ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240805

Prompt-Based Data Augmentation Using Contrastive Learning Under Scarcity of Annotated Data

Muhammad Uzair Ul Haq[a,b,c], **Davide Rigoni**[d] **and Alessandro Sperduti**[a,b,e,f]

[a]Department of Mathematics "Tullio Levi-Civita", University of Padova, Via Trieste 63, 35121 Padova, Italy
[b]Human Inspired Technology Research Centre, University of Padova, Via Luzzatti 4, 35122 Padova, Italy
[c]Amajor SB S.p.A, Via Noventana 192, 35027 Noventa Padovana, Italy
[d]Molecular Modelling Section (MMS), Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Via Marzolo 5, 35131, Padova, Italy
[e]Augmented Intelligence Center, Bruno Kessler Foundation, Via Sommarive 18, 38123 Povo, Italy
[f]Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 9, 38123 Povo, TN, Italy

Abstract. Named Entity Recognition is a crucial task in Natural Language Processing (NLP) which aims to identify the entities in text. Given an adequate amount of annotated data, Large Language Models (LLMs) have been shown to be effective in this task when fine-tuned. However, the performance of LLMs is severely affected when annotated datasets are limited. To alleviate this problem, adding synthetic data via Data Augmentation (DA) techniques is a viable approach. Even so, DA for token-level tasks suffers from two main limitations: *(i)* token-label misalignment problem; and *(ii)* quality of generated synthetic data. In this paper, we propose a novel prompt-based DA approach using contrastive learning. The proposed method can generate high-quality synthetic data while preserving the token-label correspondences. Experimental results demonstrate that the proposed approach, when compared against multiple baselines on well-known Named Entity Recognition (NER) datasets, achieves State-of-the-Art performance.

1 Introduction

Named Entity Recognition (NER) is a crucial Natural Language Processing (NLP) task when it comes to extract fine-grained information from text. NER involves identifying and categorizing entities within text. These entities play a pivotal role in tasks such as information retrieval. Large Language Models (LLMs) have shown promising performances across various NLP tasks such as text classification [34], document summarization [35], question answering [25]. LLMs can be fine-tuned to domain-specific data for any NLP task. However, to reach an acceptable level of performance, an adequate amount of data is required for fine-tuning.

Manual data annotation is frequently used by practitioners to label large datasets. However, annotating manually a sufficient amount of data is time-consuming, labour-intensive, and resource-demanding. The problem is even worse for fined-grained tasks like NER where each token in a document has to be tagged [20]. Data Augmentation (DA) is a popular approach in NLP to automatically generate synthetic data for training. In recent years, DA has received increasing attention in the research community due to the availability of LLMs and significant interest in low-resource domains [3].

There are various DA techniques in NLP. The classical approaches, such as Easy Data Augmentation (EDA) [32], alter the training instances but do not bring any diversity to the training regime. To augment the training instances with diverse examples, researchers have made use of vocabulary-based approaches such as WordNet [24], where the entity mentions are replaced by randomly sampled similar words from the WordNet vocabulary [24]. For instance, the name of the person "John" could be replaced by the words "Mark", "Wills", etc... Vocabulary-based approaches can bring diverse examples in the training regime but these replacements do not take any contextual information into account. Moreover, for simpler entities, such as the name of a person or a location, it is easy to find a vocabulary or list of entities on the web. On the contrary, for complex entities, such as soft skills and creative work, it is difficult to scrap information from the web. Apart from classical approaches, back-translation and text generation [9] are widely used in NLP for DA. However, these approaches are only viable for text classification tasks where the annotation is only at the sentence level. An example of such a case is the sentiment classification task, in which each sentence could be labelled with sentiment as "happy", "sad" or "afraid". In such scenarios, preserving the golden annotation is straightforward, as token-level label correspondence does not need to be maintained. Thus, such tasks can easily leverage the DA techniques mentioned above. However, for label-sensitive tasks, such as NER, each token of a document or sentence is tagged. Any manipulation of the input sequence might misalign the corresponding label. Therefore, preserving the gold labels becomes a critical task in DA for NER problems.

In this work, we propose a novel contrastive learning-based model for DA, dubbed Contrastive Prompt Data Augmentation (CPDA), which fine-tune prompts via contrastive learning to generate high-quality synthetic data and, thus, enhance the performance of the NER task. CPDA leverages prior information available in each sentence to guide the DA toward those sentences that better align with it. By using prior information (i.e., entity category) in the form of a prompt for each sentence, we fine-tune the LLM using contrastive learning. This allows the model to bring entities close to their prompt in the embedding space. This allows the model to generate diverse entities belonging to the same entity category, thereby, reducing the probability of sampling from different entity categories. CPDA follows a three-step process, first, it extracts the entity category (label) from the training data and prepends a prompt to each sentence. Secondly, it fine-tunes the LLM using a contrastive approach to adapt prompts to the label space, optimizing the Masked Language Modeling (MLM) objective. Lastly, the model generates new training instances, taking into account both label and contextual information.

