9, Co. Dublin, Ireland

Abstract

The Irish language has been deemed 'definitely endangered' (Moseley, 2012) and has been classified as having 'weak or no support' (Lynn, 2023) regarding digital resources in spite of its status as the first official and national language of the Republic of Ireland. This research develops the first named entity recognition (NER) tool for the Irish language, one of the essential tasks identified by the Digital Plan for Irish (Ní Chasaide et al., 2022). In this study, we produce a small gold-standard NER-annotated corpus and compare both monolingual and multilingual BERT models fine-tuned on this task. We experiment with different model architectures and low-resource language approaches to enrich our dataset. We test our models on a mix of single- and multi-word named entities as well as a specific multi-word named entity test set. Our proposed gaBERT model with the implementation of random data augmentation and a conditional random fields layer demonstrates significant performance improvements over baseline models, alternative architectures, and multilingual models, achieving an F1 score of 76.52. This study contributes to advancing Irish language technologies and supporting Irish language digital resources, providing a basis for Irish NER and identification of other MWE types.

1 Introduction

Despite being the first official and national language of the Republic of Ireland, Irish faces a stark reality - it is 'definitely endangered' (Moseley, 2012). Furthermore, it is one of the two European Union languages classified as having 'weak or no support' regarding digital resources (Lynn, 2023). Recognising this challenge, the Digital Plan for Irish (Ní Chasaide et al., 2022) outlines a broad strategy aimed at strengthening technologies tailored to the Irish language. Central to this plan is the recognition of the urgent need for a Named Entity Recognition (NER) tool for Irish. Such a tool not only facilitates various natural language processing (NLP) tasks but also represents a crucial step in providing much-needed essential digital support for the Irish language community (Ní Chasaide et al., 2022). Existing research on Irish MWEs has also highlighted Named Entities (NEs) as requiring special attention (McGuinness et al., 2020), as treatment of these constructions mirrors other MWE types, such as noun compounds (Walsh, 2023). Our research aims to address this gap in Irish language technology by developing a base NER tool specifically tailored for the Irish language, and the construction of the first gold-standard NER-annotated corpus for Irish.

NER is an information extraction task involving the identification of portions of text that refer to NEs and the categorisation of these portions into predefined groups such as location, person, organisation, or other relevant categories. While NER may seem straightforward in its concept, it presents significant challenges. Determining the category of a NE relies not only on the entity itself but also heavily on the context it appears in (Marrero et al., 2013). State-of-the-art NER tools employ neural models that are pre-trained using language modeling tasks, which mitigates the need to have an abundance of annotated corpora (Peters et al., 2018).

For Irish, a NER tool represents a pivotal step towards improving digital content and interfaces in the language, leading to an increase in its use across digital environments. In the development of this tool, a small NER-annotated corpus has been constructed from existing contemporary Irish text. This corpus is to be utilised with pre-trained language models, such as gaBERT (Barry et al., 2022) for the NER task. Due to the size of this corpus and also the size of Irish text in the pre-training of multilingual language models, we use data augmentation approaches to enhance and enrich the

Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 82–96

corpus. Furthermore, we add a conditional random fields (CRF) layer and a bidirectional long short-term memory (Bi-LSTM) CRF layer to leverage contextual understandings captured by the models while incorporating the sequential modelling capabilities of CRFs and Bi-LSTMs (Souza et al., 2020). In this report, we evaluate several BERT (Bi-directional Encoder Representation from Transformers) (Devlin et al., 2019) models for the NER task for Irish, both monolingual and multilingual, and compare the above strategies. The models are tested on both a mixed-length NE test set and also a multi-word NE (MW-NE) test set.

2 Background

2.1 Irish Resources Utilised

This research leverages the following Irish NLP resources:

2.1.1 Irish Universal Dependency Treebank

The Irish Universal Dependency Treebank (IUDT) (Lynn et al., 2023) was first constructed as a conversion of the Irish Dependency Treebank (Lynn, 2016) to the Universal Dependency labeling scheme (Nivre, 2015; Nivre et al., 2016). Each tree in the treebank is manually annotated to include part-of-speech information, syntactic dependencies, and morphological features. While Universal Dependencies do not typically capture NEs in their dependency labels, the IUDT included NE information as part of their annotations (McGuinness et al., 2020). Data from V2.8 of the the mixed-domain treebank was leveraged in our experiments (see Section 3).

2.1.2 gaBERT

gaBERT is a monolingual Irish BERT model trained on over 7.9 million Irish sentences and approximately 161 million words (Barry et al., 2022). It uses the original BERT pretraining parameters (Devlin et al., 2019) along with whole-word masking. Whole-word masking treats entire words as a single unit during the masking process, enabling the model to effectively handle languages with intricate inflection and compounding such as Irish (Barry et al., 2022). When evaluated against off-the-shelf BERT, mBERT and monolingual Irish WikiBERT model, the gaBERT model outperformed these other models in the tasks of dependency parsing and masked-token prediction for Irish (Barry et al., 2022). gaBERT was also fine-tuned for the downstream task of MWE identification (Walsh et al., 2022), achieving higher results compared to a similar fine-tuned mBERT model.

2.2 Techniques for Data Augmentation

2.2.1 Rule-Based Data Augmentation

Rule-based approaches to data-augmentation implement simple manipulations of the data. Multimodal Data Augmentation (Xu et al., 2021) introduces four simple yet effective rule-based data augmentation techniques: synonym replacement, random insertion, random swap, and random deletion. While this work primarily targets text classification, these methods have been widely adapted for NLP tasks due to their potential to enrich training datasets significantly (Xu et al., 2021). This approach was further extended by the introduction of Label-wise Token Replacement, a technique that improves data diversity by replacing a token with another of the same entity type at random (Dai and Adel, 2020). In a study on a different low-resource language—Filipino—researchers used a technique where entities were randomly inserted into sentences or entirely new sentences were crafted with these entities at their core (Chan et al., 2023). Additionally, training data augmentation was utilised by swapping the positions of two randomly selected words within sentences (Xu et al., 2021). Another innovative rule-based method is Entity List Augmentation, where an entity from a list is chosen and the list is expanded by adding other entities of the same type from the training dataset. This approach makes the entity list more comprehensive, thus exposing models to a broader array of entity types (Hu et al., 2023). Mention Replacement is a method proposed by Raiman and Miller for the task of question-answering (Raiman and Miller, 2017) and has been implemented for NER previously (Dai and Adel, 2020), where an entity of the same type is randomly selected to replace the original mention of the entity, similar to the approach of Entity List Augmentation (Hu et al., 2023).

