Are Data Augmentation Methods in Named Entity Recognition Applicable for Uncertainty Estimation?

Wataru Hashimoto, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology
{hashimoto.wataru.hq3, kamigaito.h, taro}@is.naist.jp

Abstract

This work investigates the impact of data augmentation on confidence calibration and uncertainty estimation in Named Entity Recognition (NER) tasks. For the future advance of NER in safety-critical fields like healthcare and finance, it is essential to achieve accurate predictions with calibrated confidence when applying Deep Neural Networks (DNNs), including Pre-trained Language Models (PLMs), as a real-world application. However, DNNs are prone to miscalibration, which limits their applicability. Moreover, existing methods for calibration and uncertainty estimation are computational expensive. Our investigation in NER found that data augmentation improves calibration and uncertainty in cross-genre and cross-lingual setting, especially in-domain setting. Furthermore, we showed that the calibration for NER tends to be more effective when the perplexity of the sentences generated by data augmentation is lower, and that increasing the size of the augmentation further improves calibration and uncertainty.

1 Introduction

Named Entity Recognition (NER) is a one of the fundamental tasks in Natural Language Processing (NLP) to find mentions of named entities and classify them into predefined categories. The predicted information by NER is essential for downstream tasks like event detection (Vavliakis et al., 2013), information retrieval (Cowan et al., 2015), and masking of personal user information (Kodandaram et al., 2021). Due to the demand, NER is the underlying technology for information extraction from text and documents.

Based on the recent advances in Deep Neural Networks (DNNs), NER's performance is also improved like other NLP fields. In recent years, Pre-trained Language Models (PLMs) based architectures, such as BERT (Devlin et al., 2019) and DeBERTa (He et al., 2021), have been strong baselines in many NLP tasks, including NER.

In general, however, DNNs are prone to miscalibration (Guo et al., 2017), including PLMs (Desai and Durrett, 2020); *calibration* means the predicted confidence of the model aligns with the accuracy.[1] The problem causes DNNs to make incorrect predictions with high confidence, which limits the applicability of DNNs on the number of domains where the cost of errors is high, e.g., healthcare and finance. Therefore, DNNs need to provide high prediction performance with appropriately calibrated confidence at the same time.

Confidence calibration and uncertainty estimation methods are ways to solve the miscalibration of DNNs, and have been applied in NLP tasks such as text classification (Xiao and Wang, 2019), structured prediction (Jiang et al., 2022; Reich et al., 2020), question answering (Si et al., 2022), and machine translation (Malinin and Gales, 2021). However, many methods for confidence calibration and uncertainty estimation, typically Monte-Carlo Dropout (MC Dropout) (Gal and Ghahramani, 2016), are computationally expensive due to multiple stochastic inferences, making them difficult for real-world application.

Data augmentation has also been applied for NER (Dai and Adel, 2020; Zhou et al., 2022), though, it was focusing on the generalization ability on low-resource data. In computer vision (CV) areas, data augmentation makes the model more robust to the input and leads to confidence calibrations (Wen et al., 2021; Liu et al., 2023), in which the same labels are trained on different representations of the input than the original data. Based on the findings of these previous studies, there is a possibility that data augmentation in NER can improve confidence calibration without increasing inference

[1] For example, a predicted confidence of 0.70 from a *perfectly calibrated* network should be 70% accuracy for that inputs.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18852–18867

time, in contrast to the conventional confidence calibration and uncertainty estimation methods.

In this study, we conducted comprehensive experiments to analyze the impact of data augmentation methods for NER (Dai and Adel, 2020; Zhou et al., 2022) on the confidence calibration and uncertainty in the cross-genre and cross-lingual settings on OntoNotes 5.0 (Pradhan et al., 2013) and MultiCoNER (Malmasi et al., 2022), respectively.

Our experiments yield several findings. First, some data augmentation methods in NER lead to improved confidence calibration and uncertainty estimation, especially in-domain. In particular, entity-prediction-based data augmentation (Zhou et al., 2022) and entity replacement from the same entity type (Dai and Adel, 2020) show good performance. On the other hand, common confidence calibration methods, MC Dropout or TS (Guo et al., 2017) have worse confidence calibration and uncertainty estimation performance than the data augmentation methods in NER, even though the data augmentation methods do not aim to improve confidence calibration and uncertainty estimation. Moreover, increasing the augmentation size improves performance in confidence calibration and uncertainty estimation. The improvement tends to be better the lower the perplexity of the sentences generated by the data augmentation. Our code is available on https://github.com/wataruhashimoto52/ner_da_uncertainty.

2 Related Work

Named Entity Recognition In the last decade, NER using DNNs has been widely successful; Lample et al. (2016) reported a sequence-labeling model combining bi-directional LSTM with CRF (BiLSTM-CRF). Akbik et al. (2018) proposed contextualized character-level word embeddings combined with BiLSTM-CRF. In recent years, NER models based on PLMs, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021), have achieved state-of-the-art performance.

Uncertainty Estimation In general, DNNs are prone to miscalibration and overconfidence (Guo et al., 2017) especially without pretraining (Desai and Durrett, 2020; Ulmer et al., 2022). One way to estimate uncertainty is to run multiple stochastic predictions. Deep Ensemble (Lakshminarayanan et al., 2017) trains multiple DNN models and integrates their multiple stochastic predictions to make a final prediction. MC Dropout (Gal and Ghahramani, 2016) applies Dropout (Srivastava et al., 2014) regularization at both training and inference time, and by taking multiple samples of the network outputs during inference. These are known to perform calibration well in many cases (Ovadia et al., 2019; Immer et al., 2021), but their practical use is hampered by the fact that they make multiple probabilistic predictions. A relatively lightweight calibration method is the post-hoc approach. For example, temperature scaling (Guo et al., 2017) performs calibration via dividing logits by a constant, which is a simple and lightweight baseline.

Data Augmentation Data augmentation methods are widely used in machine learning, CV, and NLP areas. More recent attention has focused on the provision of data augmentation methods to improve calibration and uncertainty. Test-time augmentation (TTA) (Ashukha et al., 2020) generates multiple samples during inference and integrates the predictions to estimate the prediction uncertainty. MixUp (Zhang et al., 2018) uses linear interpolation between two samples to augment a new sample with soft labels, which has been investigated for situations where it is effective for calibration (Zhang et al., 2022).

In NLP tasks, the impact of data augmentation on calibration in text classification has been investigated in recent study (Kim et al., 2023), but only for In-domain (ID) and not for NER. Furthermore, it has been found that predictive performance is driven by data augmentation in NER (Dai and Adel, 2020; Chen et al., 2020; Zhou et al., 2022; Chen et al., 2022; Hu et al., 2023), but these studies have focused only on the predictive performance of NER and have not evaluated for calibration and uncertainty. This is the first study to comprehensively investigate the impact of data augmentation on calibration and uncertainty in NER, both in ID and OOD (Out-of-domain) settings.

3 Methods

In this section, we describe the popular baseline methods for confidence calibration and data augmentation methods for NER. Details about existing calibration methods are described in Appendix B.

3.1 Existing Calibration Methods

Baseline Baseline uses the maximum probability from the softmax layer.

Temperature Scaling (TS) TS (Guo et al., 2017) is a post-processing technique for calibrating the confidence scores outputted by a neural network. It involves scaling the logits (i.e., the outputs of the final layer before the softmax) by a temperature parameter T before applying the softmax function to obtain the calibrated probabilities.

Label Smoothing (LS) LS (Miller et al., 1996; Pereyra et al., 2017) is prevalent regularization technique in machine learning, introduces a controlled level of uncertainty into the training process by modifying the cross-entropy loss.

Monte-Carlo Dropout (MC Dropout) MC Dropout is a regularization technique that can be used for uncertainty estimation in neural networks, which requires multiple stochastic inferences (Gal and Ghahramani, 2016). We perform 20 stochastic inferences and output their average.

3.2 Data Augmentation Methods for NER

We investigate data augmentation methods in NER (Dai and Adel, 2020; Zhou et al., 2022) for confidence calibration and uncertainty estimation.

Label-wise Token Replacement (LwTR) LwTR uses binomial distribution to determine whether a token is replaced. The chosen token is randomly replaced with another token with the same label based on label-wise token distribution on training data. Thus, LwTR keeps the original label sequence.