Our main contributions are summarized in the following:

1. we introduce a Contrastive Prompt Data Augmentation (CPDA)[1] , a novel framework that fine-tunes prompts using contrastive learning for generating augmented training data;
2. the proposed approach uses a prompt-based label-guided augmentation approach that leverages entity category information;
3. we employ a contrastive learning objective to fine-tune prompts; to the best of our knowledge, this is the first attempt that utilizes contrastive learning to fine-tune prompts for Data Augmentation;
4. comprehensive experiments conducted on three token-classifications datasets show that the proposed method achieves state-of-the-art performance under lack of annotated data.

2 Related Work

Recently, there have been many efforts to explore LLMs for DA while preserving the token-label misalignment problem. Dai et al. [5] proposed to randomly substitute entities with others entities of the same category in the dataset. They avoided the token-label misalignment issue but the entity diversity did not increase. Moreover, the substituted entity might not be suitable for the original context. Ding et al. [8] proposed using DA as a conditional generation task, generating new sentences while preserving the original targets and labels. Their approach relies upon linearized labelled sequences. During linearization, the entity labels are explicitly inserted in the sequence. This approach is controllable and allows for more diversified sentence generation. Zhou et al. [37] suggested the use of Masked Entity Language Modelling (MELM) as a DA framework for low-resource NER, which addresses the token-label misalignment issue by injecting NER labels explicitly into a sentence. This enables the fine-tuned MELM to predict masked entity tokens while explicitly conditioning their label. Such technique solves the token-label misalignment problem by injecting the label information explicitly into the model. These methods require post-processing to remove noisy samples from the augmented data.

Most recently, motivated by the GPT-4 [26], prompt-based fine-tuning is becoming a popular approach to improve the performance of LLMs for various tasks, such as text classification [10], question answering [25], and language generation [19], etc. The prompt-based learning has shown promising performances, especially when the annotated data is scarce [21]. Chen et al. [1] proposed the use of prompt-based DA for low-resource Natural Language Understanding (NLU) tasks. The authors show promising results, however, their approach is only limited to sentence-level tasks, such as text classification. Even though prompt-based approaches are quite recent, selecting an appropriate prompt for each task includes manual effort. Additionally, most recent works exploit soft prompt templates which add to model complexity and additional computational power [21].

Contrastive learning has shown promising results in Computer Vision (CV) [2, 16, 29, 28, 27], graph representations [18], and NLP [11, 15]. Contrastive representation learning aims to project similar samples closer in the embedding space while pushing apart the dissimilar samples. The same idea has been extensively applied for solving the NER Task [6]. Das et al. [6] proposed the use of contrastive learning to improve the performance of the NER task in a Few-shot setting, optimizing the inter-token distribution distance. They optimize a generalized objective of differentiating between token categories based on their Gaussian-distributed embeddings. Huang et al. [13] used contrastive learning in conjunction with prompt guiding to solve the Few-shot NER problem. The author proposed to use a prompt composed of category-specific tokens, which provides supervision signals for conducting contrastive learning to optimize token representations.

3 Method

This section defines the task addressed in this work, as well as our proposed approach. The following mathematical notation is adopted: (*i*) lower case symbols for scalars, indexes, and assignment to random variables, e.g., n and x; (*ii*) italics upper case symbols for sentences, sets, and random variables, e.g., A and X; (*iii*) bold lower case symbols for vectors, e.g., $\boldsymbol{a}$; (*iv*) bold upper case symbols for matrices and tensors, e.g., $\boldsymbol{A}$; (*v*) the position within a tensor or vector is denoted by numeric subscripts in square brackets, e.g., $\boldsymbol{A}_{[i,h,k]}$; (*vi*) calligraphic symbols for domains, e.g., $\mathcal{Q}$.

3.1 Problem Definition

Our proposed approach performs data augmentation on the input sentences. In practice, given a sentence in input, it aims to generate a different version of it where the same semantic information remains unaltered. Formally, let $D = \{(S_i, GT_i)\}_{i=1}^{d}$ be the dataset composed of d examples, where each sentence S_i is associated to its gold label annotation GT_i. The sentences are defined as $S_i = \{w_j^i\}_{j=1}^{\phi(S_i)}$, where $w_j^i \in \mathcal{V}$ is the j-th word appearing in the i-th sentence and belonging to vocabulary $\mathcal{V}$. The function $\phi(S_i)$ returns the number of words in sentence S_i. The gold label annotations are defined as $GT_i = \{c_j^i\}_{j=1}^{\phi(S_i)}$, where $c_j^i \in \mathcal{C}$ is the gold label for the word w_j^i, among the pre-defined categories in $\mathcal{C}$.