2.2.2 Back-translation Data Augmentation

An innovative technique for augmenting low-resource NER data is described by (Yaseen and Langer, 2021), who employ Back-Translation (BT) on a simulated low-resource dataset of English-German text. The method involves translating a text into another language, and then back into the

original language, to create paraphrased texts that retain the general meaning of the original sentence, while still containing the same NE labels. BT as a data-augmentation technique was also explored by (Sbaty et al., 2021), using Code-Switched data.

2.3 Architecture Augmentation

2.3.1 Addition of a Conditional Random Fields Layer

In recent studies, the incorporation of a CRF layer within BERT, positioned after the softmax layer, has demonstrated notable enhancements in NER performance (Arkhipov et al., 2019; Ge et al., 2022). Additionally, for sequence labelling tasks, the use of a Bi-LSTM-CRF on top of BERT has achieved higher performance than the addition of a linear CRF layer (Liu et al., 2023). Specifically, monolingual BERT models augmented with a CRF layer have exhibited superior performance in precision and F1 scores compared to multilingual BERT models with this augmentation, in the context of Portuguese language tasks (Souza et al., 2020). Furthermore, the integration of a word-level CRF layer has been identified as a method to further amplify the performance of these models (Arkhipov et al., 2019).

3 Data

The data we used for these experiments are comprised of 36,825 tokens and were collected from three sources: the IUDT training set, the IUDT test set (Lynn, 2022), and publicly available transcripts from Dáil proceedings (Houses of the Oireachtas, 2024).

NEs have previously been tagged in the IUDT datasets using a designated label in the morphological features; a simple script was used to filter out these sentences for use as data. The domain is balanced, containing text from news, books, websites, and other sources.

All sentences gathered from Dáil proceedings were published between October 2023 and February 2024 (Houses of the Oireachtas, 2024). This ensured the data postdated the training completion of gaBERT in 2021 (Barry et al., 2022), and so was very unlikely to have been included in the training data for this model. As the Dáil transcripts are largely English text, annotators manually filtered the text for Irish sentences, and identified sentences containing named entities from these for use as data. The Dáil text comprises of formal language often discussing proposed laws, government policies, national issues, and other parliamentary business.

While the IUDT data was tagged with a general "Named Entity" label, none of the above data sources had been previously labeled with fine-grained NE information. Annotation of the data collected was conducted using Label-Studio (Label Studio, 2020) by two annotators and carried out following specified annotation guidelines (see Appendix A), labelling entities as persons (PER), locations (LOC), and organisations (ORG), using an IOB2-tagging scheme (where B indicates the initial token of a named entity span e.g. B-LOC, I indicates a non-inital token of a named entity span e.g. I-PER, and O indicates the token is outside of a named entity span). IOB2 tagging was previously implemented in a similar task for Irish MWE identification (Walsh et al., 2022). The annotators both performed annotation on all 1,249 sentences; all discrepancies were discussed during the annotation process and a decision was made on how to annotate that token/tokens. Overall, there were very few discrepancies in the annotation.

- Training set: 1,009 sentences containing at least one named entity (758 from the IUDT training set and 251 from Dáil proceedings October - December 2023) (Lynn, 2022; Houses of the Oireachtas, 2024)
- Validation set: 100 sentences containing at least one named entity (40 from the IUDT test set and 60 from Dáil proceedings January - February 2024) (Lynn, 2022; Houses of the Oireachtas, 2024)
- Test set: 140 sentences containing at least one named entity (50 from the IUDT test set and 90 from Dáil proceedings January - February 2024) (Lynn, 2022; Houses of the Oireachtas, 2024)
- MWE-Test set: a subset of the test set containing 89 sentences, each containing at least one MW-NE (46 sentences are from the IUDT test set and 43 sentences are from the Dáil proceedings January - February 2024) (Lynn, 2022; Houses of the Oireachtas, 2024)

All datasets were curated carefully to have a balanced spread of named entity types within them. Additionally, the validation and test sets each contained a majority of unseen NEs (see Figure 1),

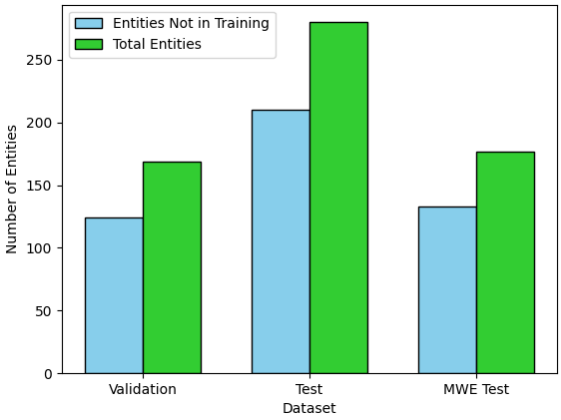

Figure 1: Named entities known or unknown from training in validation and both test sets

with approximately 75% of the NEs being unseen in the training data. This enables us to investigate the capability of testing on unseen NEs, mirroring the focus on unseen MWEs in the PARSEME Shared Task Edition 1.2 (Ramisch et al., 2020). Altogether there was an 80.79%/8.01%/11.20% train/validation/test split implemented for this task. While MW-NEs represent roughly 70% of the NEs in training, the number of single and two-word MW-NEs in the training set represent the majority of the NEs (38% single NEs and 30% two-word MW-NEs).

Table 1: Entity Counts across Datasets

Label	Train	Val	Test	(MWE-test)	Total
LOC	1444	99	275	(147)	1808
ORG	2271	177	225	(148)	2668
PER	1222	150	222	(177)	1590
O	24869	2507	3384	(2197)	30759
Total	29806	2933	4106	(2669)	36845

4 Models for the Task

We employ three BERT (Devlin et al., 2019) models (gaBERT, mBERT and XLMRoberta) to evaluate monolingual and multilingual models on the task of NER for Irish. gaBERT, a monolingual Irish BERT model, proved valuable to our research as it outperformed multilingual models on downstream tasks (Barry et al., 2022). mBERT (multilingual BERT) is a multilingual model containing Irish training data (Devlin et al., 2019). The third model used was XLMRoberta which has achieved state-of-the-art performance on sequence labelling tasks and has outperformed mBERT on cross-lingual classification on low-resource languages (Conneau et al., 2020). Irish is contained in the pre-training data for both mBERT and XLMRoberta, allowing us to evaluate their performances against monolingual gaBERT by fine-tuning these models on the NER task for the low-resource language Irish.

