Investigating the effectiveness of Data Augmentation and Contrastive Learning for Named Entity Recognition

Noel Chia
University of Mannheim
Germany
neraug@noelchia.com

Ines Rehbein
University of Mannheim
Germany
rehbein@uni-mannheim.de

Simone Paolo Ponzetto
University of Mannheim
Germany
ponzetto@uni-mannheim.de

Abstract

Data Augmentation (DA) and Contrastive Learning (CL) are widely used in NLP, but their potential for NER has not yet been investigated in detail. Existing work is mostly limited to zero- and few-shot scenarios where improvements over the baseline are easy to obtain. In this paper, we address this research gap by presenting a systematic evaluation of DA for NER on small, medium-sized and large datasets with coarse and fine-grained labels. We report results for a) DA only, b) DA in combination with supervised contrastive learning, and c) CL with transfer learning. Our results show that DA on its own fails to improve results over the baseline and that supervised CL works better on larger datasets while contrastive transfer learning (CTL) is beneficial if the target dataset is very small. Finally, we investigate how contrastive learning affects the learned representations, based on dimensionality reduction and visualisation techniques, and show that CL mostly helps to separate named entities (NEs) from non-entities.

1 Introduction

Named Entity Recognition (NER) has been widely studied in NLP and has many applications in the computational social sciences and the digital humanities. Many of these applications, however, require the adaptation to new languages or genres for which no or only small amounts of annotated data are available. A major disadvantage of supervised NER systems is their dependence on large and representative datasets for training (Li et al., 2022b). Consequently, the scarcity of labelled data has become one of the major challenges impeding the performance of NER systems, especially in highly specialised domains.

Data Augmentation (DA) seems like a compelling solution to address this problem. By applying transformations to the data, new training instances can be generated, thus reducing the amount of manually annotated data needed to train the model (Perez and Wang, 2017). Many studies have applied DA to text classification tasks, summarisation, or question answering (Li et al., 2022a; Pellicer et al., 2023), with a focus on low-resource scenarios. We are not aware of any studies that report improved results for DA over strong baselines, such as transformers, for medium to large data sizes.

Furthermore, there is a lack of research on DA for token-level tasks such as NER, where the integration of DA presents a unique challenge. Several DA techniques apply transformations directly to tokens, thus changing their contextual information. As a consequence, this process may inadvertently modify the associated entity labels, disrupting the correspondence between tokens and their intended NEs (Dai and Adel, 2020). This challenge underscores the necessity of developing augmentation strategies that preserve the entity labels while enhancing the diversity and robustness of the training data for improved NER model performance.

Another promising approach to improve model performance is contrastive learning (CL), where the model learns to position representations of instances from the same class closer together in the embedding space while representations for data points that belong to different classes are pushed further apart. CL can be used on its own but can also be combined with DA and transfer learning.

In the paper, we address the question of which of the techniques described are effective in improving results for NER on small, medium-sized and large datasets.[1] Our main contributions include:

- a systematic evaluation of DA, CL and transfer learning for NER,

[1] Our source code is openly available at https://codeberg.org/noelchia/NER-Aug

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 66–79
March 3-4, 2025 ©2025 University of Tartu Library

- an adaptation of supervised contrastive learning for token-level tasks, and
- a visual analysis of the learned representations.

2 Related Work

2.1 Data Augmentation for NER

Only a few studies have applied DA techniques to NER, focussing mostly on low-resource settings. One possible reason for this is that DA reduces overfitting and thus improves the generalisability of the model. Since overfitting is most common and severe for small datasets, we can expect the greatest benefit of DA in this context.

Dai and Adel (2020) explore simple data augmentations such as label-wise token replacement, synonym replacement, mention replacement and shuffle the order of tokens within segments on data from the biomedical and materials science domain. Their transformer-based tagger obtains improvements only for small dataset sizes (≤ 500 instances) but not when training on the full data. Ding et al. (2020) introduce an approach dubbed DAGA where they generate training examples for NER and other token-level NLP tasks using language models. Instead of producing unlabelled text, they generate new labelled training examples.

Zhou et al. (2022) propose Masked Entity Language Modelling (MELM) where they train a language model to generate NEs, conditioned on a masked sentence with NE tags. The main difference between DAGA and MELM is that DAGA generates the entire sentence, while MELM uses pre-existing instances and only replaces existing NEs by masking them and generating a new entity of the same class. Both approaches have been evaluated in low-resource scenarios.

Instead of low-resource NER, Chen et al. (2021) focus on DA for cross-domain NER, using an approach that learns textual patterns and transforms the text from a high-resource to a low-resource domain. based on denoising reconstruction, detransforming reconstruction and domain classification. Cai et al. (2023) leverage graph propagation to create new data points, based on the relationship between labelled data and unlabelled natural texts, and evaluate their method in low-resource and cross-domain settings.

Zhang et al. (2022) develop two data augmentation methods for a BART based generative NER model. Theirs is the only work we are aware of that addresses the problem of DA in in-domain settings with medium and large data sizes.

2.2 Contrastive Learning for NER

Contrastive learning (CL) is a discriminative machine learning technique that aims to create similar representations for data points that belong to the same class while pushing samples from different classes further apart in distributional space (Kumar et al., 2022). CL can be used in (semi)-supervised and unsupervised settings and is very popular because it allows the application of self-supervised learning to tasks that were previously only possible in supervised environments (Le-Khac et al., 2020; Liu et al., 2023). However, only few papers apply CL to NER, and most of these focus on few-shot learning.

Huang et al. (2022) introduce COPNER, a method to create prototypical tokens that represent each class. During contrastive training, the token representing the class forms positive pairs with NE tokens from that class while class tokens paired with words from other classes are considered as negative pairs. He et al. (2023) use a similar idea to develop a template-free prompting method for few-shot NER. Using external knowledge like textual descriptions of entity types, they generate anchors to represent the entity type. These anchors are then appended to the end of the input sentence. The authors use CL to train the encoder to produce representations of words that are similar to the corresponding entity type.