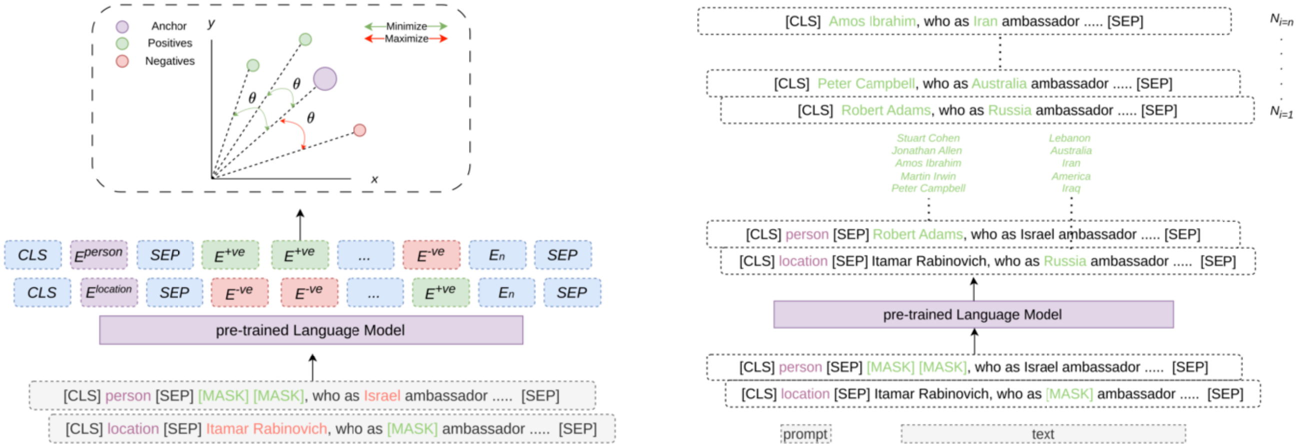
The NER task can be seen as the task of learning the function $\widetilde{f}_\theta$, that approximates the real function $f : \mathcal{S}_\mathcal{V} \times \mathcal{V} \rightarrow \mathcal{C}$, which, given a sentence $S \in \mathcal{S}_\mathcal{V}$ on the vocabulary $\mathcal{V}$, associates each word $w \in S$ to its category $c \in \mathcal{C}$. In this context, given a sentence S_i in input, our proposed DA approach aims to generate one (or more) new sentence(s) $\hat{S}_i = \{\hat{w}_j^i\}_{j=1}^{\phi(S_i)}$, where:

$$\hat{S}_i \neq S_i \quad \text{s.t.} \quad f(\hat{S}_i, \hat{w}_j^i) = c_j^i, \quad \forall j \in [1, \ldots, \phi(S_i)].$$

3.2 Our Approach: Contrastive Prompt Data Augmentation

This section presents our Contrastive Prompt Data Augmentation approach, whose aim is to condition the DA task with prior information. This information is extracted from a given sentence S, subject to DA, and it is represented as a prompt that guides the DA task towards those sentences that better align with the prior information. The prompt can direct the generation to only some type of entities, or it can introduce more information about the kind of entities' instances to generate. For example, the prompt can be represented by the entity that needs to be augmented, such as `person` for the name of a person or `location` for a location. In some cases, such as in the case of soft skills, the prompt may specify available information

[1] Source code available at https://github.com/UzairUlhaq/cpda.git


Anchor
Positives
Negatives
y
x
θ
θ
θ
Minimize
Maximize
CLS
E^person
SEP
E^+ve
E^+ve
...
E^-ve
E_n
SEP
CLS
E^location
SEP
E^-ve
E^-ve
...
E^+ve
E_n
SEP
pre-trained Language Model
[CLS] person [SEP] [MASK] [MASK], who as Israel ambassador ..... [SEP]
[CLS] location [SEP] Itamar Rabinovich, who as [MASK] ambassador ..... [SEP]
[CLS] Amos Ibrahim, who as Iran ambassador ..... [SEP]
N_{i=n}
[CLS] Peter Campbell, who as Australia ambassador ..... [SEP]
[CLS] Robert Adams, who as Russia ambassador ..... [SEP]
N_{i=1}
Stuart Cohen
Jonathan Allen
Amos Ibrahim
Martin Irwin
Peter Campbell
Lebanon
Australia
Iran
America
Iraq
[CLS] person [SEP] Robert Adams, who as Israel ambassador ..... [SEP]
[CLS] location [SEP] Itamar Rabinovich, who as Russia ambassador ..... [SEP]
pre-trained Language Model
[CLS] person [SEP] [MASK] [MASK], who as Israel ambassador ..... [SEP]
[CLS] location [SEP] Itamar Rabinovich, who as [MASK] ambassador ..... [SEP]
prompt
text

(a) Using the prompt to adapt the target label space.

(b) Data generation conditioned by prompt and context.