Mention Replacement (MR) Unlike LwTR, MR replaces an entity with another entity with the same label instead of a token. Other parts are the same as LwTR. Since entities can have multiple tokens, MR does not keep the original label sequence.

Synonym Replacement (SR) SR is similar to LwTR except that SR replaces a token with its synonym in WordNet (Miller, 1995). Since the synonym can have multiple tokens, SR does not keep the original label sequence.

Masked Entity Language Modeling (MELM) MELM (Zhou et al., 2022) performs data augmentation using a language model that predicts contextually appropriate entities for sentences in which entity parts are masked by entity markers.

4 Evaluation Metrics

We use Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Area Under Precision-Recall Curve (AUPRC) to evaluate confidence calibration and uncertainty estimation.

4.1 Expected Calibration Error (ECE)

ECE (Naeini et al., 2015) measures the difference between the accuracy and confidence of a model. Specifically, it calculates the difference between the average confidence and the actual accuracy of the model on different confidence levels. Formally, ECE is defined as:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{D}_b|}{n} \left|\mathrm{acc}(\mathcal{D}_b) - \mathrm{conf}(\mathcal{D}_b)\right|$$

where B is the number of confidence interval bins, $\mathcal{D}_b$ is the set of examples whose predicted confidence scores fall in the b-th interval, n is the total number of examples, $\mathrm{acc}(\mathcal{D}_b)$ is the accuracy of the model on the examples in $\mathcal{D}_b$, and $\mathrm{conf}(\mathcal{D}_b)$ is the average confidence of the model on the examples in $\mathcal{D}_b$.

4.2 Maximum Calibration Error (MCE)

MCE (Naeini et al., 2015) is the maximum difference between the accuracy and the confidence of the model on different confidence levels. Formally, MCE is defined as:

$$\mathrm{MCE} = \max_{b=1}^{B} \left|\mathrm{acc}(\mathcal{D}_b) - \mathrm{conf}(\mathcal{D}_b)\right|,$$

MCE takes the maximum calibration error in each bin, not the expectation; a smaller MCE means that the model's predictions are less likely to be far off in a given confidence region.

4.3 Area Under the Precision-Recall Curve (AUPRC)

AUPRC is the summary statistic the relationship between precision and recall at different thresholds. The higher the value, the higher the overall precision at a given threshold.

5 Experimental Settings

5.1 Datasets

We conducted experiments on two different NER datasets to evaluate the performance of confidence calibration methods in different settings. For the cross-genre evaluation, we used the OntoNotes 5.0

Dataset & Domain	N_{ent}	Train	Dev	Test
OntoNotes 5.0				
bc	18	11,866	2,117	2,211
bn	18	10,683	1,295	1,357
mz	18	6,911	642	780
nw	18	33,908	5,771	2,197
tc	18	11,162	1,634	1,366
wb	18	7,592	1,634	1,366
MultiCoNER				
English (EN)	6	15,300	800	10,000
German (DE)	6	-	-	10,000
Spanish (ES)	6	-	-	10,000
Hindi (HI)	6	-	-	10,000

Table 1: Dataset statistics. The table presents the number of entity types, and sequences for the train, development, and test parts of the datasets. For MultiCoNER, we randomly sampled and fixed 10,000 cases out of 200,000 test cases.

Dataset & Domain	LwTR	MR	SR	MELM (η, μ)
OntoNotes 5.0				
bc	0.3	0.7	0.3	(0.5, 0.5)
bn	0.4	0.8	0.2	(0.7, 0.3)
mz	0.7	0.4	0.5	(0.3, 0.3)
nw	0.7	0.5	0.7	(0.7, 0.7)
tc	0.4	0.4	0.1	(0.3, 0.3)
wb	0.7	0.7	0.8	(0.5, 0.7)
MultiCoNER				
English (EN)	0.2	0.8	0.4	(0.3, 0.3)

Table 2: Optimized hyperparameters in data augmentation methods in each source domain. We present the binomial distribution parameters for LwTR, SR and MR, and (η, μ) for MELM, respectively.

dataset (Pradhan et al., 2013), which consists of six different genres, broadcast conversation (bc), broadcast news (bn), magazine (mz), newswire (nw), telephone conversation (tc), and web data (wb). This dataset is commonly used for NER evaluation in a cross-domain setting (Chen et al., 2021).

For the cross-lingual evaluation, we used the MultiCoNER dataset, which is a large multilingual NER dataset from Wikipedia sentences, questions, and search queries (Malmasi et al., 2022). We selected English as the source language and English, German, Spanish, Hindi, and Bangla as the target languages. The details of the dataset statistics are provided in Table 1.

5.2 Training Details

In all experiments, we train out models on a single NVIDIA A100 GPU with 40GB of memory. We used MIT-licensed mDeBERTaV3 (microsoft/mdeberta-v3-base) (He et al., 2023) whose model size is 278M, as a multilingual transformer encoder from Hugging Face transformers (Wolf et al., 2020) pre-trained model checkpoints, and extracted entities via *sequence labeling*. Cross-entropy loss is minimized by AdamW (Loshchilov and Hutter, 2019) with a linear scheduler (Goyal et al., 2017). The batch size is 32, and gradient clipping is applied with maximum norm of 1. The initial learning rate was set to 1e-5. To avoid overfitting, we also applied early stopping with $patients = 5$.

For the temperature parameter in TS, we used Optuna (Akiba et al., 2019) to optimize the temperature parameter based on dev set loss with a search range of [0.001, 0.002, ..., 5.000] in 100 trials. In addition, we optimized the binomial distribution parameter to manipulate replacement intensity for data augmentation methods using the dev set by a grid search in the range of [0.1, 0.2, ..., 0.8]. In LS, we conducted a grid search in the range of [0.01, 0.05, 0.1, 0.2, 0.3] to optimize the smoothing parameter. In the case of MELM, mask rate η during fine tuning and mask parameter μ during generation are hyperparameters. We conducted a grid search for each hyperparameter in the range [0.3, 0.5, 0.7], as in Zhou et al. (2022). All hyperparameters in data augmentation are shown in Table 2. The implementations of LwTR, MR and SR are used several repos,[2] [3] while the implementation of MELM used the official repo.[4]

We perform each experiment 10 times using different random seeds, collect evaluation metric values, and report their average and standard deviation. For convenience, the reported values are multiplied by 100.

5.3 Evaluation Details

The NER model calibration is evaluated based on the "Event of Interests" concept introduced in the previous study (Kuleshov and Liang, 2015; Jagannatha and Yu, 2020). Since the full label space $|\mathcal{Y}|$ is large for structured prediction tasks such as NER, we focus instead on the event set $L(x)$, which is the set containing the events of interest $E \in L(x)$ obtained by processing the model output.

There are two main strategies for constructing $L(x)$: The first strategy is to construct $L(x)$ only from the events obtained by the MAP label se-

[2] https://github.com/boschresearch/data-augmentation-coling2020

[3] https://github.com/kajyuuen/daaja

[4] https://github.com/RandyZhouRan/MELM

Methods	bc		bn		mz		nw		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	18.87±0.73	23.58±1.01	11.50±0.75	16.14±1.97	15.75±0.94	20.93±0.97	11.74±0.27	16.15±0.77	31.17±1.56	33.81±1.67	28.86±1.51	34.38±1.82
TS	18.86±0.68	23.22±0.86	11.25±0.55	15.43±1.41	15.40±0.74	20.30±1.23	11.71±0.36	15.80±0.85	27.95±2.51	30.70±2.55	29.70±1.54	34.88±1.66
LS	19.29±1.04	24.11±1.57	11.42±0.52	15.31±1.24	15.59±0.85	20.91±1.30	12.05±0.20	16.83±0.36	26.46±1.36	28.89±1.42	29.34±2.25	34.86±2.22
MCDropout	18.69±0.71	23.54±1.31	11.38±0.71	15.73±1.60	15.89±0.29	21.15±0.54	11.83±0.55	16.56±1.41	29.01±2.50	31.94±2.81	28.41±1.45	33.88±1.77
LwTR (DA)	19.15±0.55	23.70±0.77	11.72±0.42	16.37±1.21	15.12±0.44	20.56±0.80	11.82±0.39	15.57±0.47	28.78±2.27	31.31±2.14	28.72±1.70	34.30±1.68
MR (DA)	19.13±0.95	23.17±1.10	11.59±0.34	15.89±0.92	14.66±1.05	19.63±1.37	11.50±0.33	15.62±0.74	28.65±3.20	31.23±3.18	27.08±1.40	32.39±1.57
SR (DA)	**18.16±0.63**	**21.99±0.91**†	11.38±0.44	15.44±0.96	15.29±0.96	20.11±1.14	11.71±0.25	16.31±0.57	27.30±4.37	29.85±4.54	29.72±0.91	34.74±1.05
MELM (DA)	18.59±0.60	22.67±0.95	**10.75±0.46**†	**14.11±0.69**†	**13.94±0.98**†	**18.50±1.22**†	**11.28±0.33**†	**15.43±0.98**	**25.71±1.73**	**28.19±1.87**	**26.58±1.48**†	**31.47±1.64**†