Das et al. (2022) use contrastive learning to train a model dubbed CONTaiNER, which models the distribution of token classes using Gaussian Embeddings. Tokens from the same class are considered as positive pairs, and all other valid pairs are assumed to be negative. Li et al. (2023) also use Gaussian embeddings, but add a cross-domain attention layer based on HaloNet (Vaswani et al., 2021). Si et al. (2022) propose Span-based Contrastive Learning with Retrieval Augmented Inference (SCL-RAI). Their model focusses on NEs that have been mislabelled as negative instances by the system.

All of the papers above either focus on few-shot scenarios or train their CL method on small data sizes of less than 5,000 instances.

3 Experimental Settings

The last section has shown that there is a severe lack of research regarding the effectiveness of DA and CL for NER in scenarios where ample training data is available. We address this gap by providing a systematic investigation of both techniques in different settings and comparing their impact in isolation and in combination with transfer learning.

Datasets We select three different-sized English datasets with coarse and fine-grained entity type distinctions. CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) is the smallest dataset with 14 thousand training examples consisting of 301 thousand tokens, encoding four NE types only (Person, Location, Organisation, Miscellaneous). The second dataset, OntoNotes Release 5.0 (Weischedel et al., 2013), is medium sized with 82 thousand instances, over 2 million tokens and encodes 18 different NE types. The largest dataset is Few-NERD (Ding et al., 2021) with more than 131 thousand sentences, 4.6 million tokens and 66 fine-grained NE types. The fine-grained NE types are further grouped into 8 coarse-grained NE types. We use the original train, dev and test splits for CoNLL 2003 and Few-NERD. The authors of OntoNotes Release 5.0 did not release the dataset with predefined train, dev and test splits, so the splits suggested in Pradhan et al. (2013) were used.

Baseline Model We chose RoBERTa (Liu et al., 2019) as our baseline model, as it yields competetive results at reasonable training costs. Our implementation uses the `RobertaForTokenClassification` architecture from the Huggingface Transformers library (Wolf et al., 2020) which adds one additional linear layer on top of RoBERTa.

3.1 Data Augmentation Methods

We adapt three common approaches to data augmentation for NER, namely round-trip translation, paraphrasing and masking.[2]

Round-Trip Translation Sennrich et al. (2016) proposed to augment monolingual training data with automatic backtranslations to increase the size of the data. Inspired by this, we performed round-trip translation, where we translate a sentence into another language and then back to the original language create a different sentence. We check the round-trip translated output by string matching every NE in the original sample to the augmented sample. If all NEs are found, then the entities are labelled based on the assumption that all string matches represent the same NE, and all other words are not NEs. The neural machine translation model chosen is No Language Left Behind (NLLB) (NLLB Team et al., 2022) and we use translations to/from German. We also experimented with French and Zulu, with very similar results.

For a task like NER that is sensitive to token-level changes, round-trip translation might result in missing or modified NE labels. Hence, checks are performed to ensure that all NE tokens are preserved before adding the augmented data to the training set (for details, see appendix A.1).

Paraphrasing We use T5 (Raffel et al., 2020) to generate paraphrases for our data (also see appendix A.2). The model has been fine-tuned by Vorobev and Kuznetsov (2023b) on the ChatGPT paraphrases dataset, which includes the Quora Question Pairs (QQP) (Iyer et al., 2017), the Stanford Question Answering Dataset (SQuAD) version 2.0 (Rajpurkar et al., 2018) and the CNN / DailyMail Dataset (Hermann et al., 2015). ChatGPT was used to create five paraphrases for each example in the three datasets to train the T5 model.

Masking Inspired by Shen et al. (2020), we randomly mask tokens to produce augmented data. Masking aims to reduce overfitting by forcing the model to learn to predict NEs even when the token or its context is masked. We add a consistency loss to the loss function to encourage the model to make similar predictions for both the original and masked instances (Eq. 1 below).

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{L}_{\mathrm{ce}}(\boldsymbol{x}_{\mathrm{masked}}, \boldsymbol{y}) + \mathcal{L}_{\mathrm{KL}}(\boldsymbol{x}, \boldsymbol{x}_{\mathrm{masked}}) \quad (1)$$

$\mathcal{L}_{\mathrm{ce}}$ denotes the cross-entropy loss, and $\mathcal{L}_{\mathrm{KL}}$ the Kullback-Leibler (KL) divergence loss. For each example $\boldsymbol{x}$ with target labels $\boldsymbol{y}$, an augmented sample $\boldsymbol{x}_{\mathrm{masked}}$ will be generated, where every token in $\boldsymbol{x}_{\mathrm{masked}}$ will have a 15% probability of being replaced by a `[MASK]` token (also see appendix A.3 for more details).

3.2 Supervised Contrastive Learning for NER

Khosla et al. (2020) propose the supervised contrastive (SupCon) loss for computer vision, a supervised variation of contrastive learning that also

[2] More detailed information on the different DA techniques and settings, including the number of augmented instances for each method and dataset, are provided in appendix A.1.

makes use of labelled images of the same class as additional positive pairs. This approach allows us to integrate contrastive learning into the downstream task, thus reducing the time requirements for task-specific fine-tuning after the CL step.

We adapt supervised contrastive learning for NER by considering each contextualised token embedding generated by RoBERTa as a training example and add two fully connected layers to the model. The objective of this training step is to maximise the similarity of the contextualised representations for tokens that belong to the same NE type, and to minimise the similarity otherwise. After the contrastive learning step, we add a new fully connected layer to the model and perform task-specific fine-tuning.