5 Experimental Set-Up

We utilised the AdamW optimiser (Loshchilov and Hutter, 2019) to fine-tune all parameters of the models. Weight-decay was implemented as a regularisation technique to prevent overfitting due to the small size of the training data. A learning rate of 3e-5 and an epsilon value of 1e-8 were chosen to strike a balance between convergence speed and stability during training. The weight decay rate was set to 0.01 for parameters subject to weight decay. Additionally, the maximum gradient norm was set to 1.0 to prevent exploding gradients. This scheduler adjusted the learning rate dynamically throughout the training process, starting with a warm-up phase of 0 steps and gradually linearly increasing the learning rate until reaching the total number of training steps. Training sentences were passed to the models randomly so that the sources of the data were shuffled. Validation and test sentences were passed to the model sequentially. Training epochs were set to 10 with a patience of 2 epochs. If the validation loss increased two times, training stopped and the epoch with the lower validation loss (before the two increases) was used for testing (Prechelt, 2012). The maximum sequence length was set to 256, with training batch size being 32. All models were trained using a T4 GPU in the Google Colab environment.

5.1 Data Augmentation

5.1.1 Random Data Augmentation

Our approach follows closely to that of Mention Replacement and Entity List Augmentation, where an entity pool is created from the entities in the training data and subsequently are added to the training data (Raiman and Miller, 2017; Hu et al., 2023). These entities are added to positions in the text following the IOB2-labelling scheme i.e. located a NE span where the previous and subsequent token are labelled O, then replaced the entire span with an NE or MW-NE (see Figure 2).

5.1.2 Data Augmentation using Back-Translation

To facilitate BT for augmenting our dataset, two models were selected from the Helsinki-

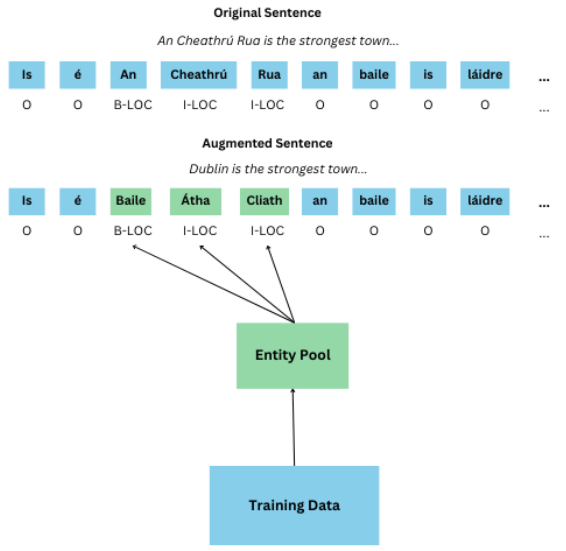

Figure 2: Random data augmentation example

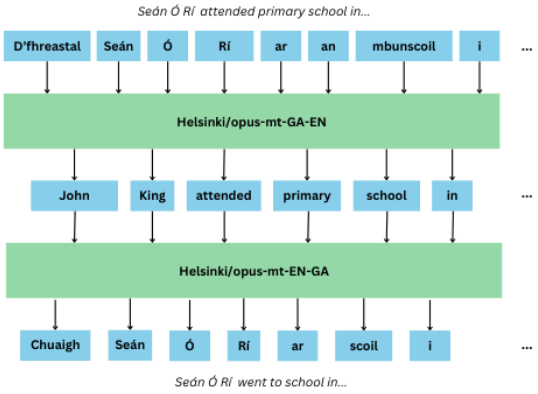

Figure 3: BT in action

NLP OPUS-MT project (Tiedemann and Thottingal, 2020). Specifically, we used the Helsinki-NLP/opus-mt-ga-en model for Irish-to-English translations and Helsinki-NLP/opus-mt-en-ga for the reverse translation from English back to Irish (Tiedemann and Thottingal, 2020). These models were selected based on their results when compared to similar models in the LoResMT 2021 Shared Task (Puranik et al., 2021; Ojha et al., 2021). The Helsinki-NLP/opus-mt-ga-en was used to translate Dáil sentences from the training set to English, and then the Helsinki-NLP/opus-mt-en-ga was used to translate them back to Irish. A total of 256 sentences sentences were backtranslated. Entities were mapped to their corresponding NE-label and the back-translated sentences were added to the training set. See Figure 3 to see backtranslation in action.

5.2 Addition of a CRF Layer

We experiment with adding a CRF layer and a Bi-LSTM CRF layer that are expected to improve the compatibility of predictions with the IOB2-tagging scheme (Ge et al., 2022). As previous work demonstrates, I- entities could not come before a B- entities (B-PER must always be before I-PER etc.) (Ge et al., 2022).

6 Evaluation

The main results from our experiments are presented in Table 2. All metrics were computed using the conlleval script[1] that considers only exact matches. This script focuses specifically on entity-level analysis, allowing for a detailed assessment of the model's ability to recognise distinct types of named entities and is similar to the seqeval (Nakayama, 2018) evaluation library utilised in a previous Irish MWE identification task (Walsh et al., 2022). It computes entity-level precision, recall, and the F1 score for each entity type, which measures the balance between precision and recall.

Additionally, it includes an overall accuracy score and a Non-O accuracy metric, which excludes the non-named entity labels from accuracy calculations to provide a deeper insight into the model's performance in identifying named entities.

7 Results

7.1 Comparison of the Baseline Models

gaBERT outperforms both mBERT and XLM-RoBERTa in most cases across the mixed-length test set and MW-NE test set, particularly excelling in the LOC-type and PER-type entities (see Table 2). While mBERT performs the best on the ORG-type entities across all metrics, XLMRoBERTa surpasses gaBERT on ORG-type entities in terms of recall and F1 scores, though not precision. Overall, monolingual gaBERT demonstrates superior performance compared to its multilingual counterparts, with mBERT and XLMRoBERTa trailing behind by a noticeable margin, except in their handling of the ORG tag.

[1] https://www.cnts.ua.ac.be/conll2000/chunking/output.html

Table 2: Table showing a subset of precision, recall and F-1 scores on the mixed-length NE test set. RDA, CRF, and BT pertain to random data augmentation, conditional random fields, and back-translation models respectively. Overall metrics include scores for O-tagged tokens.