Figure 1: Example of application of the proposed CPDA approach in the case in which the prompt corresponds to one of the entities of interest (Anchor). (a) Training of MLM module using contrastive learning: LLM is fine-tuned to maximize the cosine similarity (minimizing the θ) between prompt tokens (highlighted in purple) and positive tokens (highlighted in green) while minimizing the similarity (maximizing the θ) with negative tokens (highlighted in red). (b) Data generation module. The masked training instances are given as input to LLM along with a prompt to generate new sentences: the model predicts the masked token while conditioning on the prompt.

at sentence level on the kind of soft skill to generate, such as "technical" or "problem-solving" soft skills. This approach ensures that the augmented outputs maintain coherence with the original context while augmenting it with desired characteristics, thereby enhancing the overall effectiveness of the data augmentation process.

An example of our model workflow, in the case where the prompt corresponds to a specific entity, is given in Figure 1, where CPDA exploits contrastive learning and prompt guiding to adapt the target label space. Figure 1(a) represents the training of MLM module using contrastive learning, while Figure 1(b) refers to the data generation module.

To train a LLM for a masked language modeling task, initially, each word in a sentence must be tokenized [33]. Then, some tokens are masked and the LLM model is asked to predict the correct masked word given the context of the sentence, i.e., the unmasked words. The aim of the approach is to generate a different word/token, in correspondence of masked tokens, so to preserve the corresponding original categories. We do this by using a prompt that should help the generation process to exploit information coming either from a specific category or from general information about the input sentence. For example, in the first case, given an instance (S_i, GT_i) and a category $c \in \mathcal{C}$, our approach masks the words w^i_j where the corresponding label c^i_j is equal to c. Then it adds the token corresponding to c as a prompt to the model (Figure 1.(a)), aiming to predict new words $\hat{w}^i_j$ for each masked word w^i_j.

The category c may represent the entity of the words (like `person` of `location`) or, in the case of the soft-skill dataset, it may represent the kind of soft-skill (like `general` or `technical`) at sentence level.

Therefore, starting from a pre-trained LLM [22], and given an instance (S_i, GT_i) in input, with k masked tokens $M = \{m^i_z\}^k_{z=1}$, and a prompt $\pi(S_i)$ (defined as a function of S_i and accordingly to the above description), the model learns to maximize the probability of M conditioned on $\pi(S_i)$ and $\overline{S}_i$, i.e., the masked sentence of S_i. This can be expressed using the following equation:

$$\arg\max_{\theta} \sum_{S_i \in D, \pi(S_i)} \log \mathbb{P}_\theta \left(M|\pi(S_i), \overline{S}_i\right), \qquad (1)$$

where θ represents the parameters of the pre-trained LLM being fine-tuned, and $\mathbb{P}_\theta \left(M|\pi(S_i), \overline{S}_i\right)$ represents the probability of predicting the masked tokens M given the prompt π and LLM parameters θ.

The primary objective of the fine-tuning MLM process is to train the LLM to maximize the likelihood of predicting the correct tokens at the masked positions, conditioned on the provided prompt placeholders. This involves adjusting the LLM's parameters to effectively capture the contextual information and dependencies within the sequence, enabling accurate predictions.

To avoid the catastrophic forgetting of LLM [23], we fine-tune the MLM using two objective functions: cross-entropy loss and contrastive loss, where cross-entropy loss is given as:

$$L_{CE} = -\log \sum_{i=1}^{|\mathcal{E}|} y_i \log(\hat{y}_i), \qquad (2)$$

where y_i and $\hat{y}_i$ represent the correct and the predicted labels regarding the category index i in the domain of categories $\mathcal{E}$, respectively. While contrastive loss from [2] is adapted as:

$$L_S = -\log \frac{e^{sim(\boldsymbol{t}_i,\boldsymbol{t}_i^+)}}{\sum_{i^+\in B(i)} e^{sim(\boldsymbol{t}_i,\boldsymbol{t}_i^+)} + e^{sim(\boldsymbol{t}_i,\boldsymbol{t}_i^-)^2}}. \qquad (3)$$

where i corresponds to a single example in batch B, $\boldsymbol{t}_i^+$ and $\boldsymbol{t}_i^-$ represents the embedding of positive and negative tokens, $\boldsymbol{t}_i$ represents the embedding of anchor (prompt), and sim represent the cosine function. The square value forces the anchor and the negative term to be orthogonal, i.e., independent.