Adapting the SupCon loss for NER Tian et al. (2023) show that SupCon is similar to calculating the cross-entropy loss. Let $i \in I \coloneqq \{1, 2, ..., N\}$ be the index of a sample, and $a \in A(i) \coloneqq I \setminus \{i\}$ be the index of a different sample. $\boldsymbol{x}_i$ is a training example with its corresponding label y_i, and is mapped to projection z_i by the contrastive model. $\tau \in \mathbb{R}^+$ is a scalar temperature variable. First, a contrastive categorical distribution $\boldsymbol{q}_i$ is constructed to describe how closely z_i matches z_j for $j \in A(i)$ (see Eq. 2).

$$q_{i,j} = \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \tag{2}$$

If there is at least one element in $A(i)$, then the weighing term of the contrastive loss can be calculated similarly to the cross entropy ground-truth categorical distribution $\boldsymbol{p}_i$ as shown in Eq. 3 where the indicator function $\mathbb{1}_{\text{match}}(i, j)$ indicates whether there is a match ($y_i = y_j$).

$$p_{i,j} = \frac{\mathbb{1}_{\text{match}}(i, j)}{\sum_{a \in A(i)} \mathbb{1}_{\text{match}}(i, a)} \tag{3}$$

The supervised contrastive loss is the cross entropy between the ground-truth distribution $\boldsymbol{p}_i$ and the contrastive distribution $\boldsymbol{q}_i$, as shown in Eq. 4.

$$\mathcal{L} = \sum_{i \in I} H(\boldsymbol{p}_i, \boldsymbol{q}_i) = -\sum_{i \in I} \sum_{j \in J} p_{i,j} \log q_{i,j} \tag{4}$$

We implement the loss function in Eq. 4 for contrastive learning for NER. The projection head used for supervised learning consists of a hidden layer with ReLU activation before the final linear projection, as Chen et al. (2020) showed that this performs better than the single linear projection layer used in some contrastive learning models.

3.2.1 Contrastive Learning with DA

We test combinations of DA and CL, using masking (see section 3.1). This augmentation was chosen because of its computational efficiency, requiring only a random number generator to select words for random masking. In contrast, round-trip translation and paraphrasing both require a separate model to generate the input, making it difficult to perform the augmentation during training.

3.2.2 Contrastive Transfer Learning

Experiments on contrastive learning in other domains, such as computer vision (Chen et al., 2020), suggest that the representations produced by CL tend to be highly adaptable across different tasks and domains. We will test the hypothesis that the representations produced by training on one NER dataset can be applied to another NER dataset to improve the model's performance.

This could be useful for practical applications, especially for cases where only a small set of labelled data is available. By first performing contrastive learning on a larger dataset and then fine-tuning the learned representations on the smaller dataset, better performance could be achieved. This might be an alternative to data augmentation or could be used in combination with data augmentation to further improve results. To assess the effectiveness of CTL for NER and explore how different dataset properties affect the results, we test all six possible combinations of datasets.

4 Results

Data Augmentation We first look at the results for the three DA methods, i.e., round-trip translation, paraphrasing and masking (Table 1). All results are averaged over five runs, with the standard deviation (STDEV) shown in subscript. Statistically significant improvements over the baseline are underlined.

As shown in Table 1, the three data augmentation methods mostly fail to produce statistically significant improvements. Paraphrasing is the worst performer, often producing similar or sometimes even worse results than the baseline. One reason for this lack of improvement might be that the T5 model used for paraphrasing is trained on similar data as RoBERTa, so the paraphrased results represent a distribution of the data that RoBERTa has already seen during pre-training. Hence, the model

Dataset Size (Sentences)	100	500	1000	5000	Full
CoNLL-2003 (4 NE types)			Mean F1 Score% ± STDEV		
Baseline	83.32± 0.36	88.23 ± 0.60	89.84 ± 0.46	91.15 ± 0.23	91.98 ± 0.43
DA Translate	83.66 ± 0.81	88.40 ± 0.45	90.01 ± 0.36	91.38 ± 0.36	92.23 ± 0.39
DA Paraphrase	83.37 ± 0.65	88.21 ± 0.47	89.81 ± 0.52	91.22 ± 0.32	92.19 ± 0.61
DA Mask	80.83 ± 0.38	88.47 ± 0.33	$\underline{90.48}$ ± 0.33	91.34 ± 0.30	$\underline{92.47}$ ± 0.37
CL	82.52 ± 0.77	88.86 ± 0.49	90.06 ± 0.26	$\underline{91.53}$ ± 0.30	$\underline{92.49}$ ± 0.29
DA Mask + CL	80.90 ± 0.58	88.12 ± 0.59	89.95 ± 0.39	$\underline{91.56}$ ± 0.38	92.20 ± 0.26
OntoNotes v5 (18 NE types)			Mean F1 Score% ± STDEV		
Baseline	66.01 ± 1.58	77.90 ± 0.55	82.21 ± 0.32	85.58 ± 0.42	89.28 ± 0.25
DA Translate	66.03 ± 0.73	77.37 ± 0.50	81.63 ± 0.53	85.24 ± 0.41	89.21 ± 0.33
DA Paraphrase	66.02 ± 1.16	77.28 ± 0.66	81.58 ± 0.55	84.97 ± 0.30	88.34 ± 0.41
DA Mask	60.65 ± 0.99	77.26 ± 0.51	82.16 ± 0.33	85.80 ± 0.45	88.83 ± 0.74
CL	65.82 ± 0.79	$\underline{78.71}$ ± 0.33	82.42 ± 0.32	$\underline{86.51}$ ± 0.46	$\underline{89.76}$ ± 0.25
DA Mask + CL	65.89 ± 1.52	$\underline{78.76}$ ± 0.56	82.12 ± 0.37	86.02 ± 0.39	89.65 ± 0.55
Few-NERD (66 NE types)			Mean F1 Score% ± STDEV		
Baseline	38.77 ± 0.82	54.07 ± 0.89	58.17 ± 0.67	62.09 ± 0.41	67.90 ± 0.59
DA Translate	38.22 ± 1.69	54.27 ± 0.37	57.93 ± 0.38	62.42 ± 0.41	67.95 ± 0.70
DA Paraphrase	38.87 ± 0.72	54.16 ± 0.53	57.95 ± 0.33	62.36 ± 0.41	67.42 ± 0.97
DA Mask	35.85 ± 0.64	52.89 ± 0.57	56.66 ± 0.41	62.01 ± 0.30	63.72 ± 0.87
CL	38.46 ± 0.70	54.93 ± 0.63	$\underline{58.85}$ ± 0.43	$\underline{63.04}$ ± 0.34	$\underline{68.65}$ ± 0.24
DA Mask + CL	36.84 ± 0.86	53.15 ± 0.45	57.36 ± 0.32	62.37 ± 0.56	$\underline{68.62}$ ± 0.23

Table 1: Mean F1 scores over five runs for every data augmentation/contrastive training and dataset size combination. Underlined results show statistically significant increases over the baseline (Student's t-test, $\alpha = 5\%$).

struggles to learn new generalisable information from the examples, and this is reflected in the lack of improvement in the results.