Model	Overall			LOC			ORG			PER		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
gaBERT	71.74	**77.40**	74.46	76.61	77.51	77.06	61.11	72.79	66.44	76.88	**81.10**	78.93
gaBERT RDA CRF	**78.05**	75.05	**76.52**	**81.60**	78.70	**80.12**	**69.70**	67.65	68.66	**81.41**	77.44	**79.38**
mBERT	70.86	73.47	72.14	74.62	75.19	74.90	67.83	72.93	**70.29**	68.40	71.25	69.80
mBERT BT CRF	71.82	75.59	73.66	76.27	**80.56**	78.36	62.07	74.44	67.69	77.83	68.75	73.01
XLMRoberta	69.39	75.37	72.26	74.66	71.71	73.15	57.91	**77.78**	66.39	76.37	78.02	77.19
XLMRoberta RDA	69.18	74.02	71.52	70.13	71.05	70.59	60.16	74.40	66.52	77.92	77.59	77.75

mBERT records the highest number of false negatives, suggesting that it misclasses more tokens with an 'O' label than the other models. XLM-RoBERTa follows closely behind mBERT in identifying entity and non-entity tags, but it has a higher number of false positives. In contrast, gaBERT, though having the lowest true positives and negatives, exhibits fewer false positives and false negatives, reflecting its more conservative approach to entity prediction. While mBERT and XLM-RoBERTa show more balanced performance, they tend to miss certain entities, especially LOC-type entities. Notably, there are no sentences with common incorrect predictions across all three models, indicating the data is unlikely to contain "challenging" NE-types that are mis-categorised by all systems.

7.2 Performance by Entity Type

GaBERT-based models consistently outperform the multilingual models, with only the mBERT BT CRF model scoring higher for LOC-type entities (see Table 2) and mBERT RDA Bi-LSTM CRF scoring higher for ORG-type entities (recall of 79.32, see Appendix C). Additionally, gaBERT-based models appear to be the most robust across all entity types.

Titles or honorifics preceding PER NEs e.g. 'Bean' (Mrs. or Ms.) and 'Aire' (Minister) presented challenges for all models.

Overall, ORG-type entities were the most difficult for all models, particularly seen with low precision scores across all models. One recurring error for ORG-type NEs includes the difficulty models showed when correctly annotating the team names of regional teams e.g. 'Luimneach' (Limerick), which more commonly presents as a LOC-type entity.

MW-NE Test Set: Within multiword NEs, the same trend of gaBERT-based models dominating scores can be seen with PER-type MW-NEs category, with only one exception—mBERT RDA Bi-LSTM CRF achieves the highest precision for PER-type MW-NEs across all models on this test set (87.50, whereas the highest precision for PER achieved by a gaBERT-based model is 86.54 see Appendix D), meaning that this model is the most adept at reducing false positives for MW-NEs in this category. Additionally, mBERT variations appear to perform better on LOC-type and ORG-type MW-NEs rather than gaBERT (mBERT achieves the highest precision for LOC (76.72), mBERT BT CRF the highest recall for LOC (85.47), mBERT Bi-LSTM CRF the highest F1 score for LOC (80.11), both precision (64.84) and F1 score (71.30) for ORG, and mBERT RDA Bi-LSTM CRF the highest recall for ORG (83.89)), indicating that these models may be better at handling multi-word LOC-type and ORG-type MW-NEs than gaBERT-based models.

7.3 Effects of Data Augmentation Methods

Random data augmentation (RDA) negatively impacts recall across all models, with the most significant decline observed in mBERT, particularly for PER-type NEs, where the addition of RDA to mBERT led to the lowest results of all models for this category (precision of 66.51, recall of 58.75, and F1 score of 62.39). However, mBERT RDA sees a slight recall improvement for LOC (+4.61) due to an increase in true positives and decrease in false negatives. Precision also decreases (by 8.38 for the LOC-type NEs), suggesting that augmented data doesn't improve the model's predictive performance and makes them overconfident in their predictions. Exceptions are gaBERT RDA and XLMRoBERTa RDA, which show improvements for both the ORG- (+0.28 and +2.25 respectively and PER- (+3.50 and +1.55 respectively) type NEs. Consequently, F1 scores generally decline (see Ap-

pendix C), indicating that RDA does not enhance performance overall.

RDA on MW-NE Test Set: XLMRoBERTa RDA shows improved recall for LOC and PER, outperforming the baseline (+0.70 and +1.23 respectively) and BT (+2.09 and +0.62 respectively). In contrast, gaBERT RDA and mBERT RDA show recall degradation for PER and LOC (see Appendix D), suggesting that RDA can hinder performance for specific entity types, highlighting again that it can make models overconfident in their predictions. Notably, XLMRoBERTa maintains better precision at the overall level (+1.80 over baseline XLMRoBERTa and +1.86 over XLMRoBERTa BT), indicating it is more robust to any noise introduced by RDA.

Backtranslation (BT) leads to a more pronounced shift in model behaviour, particularly for recall, where substantial increases can be seen for mBERT on LOC-type NEs (+4.35) and gaBERT for PER-type NEs (+1.83). However, precision consistently drops across all models, particularly for mBERT (-4.35) and XLMRoBERTa on LOC-type NEs (69.74, the lowest recall for LOC of all models and a decrease of 1.97), where the models are showcasing more false positive predictions for this entity type. This highlights the fundamental trade-off with BT: it improves recall at the cost of precision, leading to more false positives. Overall F1 scores decrease due to the precision loss, although F1 scores improve for the PER tag across all models due to simultaneous increases in both recall and precision (see Appendix C).

BT on MW-NE Test Set: The inclusion of BT improves the accuracy of predictions for PER made by mBERT (76.62 vs 80.54). While precision doesn't universally improve, it does increase for PER in all models (+1.43, +3.92, and +3.35 for gaBERT BT, mBERT BT, and XLMRoBERTa BT respectively) and for ORG in XLMRoBERTa (+2.27), confirming that BT can be beneficial for specific entity types such as PER, especially where recall is prioritised, perhaps due to an increase in how often the label is seen during training.

7.4 Addition of CRF Layers

The addition of a CRF and Bi-LSTM CRF to gaBERT and mBERT yields varying improvements across both test sets. Although gaBERT generally performs well, the introduction of both a CRF or Bi-LSTM CRF leads to improvements at the overall level (increase in overall precision of 5.33 with the addition of a CRF and 2.13 with the addition of a Bi-LSTM CRF). For gaBERT, the CRF enhances precision and recall, particularly for LOC-type and PER-type NEs (see Appendix C).