Given small instances of training data for a low-resource task, we argue that using only a contrastive objective leads to catastrophic forgetting of LLM objective thereby affecting the latent space adversely. To avoid such a phenomenon, we fine-tune the parameters of LLM by total loss, which is given by:

$$L_T = \lambda L_S + (1-\lambda) L_{CE} \qquad (4)$$

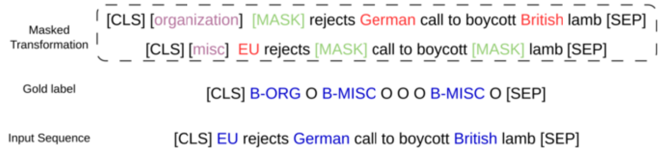

Masked Transformation
[CLS] [organization] [MASK] rejects German call to boycott British lamb [SEP]
[CLS] [misc] EU rejects [MASK] call to boycott [MASK] lamb [SEP]
Gold label
[CLS] B-ORG O B-MISC O O O B-MISC O [SEP]
Input Sequence
[CLS] EU rejects German call to boycott British lamb [SEP]

Figure 2: Proposed masking scheme. The input sequence consists of entity tokens (highlighted in blue) and context tokens (highlighted in black). We mask the tokens corresponding to entities iteratively and insert the category placeholder as a prompt at the start of the sequence. In the first iteration, we mask the tokens belonging to "misc" category, whereas in the second iteration, we mask the tokens belonging to "organization" category.

where λ is a hyper-parameter which is tuned during model selection. We explain our proposal for L_S in Section 3.2.1.

3.2.1 Prompt Tuning

Prompt tuning is widely applied to improve the performance of NER models during fine-tuning [17, 6]. Prompt-based fine-tuning can be broadly classified into two main categories, as soft-prompt tuning and hard-prompt tuning. Finding an appropriate prompt template is a challenging task for hard prompt tuning [4]. As a change in the prompt results in a different output [12]. On the other hand, soft prompt-based approaches rely on training additional model parameters [17] thereby increasing model complexity.

In this paper, without any loss of generality, we propose to use hard prompt-based approach. Given the problem definition in Section 1 and exploiting the label information from training data, we explicitly extract the category information from the training data and insert the category placeholder as prompt for each sequence. The proposed approach relieves us of any manual prompt crafting [21].

During the training process for the masked language modeling task, we begin by transforming all input examples in the training batch B using the prompt placeholder π. The placeholder π acts as a category indicator within the dataset D. For each example i in B, we then identify positive examples—those with labels similar to that specified in π—and negative examples, which have labels different from those in π. Let $\boldsymbol{t}_i$ denote the embedding of the predicted label for the masked token in example i, while $\boldsymbol{t}_i^+$ and $\boldsymbol{t}_i^-$ represent the embeddings of the positive and negative examples, respectively. To guide the model in distinguishing between similar and dissimilar labels, we employ a contrastive loss as defined in Equation (3). For each example i, the contrastive learning loss aims to learn the embedding of π by pulling semantically close examples with the same label together and making examples with a different label independent.

3.2.2 Data Generation

To generate the augmented instances of training samples, we use the fine-tuned LLM to generate new DA versions of them. Given a single training instance (S_i, GT_i) in input, with k masked tokens $M = \{m_z^i\}_{z=1}^k$ and prompt placeholders as π, the model predicts the most $K > k$ probable words for each masked token in M using Equation (1). To generate diverse entities, we randomly sample k words from the top K predictions returned by the model. After obtaining the generated sequence, we remove the prompt tokens and use the remaining parts as the augmented training data. For each sentence in the original training set, we repeat the above generation procedure N rounds to produce augmented examples.

4 Datasets and Experimental Setup

In the following section, we describe our experimental assessment of the CPDA approach. We used three NER datasets. For all the datasets used in this work, we compared our proposed CPDA approach versus the baselines obtained by using RoBERTa [22] base model, EDA [32], WordNet [24] and MELM [37].

4.1 Datasets

In the proposed study, we used three different token-classification datasets: two NER datasets CoNLL2003 [30] and WNUT-17 [7], along with SKILLSPAN [36] dataset. The CoNLL2003 dataset consists of four different entities: (*i*) $PERSON$; (*ii*) $ORGANIZATION$; (*iii*) $LOCATION$; and (*iv*) $MISCELLANEOUS$. The WNUT-17 [7] dataset consists of six different entities: (*i*) $PERSON$; (*ii*) $CORPORATION$; (*iii*) $LOCATION$; (*iv*) $CREATIVE - WORK$; (*v*) $GROUP$; and (*vi*) $PRODUCT$. This dataset consists of rare entities which are difficult to extract, and therefore, it is well suited to evaluate the proposed approach. Moreover, we made use of SKILLSPAN [36], a dataset containing soft skills, where only two types of entities are included to denote the presence or absence of a soft skill.

To simulate the low-resource scenario in our experiments, we adopted subsets of data from the original datasets. Given a training set size $d_{tr} \in \{100, 200, 300, 400, 500\}$, we sampled a training set $D_{train} = \{(S_i, GT_i)\}_{i=1}^{d_{tr}}$ and a validation set $D_{valid} = \{(S_i, GT_i)\}_{i=1}^{d_{val}}$ from the original splits of data. In our experiments, $d_{val} = d_{tr}$. The results reported on the validation set, for each size d_{tr}, are obtained by averaging the model's performance across three runs in which both D_{train} and D_{valid} are resampled. For a fair evaluation, the test set adopted in our experiments is the full test set $D_{test} = \{(S_i, GT_i)\}_{i=1}^{d_{te}}$ (i.e., no sampling is performed), and the results reported in Table 2 are obtained by averaging the model's performance across three independent runs on the test set.