Round-trip translation performs slightly better, but the improvements are also not statistically significant. Both paraphrasing and round-trip translation generate augmentations with tokens that are not NEs as we apply string matching between the NEs in the original data and the augmented examples to ensure that the labels are still valid. This means that our augmentations provide the model with different contexts for known NEs but do not actually show the model new NEs. The lack of improvement raises the question whether a more successful approach would present the model with augmented data that includes new NEs. This, however, is difficult to perform automatically without the risk of changing the NE type.

Masking, on the other hand, can be applied to both NEs and context tokens. However, the results are mixed and do not allow us to draw reliable conclusions. While we see statistically significant improvements for the CoNLL data on the full dataset and on a sample of 1000 sentences, no significant improvements were obtained on the other sample sizes or on the OntoNotes and FewNERD data.

A possible explanation could be that while masking reduces the chances of overfitting, it also increases the difficulty of the task as the model now needs to guess the NE of the masked tokens. Therefore, the technique might be better suited to easier problems with a high risk of overfitting, such as datasets with fewer NE types like CoNLL with its four coarse NE classes.

Contrastive Learning CL shows the most consistent results. At dataset sizes of above 5,000, we see statistically significant improvements for all three datasets. While data augmentation methods tend to work better on smaller datasets, our results show that contrastive learning needs more data to be beneficial. Instead of providing the model with new instances, contrastive learning improves the representations produced by the model. To learn robust and generalisable representations, large datasets are necessary to avoid overfitting.

Combining DA and CL As both approaches seem complementary, we also test the combination of DA and CL, using masking for data augmentation (Table 1, Mask + CL). While the model occasionally produces statistically significant results, the improvements are rather small. This does not necessarily mean that combining contrastive learning with data augmentation does not work in general.

Dataset Size (Sentences)	100	500	1000	5000	Full
CoNLL-2003 (4 NE types)	Mean F1 Score% ± STDEV				
Baseline	83.32 ± 0.36	88.23 ± 0.60	89.84 ± 0.46	91.15 ± 0.23	91.98 ± 0.43
CL only	82.52 ± 0.77	88.86 ± 0.49	90.06 ± 0.26	91.53 ± 0.30	92.49 ± 0.29
CTL + OntoNotes	83.75 ± 1.23	87.91 ± 0.44	89.43 ± 0.21	91.23 ± 0.14	92.19 ± 0.32
CTL + Few-NERD (coarse)	85.46 ± 0.39	89.23 ± 0.39	90.16 ± 0.20	91.39 ± 0.21	92.35 ± 0.36
CTL + Few-NERD (fine)	85.26 ± 0.65	89.02 ± 0.66	90.67 ± 0.21	91.68 ± 0.25	92.15 ± 0.28
OntoNotes v5 (18 NE types)	Mean F1 Score% ± STDEV				
Baseline	66.01 ± 1.58	77.90 ± 0.55	82.21 ± 0.32	85.58 ± 0.42	89.28 ± 0.25
CL only	65.82 ± 0.79	78.71 ± 0.33	82.42 ± 0.32	86.51 ± 0.46	89.76 ± 0.25
CTL + CoNLL	65.73 ± 1.67	78.58 ± 0.42	82.39 ± 0.81	85.71 ± 0.27	89.28 ± 0.23
CTL + Few-NERD (coarse)	68.47 ± 2.44	79.64 ± 0.16	83.07 ± 0.01	86.14 ± 0.44	88.87 ± 0.29
CTL + Few-NERD (fine)	67.55 ± 0.88	79.66 ± 0.63	83.17 ± 0.22	86.23 ± 0.55	89.48 ± 0.27
Few-NERD (66 NE types)	Mean F1 Score% ± STDEV				
Baseline	38.77 ± 0.82	54.07 ± 0.89	58.17 ± 0.67	62.09 ± 0.41	67.90 ± 0.59
CL only	38.46 ± 0.70	54.93 ± 0.63	58.85 ± 0.43	63.04 ± 0.34	68.65 ± 0.24
CTL + CoNLL	36.34 ± 0.61	54.45 ± 0.30	58.43 ± 0.41	62.45 ± 0.20	68.26 ± 0.28
CTL + OntoNotes	37.79 ± 1.58	54.79 ± 0.67	58.32 ± 0.34	62.57 ± 0.24	68.33 ± 0.21

Table 2: Mean F1 scores over five runs with and without contrastive training for different dataset sizes. The underlined results are statistically significant increases over the baseline ($\alpha = 5\%$). Few-NERD (coarse) uses the 8 coarse-grained labels, Few-NERD (fine) refers to the 66 fine-grained NE types.

More work is needed to explore data augmentations for NER to answer that question.

Contrastive Transfer Learning Table 2 shows results for CTL for all possible dataset combinations. We observe two clear trends. First, CTL works better when the *target data* is small. This is not surprising, given that there is more room for improvement when the baseline is low. The second observation is that CL needs sufficiently large *source data* to work well. This also makes sense as a larger transfer learning dataset allows the model to learn more useful representations of the data for the downstream task.

To investigate the impact of the *number of entity types* in the contrastive training set, we report results for two different settings. In the first setting, we use the eight coarse-grained NE types in Few-NERD that have some overlap with the entity inventory in CoNLL and OntoNotes,[3] the second setting includes Few-NERD's 66 fine-grained NE types.