CRF on MW-NE Test Set: The addition of a Bi-LSTM layer increases the F1 score for each model when compared to their baselines (increase of 2.53 for gaBERT and 1.98 for mBERT, with the overall F1 for gaBERT Bi-LSTM being the highest achieved of all models tested on this set (77.36)). While the addition of a CRF to mBERT does not yield performance improvements, a Bi-LSTM improves over the mBERT baseline overall and achieves a higher precision for LOC on the mixed-length test set (see Appenidx D).

CRF with Data Augmentation: When CRF and BiLSTM-CRF layers are added to models with RDA or BT, the impacts are more nuanced. For RDA, the CRF layer enhances recall, with improvements generally seen across all entity types with only a few exceptions (mBERT RDA Bi-LSTM recall on LOC-type NEs and precision on ORG-type NEs, gaBERT RDA CRF recall on ORG-type NEs, and gaBERT RDA Bi-LSTM CRF recall and F1 score on ORG-type NEs and precision on PER-type NEs). Similarly with mBERT RDA, the addition of a CRF yielded enhancements across all metrics, except recall on LOC-type NEs. The addition of a Bi-LSTM improves recall scores for both mBERT and gaBERT RDA models when calculated across all NE types (see Appendix C). Recall improvements are seen across the models with RDA at the overall level with the introduction of a Bi-LSTM CRF. Introducing CRFs helps to mitigate some of the precision loss associated with RDA and BT. Overall, the addition of CRFs generally enhances F1 scores across the models and particularly enhances performance on PER entities. The CRFs lead to a more balanced performance across both test sets and model variants. The best performing model for the mixed-length test set was gaBERT RDA CRF achieving a F1 score of 76.52. This model also performed well on the MW-NE test set, however it was outperformed by gaBERT Bi-LSTM CRF (75.09 vs 77.36).

On inspection of the results from gaBERT RDA, gaBERT RDA CRF, and gaBERT RDA Bi-LSTM CRF on the mixed-length test set, it is clear that the addition of a CRF outperforms the others by being more precise and accurate when predicting entities. The Bi-LSTM CRF architecture shows a similar performance, although it tends to produce more

false positives. gaBERT RDA faces challenges in both over-predicting and missing entities compared to its CRF variants. Many of the errors made by gaBERT RDA are due to diverging from the integrity of the IOB2 tagging scheme. For almost all of these errors, both CRF variants did not make this mistake, as they were more consistent in maintaining the correct tagging structure, ensuring proper transitions between the tags (e.g. I-tagged tokens following B-tagged tokens).

7.5 MW-NEs

On further analysis, the majority of errors made on the mixed-length test set were due to incorrect predictions and divergence from the IOB2 tagging scheme. Investigation of the results show the majority of errors were made on MW-NEs with fewer words i.e. 2 words long. This is not surprising as the test set predominantly contains NEs of less than 3 words long (38% single-word NEs and 30% two-word NEs). As stated above; models have difficulty in maintaining accurate transitions between IOB2 tags, where entities are not always properly marked as part of a continuous sequence, titles and honorifics provide challenges for the PER-type NEs, and team ORG entities that are named for the location they are based in are predicted as the latter type.

The best performing model on the MW-NE test set is the gaBERT Bi-LSTM CRF with a F1 of 77.36 whereas the highest performing model for mixed-length NEs is gaBERT RDA CRF. It is interesting that a different model performed better on the MW-NE test set given that MW-NEs make up the majority of the entities in training. While this analysis did not reveal any entity-specific patterns for MW-NEs, it is hoped that on expansion of the datasets further insights can be gleaned on how MW-NEs are handled by these models.

8 Ethical Considerations

In the current climate of large language models, and massive data resources, the importance of data sovereignty and proper usage cannot be overlooked. The IUDT data (Lynn, 2022) was in part selected for the construction of the gold-standard NER-annotated corpus as it is under a CC BY-SA 3.0 license and comprises publicly accessible textual data sourced from the New Corpus of Ireland-Irish (NCII) (Kilgarriff et al., 2006), encompassing content from various sources such as websites, books, news articles, and other media. Additionally, it includes supplementary publicly available data available under the Open Data directive (European Parliament and Council of the European Union, 2019). The Dáil proceedings used are also publicly available under this directive (European Parliament and Council of the European Union, 2019). The lower energy demands of smaller BERT models is an argument for their continued usage in such experiments, particularly for exploratory studies such as this one. Insights from this work can be applied in future studies employing larger energy- and resource-hungry models.

9 Future Directions

Several avenues for advancing the scope and efficacy of NER for Irish present themselves after this research work. Firstly, acquiring more annotated data remains paramount given the scarcity of labelled corpora for low-resource languages such as Irish. Expanding the dataset used in this task could significantly bolster the performance of the models. A promising approach to this is self-training (Zhou et al., 2023) with Irish Wikipedia (Vicipéid). This semi-supervised approach would mitigate the labour-intensive manual annotation employed in this research (Zhou et al., 2023). Also, improving the data augmentation approach could be a focal point for future enhancement. Advanced techniques such as sentence-level resampling (Wang and Wang, 2022) could provide substantial benefits. This approach involves modifying existing sentences to create new training examples, which leads to different syntactic and semantic variations to improve the model's accuracy and generalisation capabilities (Wang and Wang, 2022). Additionally, a hybrid approach leveraging existing parsers should be considered (Lynn, 2016), using the parser to identify NEs in the text while BERT-based classifiers could be trained to predict the NE label. Newer models such as the UCCIX (Tran et al., 2024) monolingual Irish model are recently available for future experiments, and increases in performance can be weighed against the energy costs of training larger models. As mentioned in the European Language Equality Project's Report on the Irish Language (Lynn, 2023), the Gaois database of Irish-language surnames (Gaois, 2020) and the national placenames and biographies database (Gaois, 2008) could also be leveraged to build a NER tool.

[5]https://ga.wikipedia.org/wiki/Pr%C3%ADomhleathanach

Futhermore, the task of NER can be integrated into a larger study on machine translation capabilities of handling these challenging constructions (Ugawa et al., 2018). Additionally, a CRF model could be used to provide a baseline for comparison with the models utilised in this work. Finally, there is scope to combine this work with ongoing research in Irish MWE processing, and in particular research on noun compounds, as both constructions could be treated simultaneously.