Table 3: Results of existing methods and data augmentation methods in OntoNotes 5.0 in ID setting. The best results are shown in bold. † indicates significantly improved than existing methods ($p < 0.05$) by using t-test.

quence prediction of the model; The second strategy is to construct $L(x)$ from all possible label sequences; The first strategy is easy to obtain events, but the coverage of events is low depending on the model's prediction. The second strategy provides a high coverage of events, but is computationally expensive to obtain events. Jagannatha and Yu (2020) is based on the first strategy, where the entities extracted by the NER model are calibrated on the basis of forecasters (e.g., gradient boosting decision trees (Friedman, 2000)), which are binary classifiers separate from the NER model. Since the training dataset for forecasters consists of entities extracted by the NER model, more entities are needed to improve the uncertainty performance of the forecasters. Therefore, for example, the top-k Viterbi decoding of the CRF is used to increase the entity coverage and the size of the forecaster's training dataset.

On the other hand, Jiang et al. (2022) is based on the second strategy, where it introduces a method to find the probability that a span has a specific entity type for datasets with short sequences, such as WikiAnn (Pan et al., 2017), with restricted token sequences and span lengths. However, this method is computationally difficult for datasets with longer token sequences and more complex label spaces, such as OntoNotes 5.0 and MultiCoNER, because the number of spans explodes. We therefore simplify the evaluation process by measuring the calibration of the entity span obtained from the NER model's MAP label sequence prediction of the model. Uncertainty performance is evaluated by taking the product of the probabilities of each token corresponding to an entity as the probability of one entity.

6 Results and Discussion

We present the performance of cross-genre and cross-lingual confidence calibration and uncertainty estimation as the main results. The cross-genre evaluations are quantified by learning on a training set in one genre and evaluating calibration and uncertainty on a test in another genre. Similarly, in the cross-lingual evaluations, we train the model in one language (in this research, we use English; `EN`) and evaluate the calibration and uncertainty on a test set in another language.

6.1 Cross-genre Evaluation

The results shown in Table 3 demonstrate ECE and MCE in OntoNotes 5.0 for NER in the ID setting, which the source domain and target domain are the same. The table results show that data augmentation methods consistently have better calibration performance than TS, LS, and MC Dropout, which have been considered to work for general classification problems, in the evaluation of calibration performance, in the ID setting. In particular, when the source genre is `tc`, MELM and other data augmentation methods show superior calibration performance, with up to 6.01 % improvement for ECE and 5.62 % improvement for MCE compared to Baseline. As shown in Table 1, the `tc` domain is not a data-poor setting, where there is sufficient training data and data augmentation is generally effective. MR and SR also show good calibration performance following MELM. Moreover, we can see that applying data augmentation methods do not increase inference time (See Appendix C Table 8). On the other hand, as Table 4 shows, when the target domain is OOD, especially when the target (e.g. OntoNotes 5.0 `wb`) is far from the source domain, the degree of improvement in the uncertainty estimation performance of data augmentation is not large, and sometimes even decreases.

We presume that the augmented data is not far from the original training set, because data augmentation methods we targeted in this study are based on the replacement of tokens or entities. Considering a recent study that indicates models tend to be more overconfident in areas with less training data (Xiong et al., 2023), we can consider calibration performance in OOD sets, especially far from

OntoNotes 5.0 (bc)										
Methods	bn		mz		nw		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	17.54±0.67	25.90±1.29	18.83±0.89	25.65±1.09	23.52±0.77	34.25±1.41	**26.20±1.23**	**28.76±1.30**	57.47±0.87	62.96±0.56
TS	17.19±0.81	24.93±1.27	19.42±1.48	26.32±1.97	23.51±1.08	33.68±1.72	26.85±2.11	29.36±2.35	57.66±1.32	62.96±1.15
LS	17.45±0.96	25.43±1.77	19.38±1.03	26.36±1.56	23.72±1.01	34.23±1.95	26.34±1.78	28.81±2.04	**56.98±1.17**	**62.51±0.91**
MC Dropout	17.50±0.66	25.77±1.58	19.22±1.21	26.39±1.16	23.67±0.73	34.51±1.59	26.32±1.10	28.66±1.12	57.51±1.29	62.80±0.90
LwTR (DA)	17.58±0.44	25.45±1.34	19.34±1.34	26.11±1.56	23.65±0.53	33.89±1.13	27.50±1.73	29.70±2.01	58.68±1.51	63.83±1.22
MR (DA)	17.43±0.62	24.99±1.36	**18.38±1.62**	**24.93±1.73**	**23.28±0.54**	33.35±1.16	26.78±2.19	28.85±2.21	59.01±0.99	64.06±0.76
SR (DA)	**17.01±0.39**	**24.45±0.74**	20.01±1.56	26.94±1.72	23.42±0.66	**33.29±1.33**	26.62±1.59	28.81±1.76	58.14±0.79	63.02±0.59
MELM (DA)	17.22±0.65	24.55±1.41	19.41±0.80	26.01±1.06	23.66±0.85	33.75±1.46	30.11±1.39	32.59±1.71	58.72±1.42	63.71±1.18

OntoNotes 5.0 (bn)										
Methods	bc		mz		nw		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	19.30±0.82	24.37±1.47	20.55±1.59	26.62±2.55	20.05±0.98	28.44±2.25	25.42±0.73	27.56±0.64	**59.02±1.16**	63.61±0.66
TS	19.20±0.88	24.18±1.75	21.21±1.14	27.20±1.72	20.34±0.73	28.80±2.12	25.33±1.28	27.57±1.27	59.11±1.06	**63.60±0.60**
LS	**18.37±0.60**	**22.52±1.41**	21.61±0.47	27.04±1.04	19.98±0.41	27.64±1.11	**24.66±0.48**	**26.69±0.44**	59.92±0.75	63.87±0.77
MC Dropout	18.76±0.97	23.34±1.56	20.91±0.96	26.62±1.82	20.04±0.57	28.25±1.62	25.21±1.27	27.52±1.17	59.09±0.99	63.63±0.54
LwTR (DA)	20.30±0.87	25.42±1.18	20.71±1.01	27.14±1.16	20.51±0.41	29.04±1.26	26.36±2.08	28.67±2.09	59.32±0.97	64.00±0.55
MR (DA)	19.78±1.26	24.35±1.85	20.19±0.47	26.08±1.07	20.42±0.60	27.83±1.74	25.69±0.77	27.75±0.81	59.57±0.96	64.13±0.50
SR (DA)	19.61±0.97	24.08±1.64	**19.79±0.75**	**25.52±1.22**	19.81±0.39	27.18±1.30	26.20±1.56	28.42±1.68	59.86±0.67	63.66±0.40
MELM (DA)	19.93±0.69	23.98±1.09	20.40±0.65	25.54±1.19	**19.73±0.65**	**26.80±1.19**†	28.47±2.14	30.59±2.15	60.51±0.57	64.44±0.33