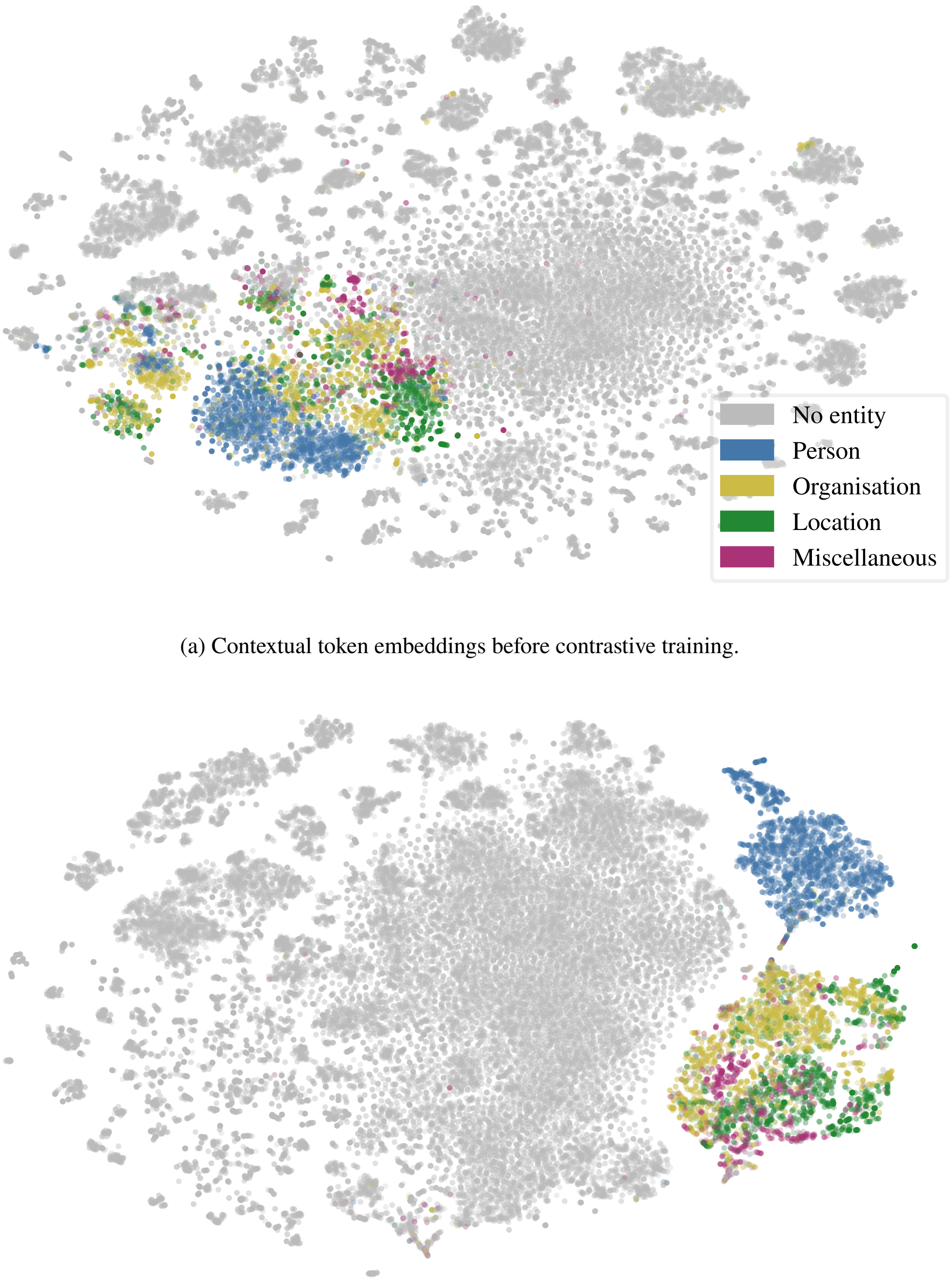
Results show that the coarse-grained entity labels only yield statistically significant improvements when the target training data is small (500 or less sentences) but fail to improve results for larger fine-tuning datasets with 1,000 or more sentences. This indicates that CL has learned more useful representations from the fine-grained information in the transfer data which is somewhat surprising, given that the coarse-grained entity types overlap with the labels in the respective target data.

5 Analysis and Visualisation

To better understand the effect of CL, we visualise the learned representations before and after the CL step. As we cannot directly plot the 768-dimensional word embeddings produced by RoBERTa on a two-dimensional graph, we apply dimensionality reduction techniques in order to obtain informative two-dimensional representations.

A popular dimensionality reduction technique is principal component analysis (PCA), which tries to reduce the dimensionality by choosing the linear combination of variables that explain the variance in the data (Jolliffe and Cadima, 2016). While PCA is quite efficient, it can only be applied when all components are linear. A method that can perform non-linear dimensionality reduction is t-SNE (van der Maaten and Hinton, 2008). However, t-SNE still has a high computational cost compared to PCA, especially when dealing with large datasets of high dimensionality. To resolve this problem, we use a combination of PCA and t-SNE for dimensionality reduction. We first apply PCA to reduce the dimensionality of the word embeddings from 768 to 50. Then, t-SNE is used to reduce the dimensions from 50 to two (see appendix B for more details).

[3] The coarse-grained entity types are PERSON, LOCATION, ORGANIZATION, ART, BUILDING, PRODUCT, EVENT, MISCELLANEOUS.



(a) Contextual token embeddings before contrastive training.

(b) Contextual token embeddings after contrastive training.

Figure 1: Visualisation of CoNLL 2003 token embeddings using a combination of PCA and t-SNE for dimensionality reduction.

Figure 1 shows the visualisations of the embeddings for the CoNLL dataset, using CL only. Results for OntoNotes and FewNERD are similar, and can be found in appendix C. For all three datasets, the separation between non-entities and NEs is greater than the separation of the representations for neighbouring NE classes. While a possible reason for this could simply be that the difference between NEs and non-entities is greater and therefore easier to learn, a more likely reason is the distribution of NEs and non-entities in the data where the latter significantly outnumber the former. In CL, this means that the model can minimise the loss by increasing the difference between the non-entities and NEs even if this comes at the expense of decreasing the difference between two different NE classes.

Hence, the lack of separation between the different NE classes can most probably be explained by the class imbalance in the data.

6 Discussion

While we found no evidence that the proposed data augmentations are effective, related work has shown that DA can be beneficial in low-resource scenarios (Dai and Adel, 2020; Ding et al., 2020; Cai et al., 2023). We also observed consistent increases in results for CL for datasets with sizes of at least 5,000 sentences. Our best results for a RoBERTa-base model with CL on OntoNotes (89.75% F1) are only slightly below the ones reported for much larger models (cf., 89.76% F1 for BART-large (Yan et al., 2021) and 90.42% F1 for a T5-base model with DA (Zhang et al., 2022)). These results are promising, given the severe lack of methods for improving the performance on larger datasets, as DA has only been successful when applied in low-resource and few-shot scenarios (Dai and Adel, 2020; Ding et al., 2020; Zhou et al., 2022; Cai et al., 2023), and the same also applies to work on contrastve learning for NER.