4.2 Experimental Setup

All the experiments are conducted in the Python environment. We used the [14] transformers repository for the model implementation. All the experiments are conducted on NVIDIA RTX $A5000$ GPU.

MLM-MODEL For training MLM, we used the method described in Section 3. The proposed approach relies on extracting category information from sentence to use as a prompt. Since, the model is trained in a contrastive fashion therefore, sampling positive and negative instances is a crucial task. For multi-category datasets such as CoNLL-2003 and WNUT-17, we followed the approach explained in Section 3, which can also be visualized in Figure 1. However, for datasets consisting of a single category, such as SKILLSPAN [36], we do not possess any information on negative tokens. Therefore, for each sentence, we randomly sample span of tokens in the range 1 to 13 (the proposed numbers correspond to the minimum and maximum length of soft skill in the dataset) and treat them as negative examples. This modification allowed us to adapt CPDA to single category datasets. We used the RoBERTa [22] base model and utilized the Language Modeling Head (LMH) implementation from the HuggingFace [14] library to adapt the model for the MLM task.

NER-MODEL For fine-tuning, we used the approach described in the original work by Vaswani et al. [31]. As mentioned in Section 3.1, we formulated the problem as a NER task. To solve these tasks, an encoder-based model such as RoBERTa [22], is required.

Table 1: Number of generated augmented sentences N and F_1 score on the validation set for each dataset and each considered model/augmentation technique. The baseline value is for $N = 0$, i.e. no augmentation. The best performance for each model/data augmentation is in bold.

Dataset	**Dataset Size**	N					
		0	1	2	3	4	5
CoNLL-2003	100	80.24	**87.34**	86.02	85.97	85.15	84.24
	200	87.17	**87.78**	86.46	85.89	85.16	84.6
	300	87.37	**88.16**	86.78	86.41	85.85	85.61
	400	88.68	**88.95**	87.91	87.56	86.58	86.43
	500	88.61	**89.50**	88.41	86.99	86.69	86.24
WNUT-17	100	28.06	**42.07**	41.18	41.53	40.62	39.81
	200	45.57	**51.16**	50.91	50.01	49.78	49.17
	300	54.14	54.91	**55.61**	54.46	54.07	53.47
	400	54.80	**58.52**	55.89	56.88	56.00	54.94
	500	57.06	**59.00**	58.33	57.64	56.74	56.02
SKILLSPAN	100	19.32	19.73	**22.61**	20.00	20.63	20.37
	200	21.38	30.48	**30.77**	28.52	28.68	30.46
	300	25.80	31.82	**37.94**	34.90	33.81	33.46
	400	30.27	41.62	**42.13**	40.84	42.01	39.02
	500	32.98	45.53	**47.39**	44.92	44.66	44.66

In fact, decoder-based generative models, such as Llama [12], GPT-4 [26], etc., are more suitable for sequence-to-sequence tasks such as machine translation and text generation[2]. We utilized the RoBERTa base [22] model for our experiments with a linear classifier for the final classification of tokens.

Hyperparameter Tuning and Model Selection To determine the optimal setting for augmentation rate N, we conducted a grid search in $\{1, 2, 3, 4, 5\}$. We used the RoBERTa base model [22] and generated the augmented data starting from the baseline dataset, following the procedure described in Section 3. The augmented data was then used to fine-tune the NER model, and its performance on the validation set was recorded. The best model was chosen based on the best F_1 score on the validation set. The selected number of augmentations for each dataset and dataset size is presented in Table 1.

5 Results and Analysis

Table 2 presents the performance metrics of our proposed methodology on the test sets of the datasets used in the experimental evaluation. Notably, our approach outperforms all the baselines used in the experimentation across different dataset sizes.

For CoNLL-2003 dataset, CPDA achieves an 1.69%, 0.6%, 0.45%, 0.52% and 0.58% gain in absolute F_1 scores for dataset size of 100, 200, 300, 400 and 500, respectively, when compared against the best performing baseline. Also, CPDA achieves an average performance gain of 10.32%, 1.24%, 1.05% and 1.4% across all five dataset sizes when compared against the baseline model, EDA, WordNet, and MELM respectively. We see a similar trend in Table 2 for WNUT-17 dataset, as well. Compared to the best performing baselines, the CPDA approach obtains 16.13%, 4.03%, 2.26%, 2.07%, and 1.26% absolute F_1 gains on dataset sizes of 100, 200, 300, 400, and 500, respectively. The CPDA approach shows an average performance gain of 6.92%, 5.73%, 6.37% and 6.08% over the baseline, EDA, WordNet, and MELM respectively. Additionally, the CPDA approach shows promising results on the SKILLSPAN dataset.Comparing against MELM [37], which is currently regarded as the best-performing baseline on SKILLSPAN, we gain absolute F_1 improvement of 3.33%, 0.58% and 3.75% for dataset sizes of 100, 300 and 400, respectively. However, for dataset sizes of 200 and 500 we perform 1.08% and 0.69% less than the MELM [37] in absolute F_1. These results underline the effectiveness of our proposed methodology in enhancing NER performance across diverse datasets, highlighting its potential for real-world applications.