10 Conclusion

We presented several architectures and training data setups for the NER for Irish task as well as a gold-standard NER-annotated corpus. We proposed and evaluated multiple NER models for Irish. Of these, the best-performing model is gaBERT with the implementation of random data augmentation and a CRF layer. Despite the limited amount of data available for fine-tuning, this model has demonstrated remarkable performance improvements compared to alternative architectures, including baseline gaBERT and the multilingual models mBERT and XLMRoberta on the task of Irish NER. With an F1 score of 76.52, the gaBERT RDA CRF model demonstrates robustness and accuracy in identifying both single- and MW-NEs in Irish text and generalises well to unseen data. Our findings further highlight several noteworthy observations. The incorporation of a CRF or BiLSTM CRF layer yielded notable performance improvements. Additionally, data augmentation strategies showed promising results at for different entity types and for general NE prediction. We hope that the gold-standard NER-annotated corpus for Irish can provide a benchmark for future research and valuable training data. Ultimately, our study sets the stage for continued progress in Irish NLP, offering a vital step forward in supporting the Irish language in the digital age.

Limitations

Due to time and resource constraints we were unable to implement the addition of CRFs to XLM-RoBERTa and only a small set of sentences were used for backtranslation. The data collection used text from transcriptions of Dáil proceedings, which represent a limited context of Irish language, e.g. formal language in a legal domain. Furthermore, as the text was filtered from a larger body of bilingual text, the Irish used likely represents a minority of speakers from the Dáil, which further narrows the scope of text type. Back-translation and random data augmentation provide a means for synthetically inflating training data in low-resource scenarios to improve model performance; however, it should be noted that these techniques can introduce new errors into the data. Future work includes exploring the impact of these data-augmentation techniques on the quality of the text (e.g. semantic coherence of the sentence, co-reference, etc.).

Acknowledgements

This research and future work on text processing for Irish is sustained through funding from the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, Research Ireland, and Údarás na Gaeltachta. The authors would also like to thank the reviewers for their detailed feedback, many of whose comments were incorporated into the final paper.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 89–93, Florence, Italy. Association for Computational Linguistics.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. 2022. gaBERT — an Irish language model. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4774–4788, Marseille, France. European Language Resources Association.

K. Chan, K.A Las Alas, C. Orcena, D.J. Velasco, Q.J. San Juan, and C. Cheng. 2023. Practical approaches for low-resource named entity recognition of Filipino telecommunications domain. In Pacific Asia Conference on Language, Information and Computation.

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.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. Preprint, arXiv:2010.11683.

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.

European Parliament and Council of the European Union. 2019. Open data directive - data.gov.ie. Data.gov.ie https://data.gov.ie/pages/open-data-directive.

Gaois. 2008. Placenames database of Ireland. logainm.ie.

Gaois. 2020. Database of Irish-language surnames. Gaois research group https://www.gaois.ie/en/surnames/info.

Yao Ge, Yuting Guo, Yuan–Chi Yang, Mohammed Ali Al–Garadi, and Abeed Sarker. 2022. A comparison of few-shot and traditional named entity recognition models for medical text. In 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), pages 84–89.

Houses of the Oireachtas. 2024. Dáil transcripts.

Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and Philip S. Yu. 2023. Entity-to-text based data augmentation for various named entity recognition tasks. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9072–9087, Toronto, Canada. Association for Computational Linguistics.

A. Kilgarriff, M. Rundell, and E. Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: Building the new corpus for Ireland. Language Resources and Evaluation, 40:127–152.

Label Studio. 2020. Label studio – open source data labeling. https://labelstud.io/.

Yafei Liu, Siqi Wei, Haijun Huang, Qin Lai, Mengshan Li, and Lixin Guan. 2023. Naming entity recognition of citrus pests and diseases based on the bert-bilstm-crf model. Expert Systems with Applications, 234.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In International Conference on Learning Representations, New Orleans, Louisiana, United States.

Teresa Lynn. 2016. Irish dependency treebanking and parsing. Phd thesis, Dublin City University, Dublin, Ireland. Available at https://doras.dcu.ie/21014/.

Teresa Lynn. 2022. Universaldependencies/ud irish-idt. https://github.com/UniversalDependencies/UD_Irish-IDT.

Teresa Lynn. 2023. Language report Irish. In European Language Equality Cognitive Technologies. Springer, Cham, Switzerland, pages 163–166. Accessed 10/12/2024.

Teresa Lynn, Jennifer Foster, Sarah McGuinness, Abigail Walsh, Jason Phelan, and Kevin Scannell. 2023. Universal Dependencies Irish Dependency Treebank (v2.12).

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: Fallacies, challenges and opportunities. Computer Standards I& Interfaces, 35(5):1–13. https://api.semanticscholar.org/CorpusID:2635684.

Sarah McGuinness, Jason Phelan, Abigail Walsh, and Teresa Lynn. 2020. Annotating MWEs in the Irish UD treebank. In Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 126–139, Barcelona, Spain (Online). Association for Computational Linguistics.

Christopher Moseley. 2012. The UNESCO Atlas of the World's Languages in Danger: Context and Process. World Oral Literature Project, University of Cambridge Museum of Archaeology and Anthropology.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

J. Nivre. 2015. Towards a universal grammar for natural language processing. Computational Linguistics and Intelligent Text Processing, 9041:3–16.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1659–1666.

Ailbhe Ní Chasaide, Neasa Ní Chiarán, Elaine Uí Dhonnchadha, Teresa Lynn, and John Judge. 2022. Digital plan for the Irish language speech and language technologies 2023-2027. https://assets.gov.ie/241755/e82c256a-6f47-4ddb-8ce6-ff81df208bb1.pdf.

Atul Kr. Ojha, Chao-Hong Liu, Katharina Kann, John Ortega, Sheetal Shatam, and Theodorus Fransen. 2021. Findings of the LoResMT 2021 shared task on COVID and sign language for low-resource languages. In Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021), pages 114–123, Virtual. Association for Machine Translation in the Americas.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. Preprint, arXiv:1802.05365.

Lutz Prechelt. 2012. Early stopping — but when? In Neural Networks: Tricks of the Trade: Second Edition, pages 53–67. Springer, Berlin, Heidelberg, Berlin, Heidelberg.

Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Thenmozi Durairaj, Anbukkarasi Sampath, Kingston Pal Thamburaj, and Bharathi Raja Chakravarthi. 2021. Attentive fine-tuning of transformers for translation of low-resourced languages @LoResMT 2021. In Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021), pages 134–143, Virtual. Association for Machine Translation in the Americas.

Jonathan Raiman and John Miller. 2017. Globally normalized reader. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1059–1069, Copenhagen, Denmark. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 107–118, online. Association for Computational Linguistics.