Our experiments failed to show that CL works for smaller datasets. However, when combined with transfer learning, the results are improved. CTL works best when the fine-tuning data size is small, making it a good complement to CL without transfer learning. Figure 2 summarises our results, showing which method might work best in different scenarios.

While the improvements we obtained are small, they are still important given that increasing the performance of a model that is already performing quite well tends to be much harder than improving the performance of a poorly performing model. In addition, data augmentation and CL can be combined, as often done in related fields like computer vision (Chen et al., 2020; He et al., 2020). This might be a promising avenue for future work on developing CL methods that work well for smaller datasets. Our experiments demonstrate that combining DA and CL is possible (see Mask + CL in Table 1) but might require more sophisticated data augmentation techniques to improve results.

Addressing Data Imbalance in CL for NER In section 5, we showed that a major problem for applying CL to NER is the data imbalance as the majority of the token labels are non-entities. One approach to address this problem could include a modifcation of the CL loss function to account for the imbalance (Cao et al., 2019; Fernando and Tsokos, 2022; Wang et al., 2020; Rezaei-Dastjerdehei et al., 2020). Assuming that equation (4) is used for the loss function, a modified loss function that includes weights is shown in equation (5), where $w_i \in \mathbb{R}^+$ is the weight for class i.

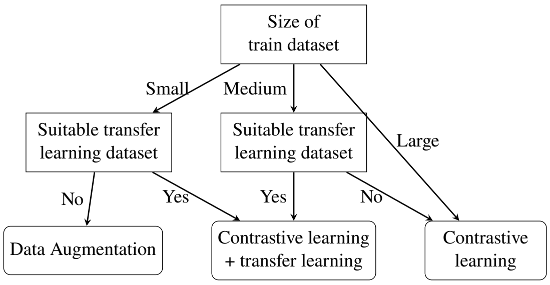

Size of train dataset
Small
Medium
Large
Suitable transfer learning dataset
Suitable transfer learning dataset
No
Yes
Yes
No
Data Augmentation
Contrastive learning + transfer learning
Contrastive learning

Figure 2: Recommendations for selecting the best approach for different-sized datasets.

$$\mathcal{L} = -\sum_{i \in I} w_i \sum_{j \in J} p_{i,j} \log q_{i,j} \tag{5}$$

There are many ways to set w_i, but one possibility is to set it to the ratio of the frequency of non-entities to the frequency of the class. This is shown in equation (6), where n_i is the frequency of class i and n_O is the frequency of the non-entity class.

$$w_i = \frac{n_O}{n_i} \tag{6}$$

This scales the loss function so that different classes can have different weights which might help encourage the model to differentiate between various types of NEs. There are many alternatives for the loss and weight functions, and the functions proposed above might not be optimal. Development and testing of a weighted loss function will be left to future work.

7 Conclusion

We presented a systematic investigation of the effect of DA, CL, and CTL for NER. Our main results can be summarised as follows. First, while DA has been shown to be effective in low-resource scenarios (specifically for pre-transformer-based taggers), we failed to demostrate an improvement in results in our experiments. CL, on the other hand, can effectively improve results over a strong RoBERTa baseline when medium to large datasets are available for fine-tuning, but has a weaker performance on smaller datasets. For small dataset sizes, contrastive transfer learning is the most promising approach but requires the existence of suitable data for transfer learning.

We hope that the insights from our experiments will foster more work on DA and CL for NER especially for medium and large datasets. To address the problem of data imbalance for NER, where the majority of the labels are non-NEs, we proposed a modification to the loss function, which we plan to explore in future work.

Acknowledgments

The work presented in this paper was funded by the German Research Foundation (DFG) under the UNCOVER project (RE3536/3-1).

References

Jiong Cai, Shen Huang, Yong Jiang, Zeqi Tan, Pengjun Xie, and Kewei Tu. 2023. Graph Propagation based Data Augmentation for Named Entity Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 110–118, Toronto, Canada. Association for Computational Linguistics.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 1597–1607. JMLR.org.

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.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A Few-shot Named Entity Recognition Dataset. 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 3198–3213, Online. Association for Computational Linguistics.

K. Ruwani M. Fernando and Chris P. Tsokos. 2022. Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951.

Kai He, Rui Mao, Yucheng Huang, Tieliang Gong, Chen Li, and Erik Cambria. 2023. Template-Free Prompting for Few-Shot Named Entity Recognition via Semantic-Enhanced Contrastive Learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. COPNER: Contrastive Learning with Prompt Guiding for Few-shot Named Entity Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First Quora Dataset Release: Question Pairs. `https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs`.

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis: A review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 18661–18673, Red Hook, NY, USA. Curran Associates Inc.

Pranjal Kumar, Piyush Rawat, and Siddhartha Chauhan. 2022. Contrastive self-supervised learning: Review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval*, 11(4):461–488.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022a. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022b. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Wei Li, Hui Li, Jingguo Ge, Lei Zhang, Liangxiong Li, and Bingzhen Wu. 2023. CDANER: Contrastive Learning with Cross-domain Attention for Few-shot Named Entity Recognition. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Gold Coast, Australia. IEEE.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2023. Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876.

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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803.

Luis Perez and Jason Wang. 2017. The Effectiveness of Data Augmentation in Image Classification using Deep Learning.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):140:5485–140:5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Mohammad Reza Rezaei-Dastjerdehei, Amirmohammad Mijani, and Emad Fatemizadeh. 2020. Addressing Imbalance in Multi-Label Classification Using Weighted Cross Entropy Loss Function. In *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, pages 333–338.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation.

Shuzheng Si, Shuang Zeng, Jiaxing Lin, and Baobao Chang. 2022. SCL-RAI: Span-based Contrastive Learning with Retrieval Augmented Inference for Unlabeled Entity Problem in NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2313–2318, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. 2023. StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.

In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. 2021. Scaling Local Self-Attention for Parameter Efficient Visual Backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904.

Vladimir Vorobev and Maxim Kuznetsov. 2023a. ChatGPT paraphrases dataset. `https://huggingface.co/datasets/humarin/chatgpt-paraphrases`.