OntoNotes 5.0 (nw)										
Methods	bc		bn		mz		tc		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	**20.65±1.79**	**25.32±2.15**	**15.24±0.65**	21.06±1.21	22.67±1.24	28.48±2.17	27.81±1.26	30.21±1.39	60.28±1.17	64.30±0.86
TS	21.08±0.75	25.80±1.01	15.61±0.46	21.63±0.80	22.76±1.01	28.92±1.47	28.02±1.61	30.21±1.90	60.37±0.89	64.61±0.68
LS	20.46±1.23	24.63±2.21	15.51±0.55	**20.80±1.70**	22.66±1.10	28.35±1.86	28.50±1.52	30.41±1.21	60.17±1.05	64.07±0.72
MC Dropout	21.25±1.84	25.98±2.09	15.58±0.98	21.59±1.71	22.38±1.10	28.34±1.67	28.05±1.70	30.19±1.79	60.64±0.94	64.63±0.57
LwTR (DA)	21.87±0.87	26.58±0.99	15.81±0.30	21.93±0.41	22.76±0.93	28.38±0.92	**27.60±0.72**	**29.48±0.45**	**59.96±0.46**	**64.06±0.40**
MR (DA)	21.70±0.27	26.29±0.30	15.55±0.87	21.38±2.16	**21.08±1.21**	**26.33±2.14**	30.35±2.69	32.44±2.82	61.16±1.06	65.12±0.80
SR (DA)	21.29±1.37	25.82±1.31	16.00±0.58	21.72±0.22	21.83±0.67	27.37±0.85	33.41±5.50	35.59±5.44	60.58±0.72	64.50±0.54
MELM (DA)	21.96±1.31	26.91±1.88	15.83±0.84	21.76±1.63	21.16±1.38	26.88±1.49	33.92±4.15	36.39±4.03	60.94±0.62	65.03±0.33

OntoNotes 5.0 (tc)										
Methods	bc		bn		mz		nw		wb	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	36.70±1.65	44.25±1.66	35.47±2.48	45.75±2.46	37.15±1.77	47.34±1.79	39.08±0.56	52.50±1.41	**46.38±1.28**	**54.29±1.37**
TS	35.69±2.21	43.34±2.18	34.15±2.65	44.48±2.56	36.38±1.79	46.71±1.43	38.59±1.53	52.58±1.38	47.20±0.92	55.31±1.10
LS	33.91±1.86	41.50±1.75	31.40±2.35	41.24±2.43	34.14±1.91	44.37±1.42	37.04±2.25	50.00±1.92	48.48±1.29	56.10±0.89
MC Dropout	35.83±2.02	43.93±1.75	33.87±2.02	44.31±1.92	36.18±2.43	46.31±2.43	38.97±0.83	52.80±1.08	46.92±2.04	54.95±2.13
LwTR (DA)	34.94±2.42	43.20±1.90	32.61±3.16	43.28±2.55	34.44±1.83	44.98±1.88	37.85±2.13	52.09±1.60	46.78±1.26	54.94±1.84
MR (DA)	35.18±2.89	42.62±2.30	33.50±3.77	42.66±3.20	34.35±2.78	44.78±2.69	37.97±2.64	50.85±3.46	48.61±1.70	55.78±1.90
SR (DA)	34.58±2.40	42.51±1.55	32.66±4.13	42.57±3.28	**32.69±3.21**	43.01±2.83	38.50±1.51	52.00±1.56	46.99±1.27	54.86±1.40
MELM (DA)	**33.05±1.75**	**40.55±2.16**	**29.46±1.55**†	**37.81±1.56**†	33.46±1.66	**42.78±2.55**	**36.79±1.27**	**49.33±2.26**	50.52±1.10	57.27±1.27

Table 4: Results of existing methods and data augmentation methods in OntoNotes 5.0 in OOD test dataset.

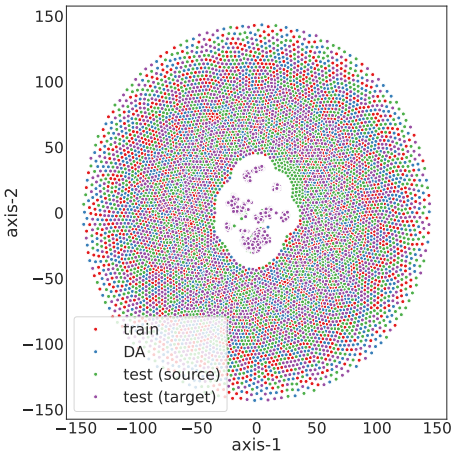

Figure 1: t-SNE plot of token embeddings of OntoNotes 5.0 bn training set (red), generated data by MELM (blue), source domain test set (green) and OntoNotes 5.0 wb test set (purple), respectively.

the source domain, will not improve by data augmentation for NER, while the performance in ID sets will be better than existing methods.

To illustrate this, we performed t-SNE (van der Maaten and Hinton, 2008) for the token embeddings with only entity token from trained Baseline model, shown in Figure 1. We can understand that the token embeddings from augmented data are near the train set or ID test set, while the OOD test sets have some poorly covered regions. Generating sentences that are distant from the training data set and semantically aligned entities from label description for uncertainty estimation is an interesting direction for future research.

AUPRC scores are shown in Table 5. In the AUPRC scores in OntoNotes 5.0, data augmentation methods are outperform existing methods in 15 cases out of 24 cases. Among the existing methods, TS shows superior performance; in data augmentation methods, MELM is not as good as in the case of calibration metrics such as ECE and

Methods	OntoNotes 5.0 (bc)						OntoNotes 5.0 (bn)					
	bc	bn	mz	nw	tc	wb	bc	bn	mz	nw	tc	wb
Baseline	94.72±0.21	95.13±0.43	96.40±0.40	93.27±0.41	92.69±0.57	93.03±0.56	**95.12±0.30**	97.23±0.20	95.83±0.45	95.29±0.27	93.62±0.59	93.13±0.40
TS	**94.89±0.59**	95.14±0.35	96.15±0.51	93.26±0.45	92.78±1.01	92.97±0.83	95.05±0.39	**97.38±0.17**	95.33±0.31	95.23±0.20	**93.96±0.51**	**93.25±0.29**
LS	94.74±0.54	95.09±0.37	96.15±0.30	93.15±0.43	92.60±0.79	92.73±0.36	94.99±0.22	97.32±0.20	95.60±0.22	95.11±0.37	93.49±0.43	92.90±0.47
MC Dropout	94.71±0.31	95.09±0.18	96.07±0.24	93.11±0.43	92.76±0.67	92.88±0.33	95.03±0.34	97.30±0.18	95.78±0.46	95.29±0.19	93.80±0.44	93.22±0.35
LwTR (DA)	94.53±0.28	95.02±0.37	96.22±0.33	93.23±0.23	92.76±0.64	92.91±0.52	94.36±0.54	97.29±0.14	95.74±0.16	95.15±0.20	93.64±0.51	93.08±0.49
MR (DA)	94.44±0.29	94.88±0.24	**96.53±0.43**	**93.4±0.29**	**92.82±0.60**	92.74±0.42	94.57±0.50	97.20±0.19	**96.27±0.31**†	95.11±0.22	93.64±0.55	92.91±0.52
SR (DA)	94.44±0.35	95.09±0.32	95.70±0.40	93.21±0.37	93.24±0.43	**93.06±0.39**	94.76±0.65	97.28±0.15	95.85±0.33	95.30±0.17	93.78±0.63	93.06±0.24
MELM (DA)	94.51±0.16	**95.15±0.34**	96.01±0.29	93.09±0.44	92.64±0.52	92.90±0.47	94.34±0.47	97.24±0.21	96.18±0.32	**95.32±0.32**	93.51±0.50	92.97±0.48
Methods	OntoNotes 5.0 (nw)						OntoNotes 5.0 (tc)					
	bc	bn	mz	nw	tc	wb	bc	bn	mz	nw	tc	wb
Baseline	94.60±0.80	96.36±0.32	95.22±0.48	97.81±0.12	93.32±0.44	93.29±0.46	87.10±1.25	89.22±0.71	84.94±1.61	81.28±2.58	93.45±0.77	89.62±1.10
TS	94.50±0.40	**96.36±0.32**	95.34±0.39	97.74±0.18	93.15±0.52	93.33±0.37	**87.74±1.12**	89.45±0.47	85.95±1.65	82.50±1.35	93.11±0.98	89.93±0.88
LS	**94.65±0.30**	96.23±0.24	95.19±0.57	97.70±0.09	93.05±0.43	93.39±0.41	87.07±1.00	89.57±0.76	**86.67±1.75**	82.79±1.09	92.75±1.06	90.66±0.61
MC Dropout	94.37±0.92	96.32±0.23	95.27±0.31	97.81±0.24	93.40±0.25	93.15±0.47	87.25±0.73	89.02±1.08	85.12±1.62	81.95±2.56	93.36±0.89	90.05±0.84
LwTR (DA)	94.11±0.68	96.33±0.22	95.36±0.29	97.79±0.31	**94.11±0.27**†	92.76±0.25	86.95±0.61	89.74±0.72	86.20±1.67	83.08±1.78	93.70±0.64	90.28±0.55
MR (DA)	93.43±0.13	96.18±0.33	95.01±0.69	97.69±0.12	93.15±0.60	92.67±0.32	86.78±1.12	**90.06±0.61**	86.36±1.64	**83.81±2.79**	**93.69±0.61**	**90.69±1.23**
SR (DA)	94.18±0.92	96.21±0.30	95.45±0.30	97.87±0.14	93.41±0.23	93.39±0.29	86.78±1.49	89.61±0.56	86.42±2.36	81.83±2.85	93.53±0.72	90.04±0.97
MELM (DA)	94.07±0.67	96.09±0.14	**95.67±0.71**	**97.83±0.12**	92.84±0.73	**93.43±0.64**	86.38±1.16	89.05±1.18	86.65±1.37	81.89±2.77	93.30±0.59	89.12±1.47