C. Sbaty, I. Omar, F. Wasfalla, M. Islam, and S. Abdennadher. 2021. Data augmentation techniques on Arabic data for named entity recognition. Procedia Computer Science, 189:292–299.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Portuguese named entity recognition using bert-crf. Preprint, arXiv:1909.10649.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024. Uccix: Irish-excellence large language model. Preprint, arXiv:2405.13010.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abigail Walsh. 2023. The Automatic Processing of Multiword Expressions in Irish. Phd thesis, Dublin City University.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2022. A BERT's eye view: Identification of Irish multiword expressions using pre-trained language models. In Proceedings of the 18th Workshop on Multiword Expressions @LREC2022, pages 89–99, Marseille, France. European Language Resources Association.

Xiaochen Wang and Yue Wang. 2022. Sentence-level resampling for named entity recognition. In North American Chapter of the Association for Computational Linguistics.

Nan. Xu, W. Mao, P. Wei, and D. Zeng. 2021. MDA: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks. IEEE Intelligent Systems, 36(6):3–12.

Usama Yaseen and Stefan Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), pages 352–358, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. 2023. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning.

Appendix

A Gitlab

The scripts and annotation guidelines used in this research can be found in the GitHub repository: Named Entity Recognition for the Irish Language Gitlab Repo.

B Annotation Guidelines

- **Tags Discussed:** Person (PER), Location (LOC) and Organisation (ORG)
- **Tagging Scheme:** IOB2

B.1 Person

- A person's name or family name (real or fictional even if spelled incorrectly) e.g. Micheál (B- PER) Martin (I-PER)
- Gods (when having a single reference and capitalised)

  (1) Buíochas le Dia (B-PER)

  Thanks to God (B-PER)

- A person's initials e.g. M.M. (B-PER)

- Do not tag titles as a name or part of a name

  (2) Chonaic (O) mé (O) Dr.(O) O'Sullivan (B-PER) inné (O), An(O) Taoiseach (O) Leo (B-PER) Varadkar (I-PER), An (O) tUasal (O) Mac (B-PER) Gabhann (I-PER), Mary (B-PER) Lou (I-PER) McDonald(I-PER) T.D. (O)

  I saw Dr. (O) O'Sullivan (B-PER) yesterday (O), the (O) Prime (O) Minister (O) Leo (B-PER) Varadkar (I-PER), Mr. (O) Mac (B-PER) Gabhann (I-PER), Mary (B-PER) Lou (I-PER) McDonald (I-PER) T.D. (O)

- Do tag if used as a name mention

  (3) Dúirt (O) An (O) Taoiseach (B-PER) go (O) bhfuil (O) . . .

  The (O) Prime (B-PER) Minister (I-PER) said that ...

B.2 Location

- Geographical places, facilities or buildings e.g. countries, cities, towns, airports, hotels, roads etc.
- When two locations are consecutive, tag separately

  (4) Tá (O) mé (O) i (O) mo (O) chónaí (O) i (O) Sord (B-LOC), Baile (B-LOC) Átha (I-LOC) Cliath (I-LOC)

  I (O) live (O) in (O) Swords (B-LOC), Dublin (B-LOC)

- Tag whole postal addresses as one

  (5) 27 (B-LOC) Bóthar (I-LOC) na (I-LOC) Foraoise (I-LOC), Caisleán (I-LOC) an (I-LOC) Chomair (I-LOC), Cill (I-LOC) Chainnigh (I-LOC)

  27 (B-LOC) Forest (I-LOC) Road (I-LOC), Castle (I-LOC) Comer(I-LOC), Kilkenny (I-LOC)

B.3 Organisation

- Named collections of people (organisations, institutions, firms, political parties, unions, groups)

  (6) Is (O) é (O) Simon (B-PER) Harris (I-PER) ceannaire (O) Fhine (B-ORG) Gael (I-ORG)

  Simon (B-PER) Harris (I-PER) is (O) the (O) leader (O) of (O) Fine (B-ORG) Gael (I-ORG)

(7) Fáiltím (O) roimh (O) an (O) nuacht (O) is (O) deireanaí (O) a (O) chuala (O) muid (O) ar (O) maidin (O) ó (O) Citylink (B-ORG) go (O) mbeidh (O)...

I (O) welcome (O) the (O) latest (O) news (O) that (O) we (O) heard (O) this (O) morning (O) from (O) Citylink (B-ORG) that (O) there (O) will (O) be (O)...

- Names of places when they act as administrative units or sports teams

  (8) Chaill (B-ORG) Baile (I-ORG) Átha (I-ORG) Cliath (I-ORG) in (O) aghaidh (O) Gaillimh (B-ORG) an (O) seachtain (O) seo (O) caite (O)

  Dublin (B-ORG) lost (O) against (O) Galway (B-ORG) last (O) week (O)

- Include corporate designators like Co. and Ltd. as part of the name

  (9) Is (O) gnólacht (O) dlí (O) iad (O) Johnson (B-ORG) & Co. (I-ORG)

  Johnson (B-ORG) & Co. (I-ORG) is (O) a (O) law (O) firm

- Only tag brands if referring to the organisation itself, not as a brand label

  (10) Tá (O) bróga (O) nua (O) eisithe (O) ag (O) Nike (B-ORG).

And not in the following:
Ghortaigh (O) mé (O) mo (O) chosa (O) mar (O) chaith (O) mé (O) Nike (O)

Nike (B-ORG) has (O) released (O) new (O) shoes (O).
And not in the following:
I (O) hurt (O) my (O) foot (O) because (O) I (O) wore (O) Nike (O)

Other points to note: Inclusion of non-name tokens should be tagged

(11) Nua-Eabhrac-bhunaithe (B-LOC)

New-York-based (B-LOC)