Vladimir Vorobev and Maxim Kuznetsov. 2023b. A paraphrasing model based on ChatGPT paraphrases. `https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base`.

Chen Wang, Chengyuan Deng, and Suzhen Wang. 2020. Imbalance-XGBoost: Leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, 136:190–197.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

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.

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.

Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. 2022. De-Bias for Generative Extraction in Unified NER Task. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 808–818, Dublin, Ireland. Association for Computational Linguistics.

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.

Appendices

A Details for Data Augmentation

Table 3 shows how the different DA techniques affect the size of the training data in our experiments. Please note that the dataset size for Masking and CL always remains constant.

A.1 Consistency Checks for Round-Trip Translation

We check the round-trip translated output by string matching every named entity in the original sample to the augmented sample. If all named entities are found, then the entities are labelled based on the assumption that all string matches represent the same named entity, and all other words are not named entities. The neural machine translation model chosen is No Language Left Behind (NLLB) (NLLB Team et al., 2022) and we use round-trip translation to/from German. We also experimented with French and Zulu, with very similar results.

A.2 Paraphrasing

The model used for paraphrasing is T5 (Raffel et al., 2020). To generate the augmented sentence, "paraphrase: " is prepended to each original example and given to the T5 model as input. The model has been fine-tuned by Vorobev and Kuznetsov (Vorobev and Kuznetsov, 2023b) on the ChatGPT paraphrases dataset (Vorobev and Kuznetsov, 2023a), which uses the Quora Question Pairs (QQP) dataset (Iyer et al., 2017), Stanford Question Answering Dataset (SQuAD) version 2.0 (Rajpurkar et al., 2018) and the CNN / DailyMail Dataset (Hermann et al., 2015). ChatGPT was used to create five paraphrases for each example in the three datasets to train the T5 model.

A.3 Masking

The masking rate is selected based on the design of BERT, which uses the same masking rate for its mask language modelling training. However, this masking method is not exactly the same as that performed by BERT, which only replaces the

Dataset Size		Original	100	500	1000	5000	Full
			Round-Trip Translation via German				
CoNLL-2003	(4 NE types)	14,041	158	785	1,587	7,966	22,348
OntoNotes	(18 NE types)	82,122	167	872	1,714	8,638	141,314
Few-NERD	(66 NE types)	131,767	165	837	1,689	8,394	219,969
			Paraphrasing				
CoNLL-2003	(4 NE types)	14,041	177	898	1,801	9,051	25,310
OntoNotes	(18 NE types)	82,122	177	896	1,782	8,911	146,041
Few-NERD	(66 NE types)	131,767	173	909	1,791	9,056	238,500

Table 3: Number of augmented instances used for training for the different DA techniques (round-trip translation, paraphrasing) and dataset sizes (100, 500, 1,000, 5,000, full dataset).

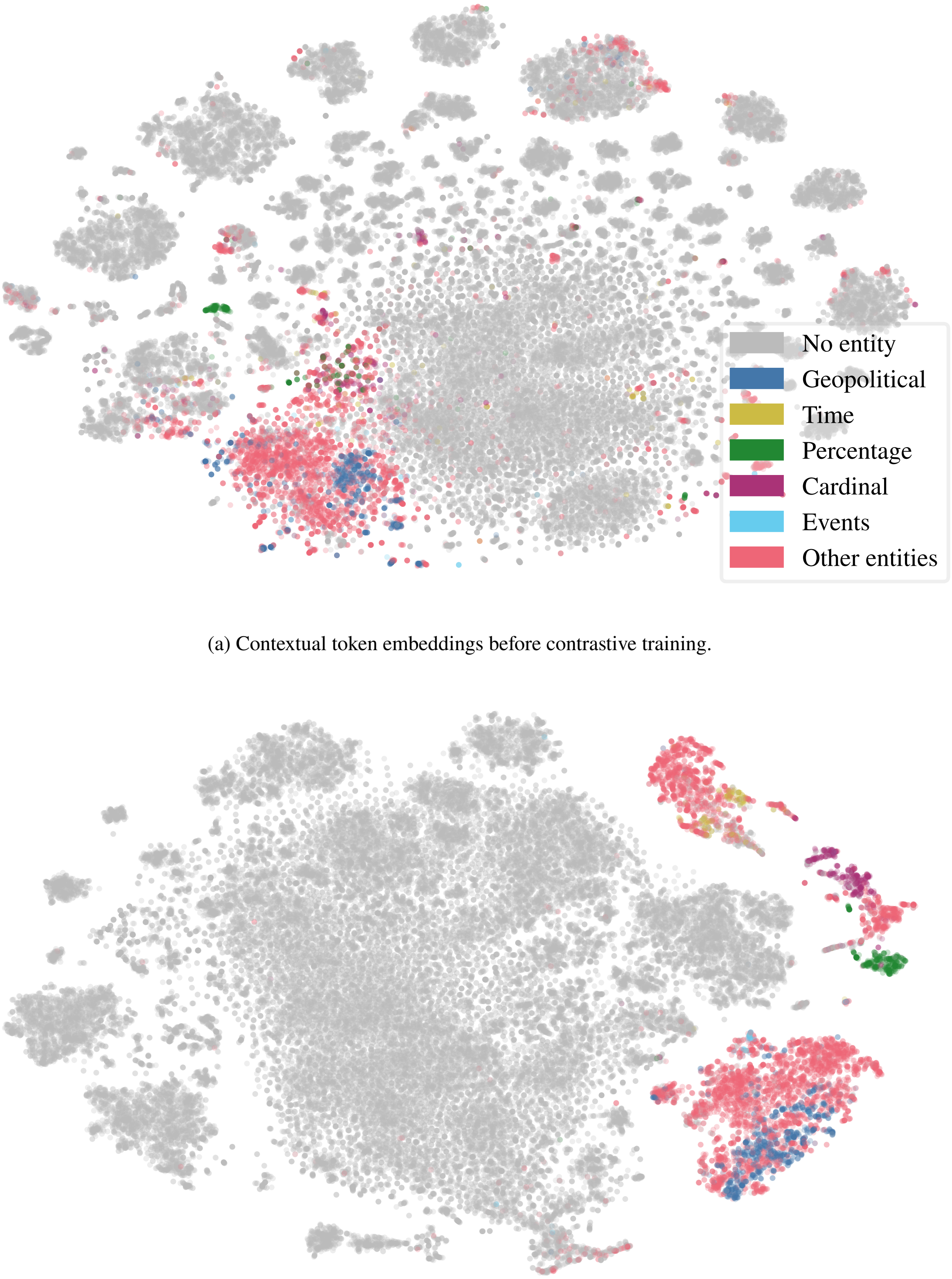
chosen token with `[MASK]` 80% of the time. There is a 10% chance the token will be replaced by a random token and a remaining 10% chance the token will remain unchanged. This is not done because the initial tests show that replacing with the `[MASK]` token is already a complicated enough task, and the addition of random tokens might cause the model to perform slightly worse.