Table 5: AUPRC scores of existing methods and data augmentation methods in OntoNotes 5.0.

MCE, and MR tends to show superior uncertainty performance. Calibration and scores based on AUC measure different points of uncertainty (Galil et al., 2023), therefore we assume that uncertainties that can be improved vary depending on the methods.

6.2 Cross-lingual Evaluation

The results of cross-lingual transfer in MultiCoNER are shown in Table 6 with English as the source language. MR performs better in uncertainty performance for the ID situation. In contrast to the calibration and uncertainty performance in the cross-genre setting, both MR and SR show better calibration and uncertainty in the OOD setting. In Jiang et al. (2022), the result shows that the larger the linguistic distance (Chiswick and Miller, 2005), the more lenient the calibration and uncertainty estimation tends to be, and similar trends are obtained in this experiment. Unlike the discussion in Section 6.1, the uncertainty performance by data augmentation is also good for OOD in cross-lingual setting because the areas where only target set exist is limited in MultiCoNER (illustrated in Appendix G). On the other hand, MELM, which tends to show excellent calibration performance in cross-genre calibration, does not show good performance in cross-lingual settings.

The amount of data for each language in the CC100 (Conneau et al., 2020) dataset used to train the base model, mDeBERTaV3, was highest for English, followed by German, Spanish, Hindi, and Bangla which correlates with the trend of the calibration results. Moreover, as mentioned in Limisiewicz et al. (2023), languages that tend to have vocabulary overlap between languages in tokenization perform better in cross-lingual transfer in NER. Similar effects may be observed in confidence calibration and uncertainty estimation.

6.3 Detailed Analyzes

We investigate the effects of entity overlap rates and the perplexity of the generated sentences to gain a better understanding of the confidence calibration and uncertainty estimation performance of data augmentation methods for NER. We also investigate the impact of data augmentation size in several settings.

6.3.1 Impact of Augmentation Size

To investigate the impact of data augmentation size on calibration and uncertainty performance, we analyze the trend of evaluation metrics in `tc` $\rightarrow$ `mz` scenario of OntoNotes 5.0 and `EN` $\rightarrow$ `ES` scenario of MultiCoNER, respectively. Figure 2 and 3 illustrate the results in the ID and OOD settings, respectively. In many cases, MR improves the calibration and uncertainty performance by increasing data.[5] SR consistently improves as the dataset size doubles, whereas LwTR demonstrates only marginal improvement or even worsens as the dataset size increases. Finally, MELM improves further for OntoNotes 5.0 `tc`, which shows excellent performance, and deteriorates further for MultiCoNER `EN`, which shows poor performance.

These results show that the calibration algorithm with the best performance for cross-domain transfers is likely to have better performance as the augmentation size is increased. On the other hand, increasing the augmentation size in MR improves the calibration and uncertainty performance compared to similar other data augmentation methods.

Since data augmentation by MR and MELM is performed only on the entity region, the uncertainty estimation performance is relatively less adversely affected by increasing the data augmentation size.

[5]Note that we have not discussed about the absolute values of the uncertainty estimation performance.

Methods	MultiCoNER (EN)											
	EN			DE			ES			HI		
	ECE (↓)	MCE (↓)	AUPRC (↑)	ECE (↓)	MCE (↓)	AUPRC (↑)	ECE (↓)	MCE (↓)	AUPRC (↑)	ECE (↓)	MCE (↓)	AUPRC (↑)
Baseline	28.29±0.30	30.51±0.39	93.04±0.18	31.31±0.52	34.91±0.83	91.97±0.23	31.22±0.28	33.70±0.39	90.87±0.27	46.84±1.64	48.13±1.51	82.04±2.24
TS	28.46±0.43	30.70±0.52	93.13±0.17	31.45±0.70	35.08±1.05	92.02±0.24	31.24±0.41	33.77±0.38	90.92±0.18	46.83±1.38	48.35±1.25	83.01±1.45
LS	28.50±0.57	30.60±0.68	93.12±0.13	31.50±0.64	34.81±0.66	91.93±0.26	31.43±0.58	33.83±0.67	90.82±0.10	46.36±1.23	47.95±1.03	84.00±1.60
MC Dropout	28.57±0.34	30.83±0.54	92.97±0.34	31.64±0.48	35.24±0.68	91.86±0.37	31.47±0.42	33.98±0.40	90.79±0.22	47.42±1.30	48.77±1.23	81.39±3.30
LwTR (DA)	28.17±0.54	30.48±0.77	92.80±0.28	31.13±0.59	34.60±0.78	91.57±0.34	31.10±0.35	33.61±0.51	90.66±0.27	46.70±1.47	47.95±1.30	82.57±1.96
MR (DA)	**28.01±0.42**	**30.08±0.49**†	**93.30±0.24**	**31.12±0.74**	34.71±0.81	**92.05±0.20**	**30.75±0.34**†	**33.24±0.36**†	**91.03±0.15**	46.96±1.20	48.28±1.12	81.75±2.52
SR (DA)	28.15±0.42	30.36±0.48	93.08±0.26	31.17±0.39	**34.42±0.70**	92.02±0.39	31.60±0.55	33.86±0.56	90.65±0.33	**45.85±0.53**	**47.38±0.47**	**84.91±0.91**
MELM (DA)	28.53±0.38	30.68±0.43	92.72±0.22	32.61±0.49	36.14±0.65	91.17±0.29	32.09±0.44	34.38±0.52	90.14±0.30	47.91±1.79	49.18±1.79	81.13±2.41

Table 6: Results of existing calibration methods and data augmentation methods in MultiCoNER.

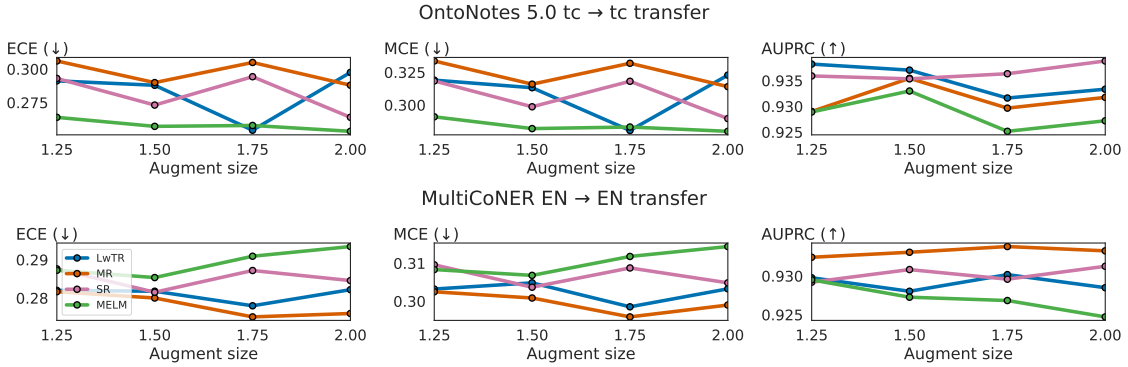

Figure 2: Average values of evaluation metrics for each data augmentation method in ID settings.