Analyzing the results in more depth, we observe that CPDA outperforms baselines like EDA, WordNet, and MELM for the CoNLL-2003 and WNUT-17 datasets where the average performance gain is 2.82% and 5.15% (combining the $F1$ scores of all five dataset sizes) in absolute $F1$ score, respectively. Since, these datasets consist of general entities such as the name of a person, organization, location, group, etc... these entities are simpler to generalize since they follow simple semantic rules. For example, the name of a person, organization, and location are all nouns and appear in place of the subject in most sentences. Whereas, the SKILLSPAN dataset consists of soft skills that are made of multiple tokens and even phrases. These entities do not possess any clear definition, thereby, making the learning task more difficult. Despite the complexity of soft skills, the proposed approach shows promising performance over the baseline models also in this dataset. In SKILLSPAN dataset, CPDA demonstrates superior performance compared to the baselines for three out of five dataset sizes, with a slight degradation in performance for the remaining two sizes.

6 Error Analysis

This section aims to highlight the differences in predictions between our approach CPDA and MELM on the SKILLSPAN dataset, where MELM achieves a higher F1 score in two out of five cases. Specifically, the comparison aims to determine how often MELM and our approach agree or disagree in predicting soft skills, thereby seeking new insights which are not evident from the metrics of Precision and Recall reported in Table 2. To achieve this, we counted the number of entities correctly classified as soft skills by each model for each gold label in the dataset, in order to assess how often the two models agreed or disagreed. Results are reported in Table 4.

In order to highlight the main differences in classification capabilities of the models adopted in this work, in Table 5 we have reported qualitative examples obtained from the test set of SKILLSPAN dataset. The table compares the predictions of the RoBERTa model fine-tuned on the baseline dataset, the MELM augmented dataset,

[2] It is not straightforward to train/fine-tune these models for NER tasks.

Table 2: The F_1, precision and recall along with standard deviation are reported on the test set. The values are averaged over three different random initializations. N represents the size of the dataset subset.