B Visualising Word Embeddings with PCA and t-SNE

A problem faced when creating the scatter plots after dimensionality reduction is that every word in the test set becomes a point on the plot, so there is a huge number of points found on the plot. This causes the points in the plot to overlap and block each other, making the plot difficult to read. Increasing the transparency of the points and making them slightly smaller was sufficient to make the CoNLL 2003 plot readable. However, a random sample of 50,000 points needed to be taken from the OntoNotes v5 and Few-NERD test set, because these sets were much bigger. The sampling was only done right before plotting to avoid any information loss when performing PCA or t-SNE.

C Visualisations for OntoNotes and FewNERD

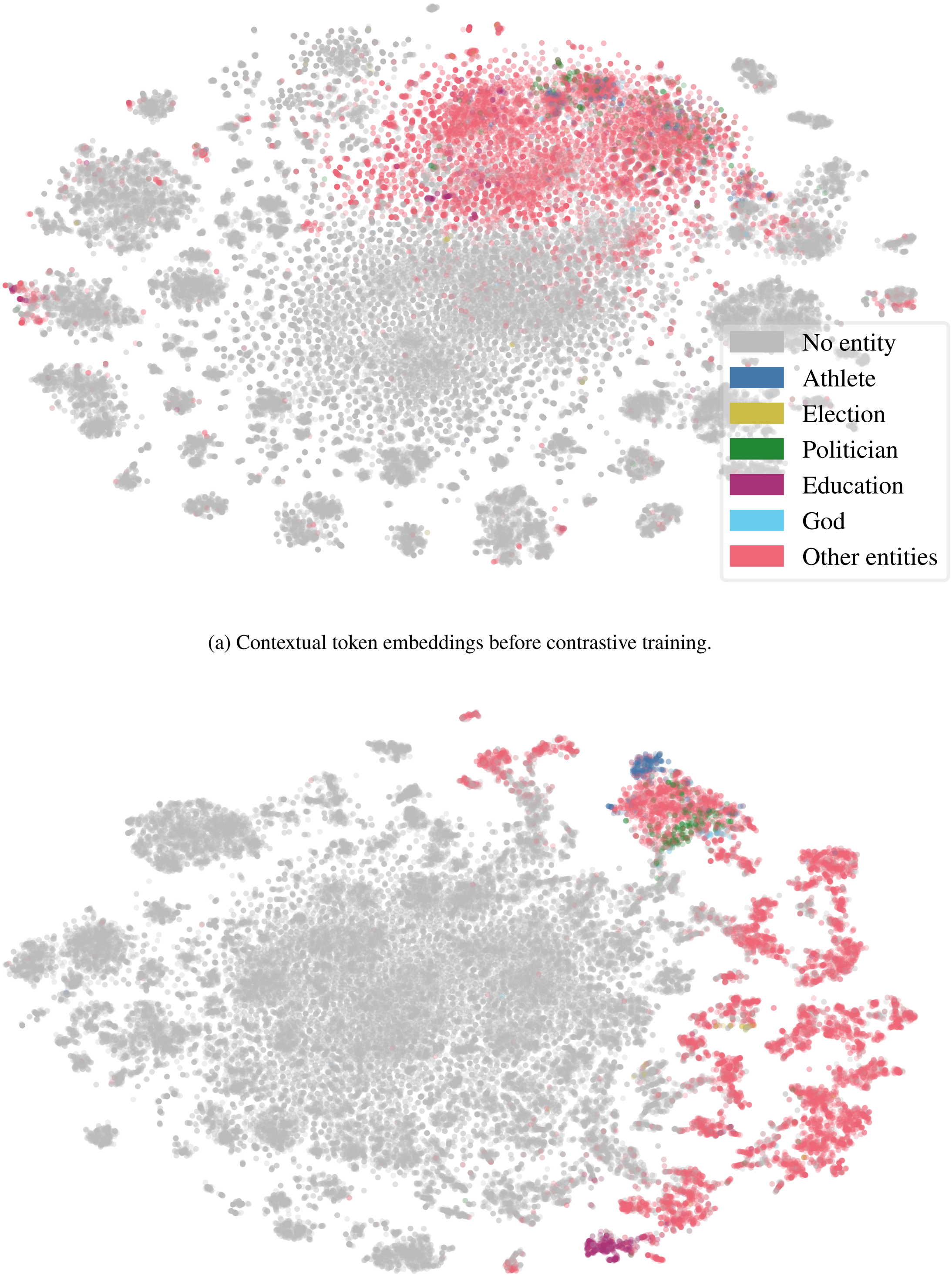
The OntoNotes v5 and Few-NERD datasets contain 18 and 66 entity classes respectively. This makes it impossible to find different colours that have good contrast for every entity class, and on the scatter plot, it is difficult to tell so many classes apart. To solve this problem, only five named entity types will have a unique colour, and the rest are grouped together as “other entities”. One of the five selected has a good F1 score after supervised fine-tuning, and another has very poor scores. These two classes ensure that the best and worst-case scenarios are shown in the plot. The remaining three entity types are randomly selected to give a more representative picture of the rest of the entity types.



(a) Contextual token embeddings before contrastive training.

(b) Contextual token embeddings after contrastive training.

Figure 3: OntoNotes v5 token embeddings using a combination of PCA and t-SNE for dimensionality reduction.



(a) Contextual token embeddings before contrastive training.

(b) Contextual token embeddings after contrastive training.

Figure 4: Few-NERD token embeddings using a combination of PCA and t-SNE for dimensionality reduction.