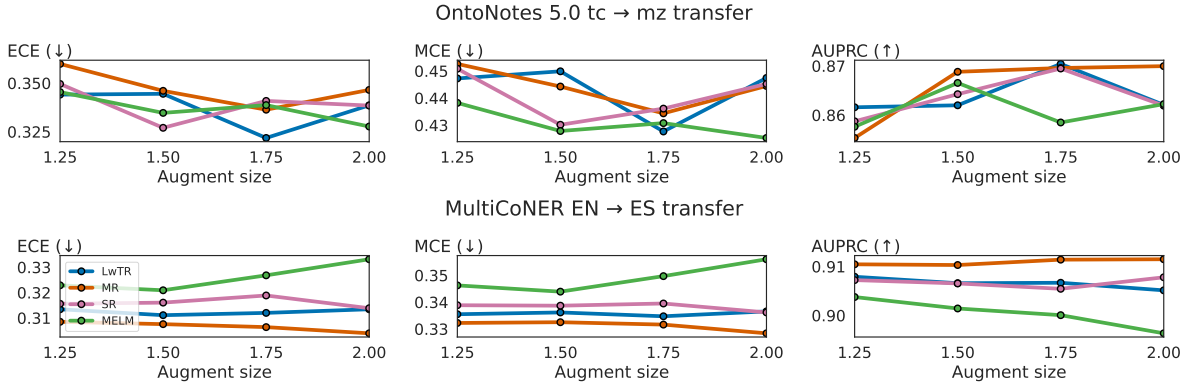

Figure 3: Average values of evaluation metrics for each data augmentation method in OOD settings.

Methods	OntoNotes 5.0 (bc)	OntoNotes 5.0 (bn)	OntoNotes 5.0 (nw)	OntoNotes 5.0 (tc)	MultiCoNER (EN)
LwTR	7.05	7.59	8.28	7.33	6.78
MR	**5.36**	**5.27**	**5.27**	**5.83**	**5.83**
SR	5.91	6.35	6.62	6.02	6.35
MELM	5.56	5.65	5.55	5.90	6.14
(Train)	5.18	4.84	4.86	5.80	5.54

Table 7: Sentences perplexities generated by the data augmentation method in each dataset. Each data augmentation method is performed to increase the training data. Bold means the lowest score in data augmentation methods.

On the other hand, in SR and LwTR, data augmentation that replaces tokens may often inject tokens with inappropriate parts of speech for that sentence, so increasing the data augmentation size often leads to a degradation of uncertainty estimation performance.

6.3.2 Impact of Perplexities for Augmented Sentences

To investigate the influence of replacement units on data augmentation for NER as mentioned in Section 6.3.1, we measured the perplexity of the augmented sentences using GPT-2 (Radford et al., 2019). The average perplexities of the augmented

sentences and the average perplexities of the original training set for each dataset are shown in Table 7. Lower perplexity from augmented sentences tends to improve calibration performance and uncertainty performance. Consistently, the average perplexity of the sentences generated by MR is the lowest. Since MR performs substitutions on an entity-by-entity basis and does not affect the structure of the sentence itself, it has the lowest perplexity among the data augmentation methods in NER.[6] MELM has the second lowest perplexity after MR, and may be adversely affected by generated entities that are adapted to the context but not actually present.

7 Conclusion

In this paper, we investigated the impact of data augmentation on the confidence calibration and uncertainty estimation in NER in terms of genre and language, using several metrics. First, we find that MELM, MR, and SR lead to better calibration and uncertainty performance in the ID setting consistently. On the other hand, in the OOD setting, uncertainty estimation by data augmentation is less effective, especially when the target domain is far from the source domain. Second, our results suggest that the lower the perplexity of the augmented data, as in MR, the further better the calibration and uncertainty performance as the augmentation size is increased. Data augmentation methods for NER do not require changes to the model structure and only require more data to improve entity-level calibration and performance without the need to change the model structure. Our findings indicate the effectiveness of uncertainty estimation through data augmentation for NER, and will be expected to stimulate future research based on their limitations.

Limitations

While this experiment provided valuable insights into the impact of data augmentation on confidence calibration and uncertainty estimation in NER across different genres and languages, there are several limitations that should be acknowledged.

Source Language Due to resource limitations, the experiment was limited to evaluation with English as the source language. To effectively investigate the calibration and uncertainty of zero-shot cross-lingual transfer, it is important to expand the investigation to include a wider range of languages as the source language. Therefore, future research should prioritize the investigation of calibration and uncertainty performance using different languages as the source for zero-shot cross-lingual transfer.

Evaluation of Uncertainty for Entities As mentioned in Section 5.3, regarding the calibration and uncertainty evaluation policy, we simply evaluated an entity span as a single data instance, but a rigorous evaluation method that performs evaluation while considering multiple span candidates has been proposed (Jiang et al., 2022). Establishing span-level NER calibration evaluation methods that can efficiently and comprehensively evaluate calibration and uncertainty for entity types for datasets with many entity types and long sequence lengths is a topic for future research.

NER Paradigm We broadly evaluated the calibration and uncertainty performance in both cross-genre and cross-lingual settings on data augmentation for NER, but only using sequence labeling-based methods. Recently, other paradigms in NER have been proposed such as the span-based methods (Fu et al., 2021) and the generation-based methods (Yan et al., 2021) including BART (Lewis et al., 2020) or Large Language Models (LLM) (Xu et al., 2024), which are also applicable to nested-NER. In the future, the calibration or uncertainty performance of these methods could be evaluated.

Other Data Augmentation Methods In this study, we focused on the data augmentation methods based on token or entity replacement. On the other hand, paraphrase-based data augmentation methods using such as LLM have attracted attention (Ding et al., 2024). By using LLM, it is also possible to generate entities that correspond to a specified entity type (Ye et al., 2024). To Investigate these in the context of uncertainty estimation also will be an interesting research.

Ethical Considerations

In this study, we used existing datasets that have cleared ethical issues. Furthermore, the data augmentation methods we used for uncertainty estimation are substitution-based methods except for MELM, and MELM generated entities from existing datasets that have no ethical issues. Therefore,

[6] As shown in Appendix I, not only the uncertainty performance but also the prediction performance could be affected by preserving the structure of a sentence.

it is unlikely that toxic sentences would be generated.

Acknowledgements

The authors also acknowledge the Nara Institute of Science and Technology's HPC resources made available for conducting the research reported in this paper.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251, Online. Association for Computational Linguistics.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2022. Style transfer as data augmentation: A case study on named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1827–1841, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

B.R. Chiswick and Paul Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26:1–11.

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.

Brooke Cowan, Sven Zethelius, Brittany Luk, Teodora Baras, Prachi Ukarde, and Daodao Zhang. 2015. Named entity recognition in travel-related search queries. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3935–3941. AAAI Press.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

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

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Jerome H. Friedman. 2000. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning*

Research, pages 1050–1059, New York, New York, USA. PMLR.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. 2023. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers? In *The Eleventh International Conference on Learning Representations*.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. 2021. Improving predictions of bayesian neural nets via local linearization. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 703–711. PMLR.

Abhyuday Jagannatha and Hong Yu. 2020. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, Online. Association for Computational Linguistics.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2022. Calibrating zero-shot cross-lingual (un-)structured predictions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2648–2674, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaeyoung Kim, Dongbin Na, Sungchul Choi, and Sungbin Lim. 2023. Bag of tricks for in-distribution calibration of pretrained transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 551–563, Dubrovnik, Croatia. Association for Computational Linguistics.

Satwik Ram Kodandaram, Kushal Honnappa, and Kunal Soni. 2021. Masking private user information using natural language processing. *International Journal of Advance Research in Computer Science and Management*, 7:1753–1763.

Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, page 6405–6416.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Shen Yan, Laura Leal-Taixé, James Hays, and Deva Ramanan. 2023. Soft augmentation for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16241–16250.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In

International Conference on Learning Representations.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David J. Miller, Ajit V. Rao, Kenneth M. Rose, and Allen Gersho. 1996. A global optimization technique for statistical classifier design. *IEEE Trans. Signal Process.*, 44:3108–3122.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2901–2907. AAAI Press.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the Inter- national Conference on Learning Representations (Workshop)*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Steven Reich, David Mueller, and Nicholas Andrews. 2020. Ensemble Distillation for Structured Prediction: Calibrated, Accurate, Fast—Choose Three. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5583–5595, Online. Association for Computational Linguistics.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Konstantinos N. Vavliakis, Andreas L. Symeonidis, and Pericles A. Mitkas. 2013. Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering*, 88:1–24.

Yeming Wen, Ghassen Jerfel, Rafael Muller, Michael W Dusenberry, Jasper Snoek, Balaji Lakshminarayanan, and Dustin Tran. 2021. Combining ensembles and data augmentation can harm your calibration. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Miao Xiong, Ailin Deng, Pang Wei Koh, Jiaying Wu, Shen Li, Jianqing Xu, and Bryan Hooi. 2023. Proximity-informed calibration for deep neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *arXiv preprint arXiv:2312.17617*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. 2022. When and how mixup improves calibration. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26135–26160. PMLR.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.

A Licenses of Datasets

OntoNotes 5.0 can be used for research purposes as described in https://catalog.ldc.upenn.edu/LDC2013T19. MultiCoNER dataset is licensed by CC BY 4.0 as described in https://aws.amazon.com/marketplace/pp/prodview-cdhrtt7vq4hf4.

B Details of Existing Calibration Methods

In this section, we describe the popular baseline methods for confidence calibration. We use the following notations: $\boldsymbol{z}_i$ denotes the logits for class i, p_i denotes the calibrated probability for class i, y_i denotes the label for class i, and K denotes the number of classes.

Methods	Inference time [s]
Baseline	14.90 ± 0.10
TS	15.53 ± 0.92
LS	14.94 ± 0.24
MC Dropout	271.77 ± 1.81
LwTR	14.91 ± 0.10
MR	14.93 ± 0.17
SR	14.83 ± 0.12
MELM	14.89 ± 0.14

Table 8: Inference time for each algorithm on MultiCoNER EN full test data.

B.1 Temperature Scaling (TS)

TS (Guo et al., 2017) is a post-processing technique for calibrating the confidence scores outputted by a neural network. It involves scaling the logits (i.e., the outputs of the final layer before the softmax) by a temperature parameter T before applying the softmax function to obtain the calibrated probabilities. The softmax function takes a vector of logits $\boldsymbol{z}$ and returns a distribution $\boldsymbol{p}$:

$$p_i = \frac{\exp(\boldsymbol{z}_i/T)}{\sum_{j=1}^{K}\exp(\boldsymbol{z}_j/T)}.$$

B.2 Label Smoothing (LS)

LS (Miller et al., 1996; Pereyra et al., 2017) is a regularization technique used to improve the calibration and generalization performance of the model. By introducing a small degree of uncertainty in the target labels during training, label smoothing mitigates overfitting and encourages the model to learn more robust and accurate representations, ultimately contributing to improved overall performance on the task at hand. LS is characterized by introducing a smoothing parameter ϵ and smoothed label y_i^{LS} as follows,

$$y_i^{LS} = y_i(1-\epsilon) + \frac{\epsilon}{K}.$$

B.3 Monte-Carlo Dropout (MC Dropout)

MC Dropout is a regularization technique that can be used for uncertainty estimation in neural networks (Gal and Ghahramani, 2016). In this method, we need to run the model M times with different dropout masks and take the average softmax output over all the runs (We use $M = 20$). The procedure can be represented using the following formula:

$$p_i = \frac{1}{M}\sum_{t=1}^{M}\frac{\exp(\boldsymbol{z}_i^{(t)})}{\sum_{j=1}^{K}\exp(\boldsymbol{z}_j^{(t)})}.$$

Methods	bc		bn		nw		tc	
	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)	ECE (↓)	MCE (↓)
Baseline	**27.07**	33.52	**26.08**	31.17	26.66	31.35	37.66	46.32
TS	27.25	33.41	26.17	31.17	26.68	31.34	36.66	45.52
LS	27.19	33.57	25.88	**30.49**	26.52	30.67	35.24	43.68
MC Dropout	27.15	33.61	25.90	30.85	26.62	31.18	36.80	45.71
LwTR (DA)	27.65	33.78	26.49	31.78	27.28	31.67	35.90	44.97
MR (DA)	27.33	33.22	26.21	31.00	**26.26**	**30.53**	36.38	44.65
SR (DA)	27.23	**33.08**	26.11	30.72	27.47	31.89	35.24	43.57
MELM (DA)	27.95	33.88	26.63	30.91	27.62	32.09	**34.83**	**42.65**

Table 9: ECE and MCE averaged over all target domain results in OntoNotes 5.0.

C Inference Time

Table 8 shows the results of the inference time on MultiCoNER EN set. We can see that data augmentation methods do not affect the computational overhead during inference clearly.

D Full Averaged Results on OntoNotes 5.0

To briefly summarize the many values in Table 3 and 4, we averaged the ECE and MCE scores for each method and domain, shown them in Table 9. From this table, we can see that data augmentation methods are slightly worse than existing methods some cases when averaging all settings , while in others, especially nw and tc, data augmentation methods are better on average.

E More Results about Test Set Duplication

Table 10 shows the results of the percentage increase in entity duplication that are new overlaps with each target domain's test set when applying each data augmentation method except MR, where the source domains are bc, bn, and nw. In all cases there is only a small increase. These results and the MR, which shows good calibration and uncertainty performance indicated from Section 6.1 and 6.2, do not increase the number of new entities in the training data set suggest that the entity overlap rate does not affect calibration and uncertainty estimation.

F Impact of New Entities via Data Augmentation

To investigate the impact of new entities added by data augmentation methods on calibration performance, we measured the percentage of new entities added in the training data and the percentage of new entities that overlap with the test set. Table 11 shows the percentage of new entities increased by data augmentation with the train set as the source domain in each dataset. In all data sets, MELM has observed the most increase of the new entities in the augmented data set. On the other hand, MR that shows good calibration performance followed by MELM does not increase the number of new entities because the replacement is based on the entities in the original training data. Furthermore, the entities generated have little overlap with the target domain, as shown in Table 12. Therefore, new entities by data augmentation methods for NER are likely to have no effect on calibration performance or uncertainty performance.

Methods	OntoNotes 5.0 (bc)					
	bc	bn	mz	nw	tc	wb
LwTR	0.27	0.26	0.00	0.14	1.83	0.30
SR	0.00	0.18	0.00	0.14	0.00	0.15
MELM	0.41	0.53	0.19	0.17	0.91	0.45
Methods	OntoNotes 5.0 (bn)					
	bc	bn	mz	nw	tc	wb
LwTR	0.55	0.35	0.19	0.35	0.91	0.60
SR	0.55	0.26	0.19	0.21	0.00	0.45
MELM	0.68	0.35	0.37	0.10	0.46	0.30
Methods	OntoNotes 5.0 (nw)					
	bc	bn	mz	nw	tc	wb
LwTR	0.96	1.23	0.37	0.52	5.02	1.34
SR	0.41	0.09	0.56	0.21	0.46	1.04
MELM	1.10	0.79	1.48	0.55	1.37	0.45

Table 10: The percentage of new entities generated by each data augmentation method using the training set in the case of the source domain bc, bn and nw.

G t-SNE Plot for MultiCoNER Dataset

To overview of the ID and OOD data instances in the MultiCoNER dataset, t-SNE plot is shown in Figure 4.

H Results for Low-resource Language

To investigate the uncertainty estimation performance for low-resource language, we additionaly show the results of 10,000 examples of Bangla (BN) from MultiCoNER dataset in Table 13 when source language is EN. The results show that data augmentation is also effective in uncertainty estimation for

Methods	OntoNotes 5.0 (bc)	OntoNotes 5.0 (bn)	OntoNotes 5.0 (nw)	OntoNotes 5.0 (tc)	MultiCoNER (EN)
LwTR	27.77	32.69	38.65	19.83	18.46
MR	0.00	0.00	0.00	0.00	0.00
SR	25.23	26.34	35.13	8.56	20.45
MELM	45.26	45.95	43.37	34.75	37.64

Table 11: The percentage increase in new entities when each data augmentation method is performed on the original train set.