C Results on Mixed-Length Test Set

D Results on MW-NE Test Set

Table 3: Results on Mixed-Length Test Set

Model	Overall				LOC			ORG			PER		
	Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
gaBERT	94.84	71.74	**77.40**	74.46	76.61	77.51	77.06	61.11	72.79	66.44	76.88	81.10	78.93
gaBERT CRF	95.25	77.07	75.27	76.16	80.37	77.51	78.92	65.00	66.91	65.94	**84.52**	79.88	82.13
gaBERT Bi-LSTM CRF	94.82	73.87	76.55	75.18	77.50	73.37	75.38	67.97	76.47	**71.97**	75.72	79.88	77.74
gaBERT RDA	94.72	71.11	75.05	73.03	71.51	75.74	73.56	61.39	71.32	65.99	80.38	77.44	78.88
gaBERT RDA CRF	**95.44**	**78.05**	75.05	**76.52**	**81.60**	78.70	**80.12**	**69.70**	67.65	68.66	81.41	77.44	79.38
gaBERT RDA Bi-LSTM CRF	95.02	74.59	76.97	75.76	79.88	77.51	78.68	63.89	67.65	65.71	78.41	84.15	81.18
gaBERT BT	94.67	69.19	76.12	72.49	67.20	73.96	70.42	61.15	70.59	65.53	78.61	82.93	80.71
gaBERT BT CRF	95.15	74.59	76.97	75.76	75.00	76.33	75.66	63.01	67.65	65.25	84.34	**85.37**	**84.85**
gaBERT BT Bi-LSTM CRF	94.59	72.08	73.77	72.92	73.78	71.6	72.67	60	63.97	61.92	80.7	84.15	82.39
mBERT	93.28	70.86	73.47	72.14	74.62	75.19	74.90	67.83	72.93	70.29	68.40	71.25	69.80
mBERT CRF	93.02	69.05	72.13	70.56	76.96	72.63	74.74	61.69	71.43	66.20	66.54	72.08	69.20
mBERT BiLSTM CRF	93.33	73.22	72.24	72.73	79.13	74.68	76.84	65.97	71.43	68.59	72.81	69.17	70.94
mBERT RDA	92.22	64.10	69.68	66.77	66.24	79.80	72.39	58.90	64.66	61.65	66.51	58.75	62.39
mBERT RDA CRF	93.57	72.28	74.14	73.20	77.27	78.26	77.76	62.34	72.18	66.90	77.31	69.58	73.25
mBERT RDA Bi-LSTM CRF	92.90	69.42	74.24	71.70	77.28	75.70	76.49	58.61	**79.32**	67.41	73.49	65.83	69.45
mBERT BT	93.01	68.81	74.02	71.32	74.58	79.54	76.98	60.45	70.68	65.16	69.62	68.75	69.18
mBERT BT CRF	93.41	71.82	75.59	73.66	76.27	**80.56**	78.36	62.07	74.44	67.69	77.83	68.75	73.01
mBERT BT Bi-LSTM CRF	92.60	68.72	69.79	69.25	75.60	72.89	74.22	57.70	71.80	63.99	73.89	62.50	67.72
XLMRoberta	93.44	69.39	75.37	72.26	74.66	71.71	73.15	57.91	77.78	66.39	76.37	78.02	77.19
XLMRoberta RDA	93.63	69.18	74.02	71.52	70.13	71.05	70.59	60.16	74.40	66.52	77.92	77.59	77.75
XLMRoberta BT	93.87	67.59	72.41	69.92	69.74	69.74	69.74	58.04	71.5	64.07	75.11	76.72	75.91

Table 4: Results on MW-NE Test Set

Model	Overall				LOC			ORG			PER		
	Accuracy	P	R	F1	P	R	F1	P	R	F1	P	R	F1
gaBERT	95.05	70.39	79.85	74.83	71.00	83.53	76.76	57.58	74.03	64.77	81.90	81.13	81.52
gaBERT CRF	95.49	73.33	77.99	75.59	74.44	78.82	76.57	57.61	68.83	62.72	86.41	83.96	85.17
gaBERT Bi-LSTM CRF	94.99	74.06	**80.97**	**77.36**	74.42	75.29	74.85	62.63	80.52	70.45	84.26	**85.85**	85.05
gaBERT RDA	94.86	67.21	76.49	71.55	64.15	80.00	71.20	54.08	68.83	60.57	83.17	79.25	81.16
gaBERT RDA CRF	**95.65**	**75.09**	77.61	76.33	75.82	81.18	78.41	61.9	67.53	64.6	85.29	82.08	83.65
gaBERT RDA Bi-LSTM	95.08	71.77	78.73	75.09	72.04	78.82	75.28	57.61	68.83	62.72	83.49	85.85	84.65
gaBERT BT	94.77	65.92	77.24	71.13	58.88	74.12	65.62	54.55	70.13	61.36	83.33	84.91	84.11
gaBERT BT CRF	95.49	72.79	79.85	76.16	72.92	82.35	77.35	58.24	68.83	63.10	85.05	85.85	85.45
gaBERT BT Bi-LSTM CRF	94.89	70.73	75.75	73.15	67.74	74.12	70.79	55.56	64.94	59.88	86.54	84.91	**85.71**
mBERT	94.08	71.54	77.64	74.46	**76.72**	81.01	78.80	62.30	79.87	70.00	76.62	71.95	74.21
mBERT CRF	93.76	68.85	76.83	72.62	73.96	79.33	76.55	58.38	77.18	66.47	75.62	73.78	74.69
mBERT Bi-LSTM CRF	94.49	73.63	79.47	76.44	76.26	84.36	**80.11**	**64.84**	79.19	**71.30**	80.79	74.39	77.46
mBERT RDA	93.29	62.90	71.34	66.86	60.76	80.45	69.23	55.38	72.48	62.79	78.57	60.37	68.28
mBERT RDA CRF	94.35	72.81	77.85	75.25	71.77	83.80	77.32	62.92	75.17	68.50	87.05	73.78	79.87
mBERT RDA Bi-LSTM CRF	93.71	68.95	77.64	73.04	72.86	81.01	76.72	55.07	**83.89**	66.49	**87.50**	68.29	76.71
mBERT BT	93.96	69.31	78.05	73.42	71.77	83.80	77.32	58.16	76.51	66.09	80.54	73.17	76.68
mBERT BT CRF	93.98	70.83	78.46	74.45	73.56	**85.47**	79.07	59.79	77.85	67.64	81.82	71.34	76.22
mBERT BT Bi-LSTM CRF	93.59	67.97	74.19	70.94	72.45	79.33	75.73	55.98	78.52	65.36	80.30	64.63	71.62
XLMRoberta	93.65	67.86	76.36	71.86	69.43	76.22	72.67	55.15	77.78	64.54	79.87	75.46	77.60
XLMRoberta+data_aug	94.23	69.66	77.07	73.18	65.48	76.92	70.74	60.67	77.78	68.16	83.33	76.69	79.87
XLMRoberta+backt	93.65	67.8	75.65	71.51	63.69	74.83	68.81	57.42	76.07	65.44	83.22	76.07	79.49