Methods	bc	bn	mz	nw	tc	wb
LwTR	0.00	0.00	0.00	0.10	0.00	0.00
SR	0.14	0.00	0.00	0.10	0.00	0.15
MELM	0.27	0.35	0.19	0.14	0.00	0.30

Table 12: The percentage increase in entity duplication in the case of the source domain tc that are new overlaps with each target domain's test set when applying each data augmentation method except MR. More results are in Appendix E.

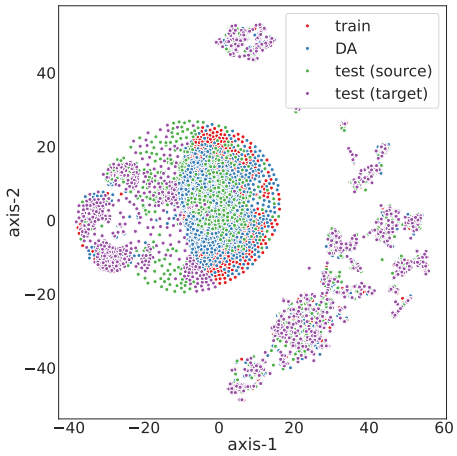

Figure 4: t-SNE plot of token embeddings of MultiCoNER EN training set (red), generated data by SR (blue), source domain test set (green) and MultiCoNER HI test set (purple), respectively.

Methods	ECE (↓)	MCE (↓)	AUPRC (↑)
Baseline	49.60±2.02	51.32±1.96	79.49±2.21
TS	48.85±1.89	50.60±1.60	79.09±4.22
LS	48.00±1.97	49.91±1.54	79.60±3.51
MC Dropout	49.29±2.20	50.93±2.14	78.31±2.52
LwTR (DA)	48.66±1.35	50.22±1.36	80.93±1.75
MR (DA)	49.54±2.65	51.20±2.65	79.17±2.97
SR (DA)	**47.67±0.98**	**49.46±0.88**	**81.96±1.35**
MELM (DA)	50.77±0.88	52.15±0.81	75.55±2.59

Table 13: Results of existing methods and data augmentation methods in MultiCoNER BN.

low-resource language.

I F_1 Scores

Table 14 and 15 show F_1 scores. Note that in many cases, data augmentation methods do not degrade predictive performance itself, but MELM often significantly degrades predictive performance in some cases, especially when the source domains are nw and tc. Considering Section 6.1 and 6.2, MR improves calibration and uncertainty performance in many cases without degrading predictive performance.

Methods	OntoNotes 5.0 (bc)						OntoNotes 5.0 (bn)					
	bc	bn	mz	nw	tc	wb	bc	bn	mz	nw	tc	wb
Baseline	81.39±0.78	80.86±1.03	81.61±1.36	75.49±0.90	68.83±1.27	45.74±0.74	80.74±1.21	90.25±0.36	81.47±0.96	81.04±0.64	72.36±1.88	46.86±0.52
TS	81.10±0.94	81.19±0.89	80.80±1.37	75.14±1.60	69.20±2.73	45.58±1.02	81.31±1.18	**90.37±0.49**	80.96±1.32	**81.13±0.62**	71.83±1.76	46.50±0.69
LS	81.21±1.11	81.17±0.91	81.43±1.33	75.30±1.26	69.64±1.45	45.75±0.82	**82.08±0.62**	90.32±0.36	81.22±0.52	80.95±0.37	72.45±1.38	46.69±0.60
MC Dropout	81.49±0.80	81.06±0.71	81.12±0.63	75.24±1.02	69.53±1.78	45.73±0.46	81.55±0.63	90.21±0.36	80.80±1.10	81.11±0.46	**73.13±1.97**	46.71±0.60
LwTR (DA)	80.85±0.82	80.91±0.93	81.45±1.08	75.33±0.82	68.40±0.94	45.53±0.84	79.43±1.13	89.98±0.40	80.75±0.67	80.33±0.31	69.62±1.80	46.23±0.54
MR (DA)	80.93±0.61	80.88±0.61	**82.02±0.66**	**75.66±0.79**	69.49±1.78	45.38±0.72	79.93±1.43	90.07±0.23	**81.70±0.61**	80.54±0.50	72.44±1.46	46.45±0.47
SR (DA)	**81.52±0.69**	**81.20±0.78**	79.93±0.95	75.08±0.89	**69.86±1.30**	**46.04±0.57**	80.24±1.44	90.05±0.21	80.92±0.93	80.84±0.42	70.80±1.66	**46.98±0.61**
MELM (DA)	81.08±0.37	80.81±0.97	80.11±0.98	74.74±1.24	66.68±1.18	45.19±1.05	79.23±0.64	90.26±0.38	81.48±0.65	80.66±0.79	68.42±1.65	46.36±0.44
Methods	OntoNotes 5.0 (nw)						OntoNotes 5.0 (tc)					
	bc	bn	mz	nw	tc	wb	bc	bn	mz	nw	tc	wb
Baseline	74.34±4.10	83.08±1.19	73.56±3.31	90.08±0.31	**72.59±1.34**	46.47±0.59	55.29±2.01	59.13±2.80	50.68±3.51	46.14±4.31	69.52±1.45	40.85±1.36
TS	75.34±1.67	83.02±0.98	75.01±2.21	90.04±0.24	71.98±1.17	46.29±0.87	**56.81±2.05**	59.04±2.95	52.98±3.34	48.85±3.26	67.45±2.30	41.12±1.27
LS	**76.60±1.65**	**83.27±1.49**	**75.79±2.00**	90.20±0.26	71.91±2.67	46.68±0.69	53.98±3.40	56.12±6.02	51.17±5.94	48.62±4.82	66.01±3.26	40.63±1.83
MC Dropout	75.07±2.84	82.69±2.11	73.79±2.23	89.98±0.56	71.96±1.43	46.25±0.92	55.16±1.70	58.95±2.87	51.11±3.75	47.31±4.48	69.15±3.05	40.57±1.44
LwTR (DA)	74.80±1.57	83.01±0.41	75.01±3.35	89.79±0.28	70.85±1.13	**46.78±0.54**	54.01±2.14	**60.86±2.89**	**53.89±3.76**	**50.20±3.77**	**69.53±1.60**	40.80±0.97
MR (DA)	73.57±1.09	81.52±2.09	71.43±3.80	89.90±0.34	68.31±3.52	44.88±1.38	53.73±2.35	57.46±3.70	52.74±3.27	46.90±4.87	68.57±2.71	40.50±1.79
SR (DA)	73.64±3.45	82.03±2.14	72.25±4.88	**90.24±0.11**	66.18±4.59	46.38±1.45	53.41±2.46	58.54±3.20	53.08±4.85	46.48±7.08	68.13±1.41	**41.20±1.23**
MELM (DA)	73.46±2.46	82.22±1.23	75.56±2.60	89.94±0.18	62.43±2.95	45.19±0.97	48.01±5.27	49.59±6.16	48.93±4.11	42.09±5.61	63.46±2.28	36.16±3.76

Table 14: F_1 scores of existing calibration methods and data augmentation methods in OntoNotes 5.0.

Methods	EN	DE	ES	HI
Baseline	68.80±0.38	64.91±0.60	63.53±0.41	37.33±3.77
TS	68.51±0.52	64.70±0.90	63.41±0.45	37.90±2.79
LS	69.17±0.55	65.37±0.51	63.83±0.32	39.93±3.50
MC Dropout	68.56±0.96	64.70±0.87	63.39±0.69	36.38±5.89
LwTR (DA)	68.86±0.82	64.95±0.64	63.52±0.85	38.24±3.11
MR (DA)	**69.71±0.72**	**65.37±0.57**	**64.25±0.62**	37.53±4.03
SR (DA)	68.81±0.41	64.75±0.86	63.85±0.46	**42.31±1.45**
MELM (DA)	68.57±0.54	63.40±0.49	62.76±0.64	37.78±3.16

Table 15: F_1 scores of existing calibration methods and data augmentation methods in MultiCoNER.