N	Dataset	CoNLL2003			WNUT-17			SKILLSPAN		
		P	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$
100	Baseline	$61.26_{\pm 2.41}$	$65.65_{\pm 1.34}$	$63.37_{\pm 1.79}$	$28.43_{\pm 9.03}$	$9.92_{\pm 9.14}$	$13.57_{\pm 10.89}$	$17.91_{\pm 1.89}$	$18.17_{\pm 3.75}$	$17.89_{\pm 2.64}$
	EDA	$75.00_{\pm 1.81}$	$76.55_{\pm 2.37}$	$74.72_{\pm 1.74}$	$24.38_{\pm 17.33}$	$12.17_{\pm 10.75}$	$14.37_{\pm 14.41}$	$19.78_{\pm 3.04}$	$20.25_{\pm 4.59}$	$19.79_{\pm 3.10}$
	WordNet	$73.62_{\pm 2.63}$	$77.53_{\pm 3.13}$	$75.04_{\pm 2.73}$	$27.13_{\pm 22.20}$	$13.22_{\pm 10.41}$	$16.91_{\pm 14.77}$	$19.91_{\pm 2.30}$	$21.76_{\pm 3.30}$	$20.72_{\pm 2.48}$
	MELM	$72.42_{\pm 2.06}$	$77.10_{\pm 0.73}$	$74.67_{\pm 1.26}$	$33.28_{\pm 4.75}$	$12.69_{\pm 9.65}$	$17.08_{\pm 10.97}$	$17.61_{\pm 1.97}$	$19.85_{\pm 6.88}$	$18.28_{\pm 4.20}$
	CPDA	$75.09_{\pm 1.68}$	$78.49_{\pm 2.17}$	$\mathbf{76.73}_{\pm \mathbf{1.22}}$	$44.02_{\pm 6.57}$	$27.01_{\pm 7.38}$	$\mathbf{33.21}_{\pm \mathbf{7.52}}$	$20.27_{\pm 1.99}$	$23.67_{\pm 3.44}$	$\mathbf{21.61}_{\pm \mathbf{1.18}}$
200	Baseline	$65.33_{\pm 3.59}$	$70.38_{\pm 1.93}$	$67.74_{\pm 2.79}$	$43.75_{\pm 5.06}$	$28.04_{\pm 8.15}$	$33.80_{\pm 7.37}$	$21.96_{\pm 4.72}$	$21.26_{\pm 6.86}$	$21.41_{\pm 5.42}$
	EDA	$80.24_{\pm 1.63}$	$83.95_{\pm 2.57}$	$81.56_{\pm 1.92}$	$46.83_{\pm 5.36}$	$28.99_{\pm 10.46}$	$35.08_{\pm 4.88}$	$25.79_{\pm 6.65}$	$28.79_{\pm 11.13}$	$26.91_{\pm 8.71}$
	WordNet	$81.60_{\pm 1.16}$	$84.24_{\pm 1.33}$	$81.89_{\pm 1.13}$	$45.62_{\pm 6.08}$	$28.07_{\pm 7.54}$	$34.35_{\pm 6.97}$	$26.10_{\pm 3.64}$	$29.99_{\pm 6.94}$	$27.75_{\pm 4.72}$
	MELM	$78.66_{\pm 2.12}$	$81.73_{\pm 1.89}$	$80.17_{\pm 1.95}$	$43.05_{\pm 4.10}$	$29.60_{\pm 7.86}$	$34.77_{\pm 6.60}$	$26.54_{\pm 3.58}$	$36.55_{\pm 2.87}$	$\mathbf{30.53}_{\pm \mathbf{2.30}}$
	CPDA	$81.15_{\pm 0.74}$	$83.86_{\pm 1.33}$	$\mathbf{82.49}_{\pm \mathbf{0.93}}$	$46.54_{\pm 3.41}$	$33.77_{\pm 3.04}$	$\mathbf{39.11}_{\pm \mathbf{3.02}}$	$29.03_{\pm 3.47}$	$30.12_{\pm 3.38}$	$29.45_{\pm 2.90}$
300	Baseline	$67.12_{\pm 3.88}$	$71.42_{\pm 3.17}$	$69.20_{\pm 3.50}$	$47.35_{\pm 5.92}$	$33.89_{\pm 8.07}$	$39.12_{\pm 7.12}$	$25.53_{\pm 4.60}$	$27.23_{\pm 2.59}$	$26.21_{\pm 3.46}$
	EDA	$80.58_{\pm 1.74}$	$86.64_{\pm 0.96}$	$82.56_{\pm 1.33}$	$51.54_{\pm 6.53}$	$33.77_{\pm 5.77}$	$40.60_{\pm 5.47}$	$31.13_{\pm 4.58}$	$34.90_{\pm 7.21}$	$32.56_{\pm 4.56}$
	WordNet	$83.00_{\pm 1.38}$	$85.93_{\pm 1.26}$	$83.44_{\pm 1.18}$	$49.44_{\pm 8.86}$	$29.34_{\pm 10.23}$	$36.33_{\pm 2.71}$	$29.82_{\pm 2.25}$	$36.51_{\pm 4.33}$	$32.75_{\pm 2.64}$
	MELM	$80.97_{\pm 0.92}$	$84.00_{\pm 1.13}$	$82.45_{\pm 0.90}$	$45.54_{\pm 6.47}$	$34.28_{\pm 9.26}$	$38.70_{\pm 8.24}$	$32.04_{\pm 3.39}$	$40.18_{\pm 4.09}$	$35.46_{\pm 2.51}$
	CPDA	$82.62_{\pm 1.41}$	$85.23_{\pm 0.82}$	$\mathbf{83.89}_{\pm \mathbf{0.67}}$	$49.63_{\pm 3.26}$	$38.19_{\pm 5.41}$	$\mathbf{42.86}_{\pm \mathbf{2.98}}$	$36.55_{\pm 1.78}$	$35.89_{\pm 4.65}$	$\mathbf{36.04}_{\pm \mathbf{2.34}}$
400	Baseline	$76.92_{\pm 0.08}$	$80.29_{\pm 0.54}$	$78.57_{\pm 0.29}$	$52.84_{\pm 2.16}$	$34.88_{\pm 3.85}$	$41.88_{\pm 2.74}$	$26.70_{\pm 4.44}$	$31.77_{\pm 6.70}$	$28.82_{\pm 4.92}$
	EDA	$83.71_{\pm 0.77}$	$86.19_{\pm 0.67}$	$83.89_{\pm 0.61}$	$53.42_{\pm 6.76}$	$37.56_{\pm 5.26}$	$43.03_{\pm 3.57}$	$34.73_{\pm 4.86}$	$41.10_{\pm 5.61}$	$37.43_{\pm 4.28}$
	WordNet	$83.87_{\pm 1.23}$	$85.89_{\pm 0.71}$	$83.83_{\pm 0.90}$	$55.44_{\pm 1.85}$	$33.86_{\pm 3.94}$	$41.91_{\pm 3.23}$	$35.83_{\pm 2.75}$	$41.27_{\pm 4.98}$	$38.31_{\pm 3.52}$
	MELM	$82.60_{\pm 0.93}$	$86.24_{\pm 0.80}$	$84.38_{\pm 0.78}$	$52.52_{\pm 2.03}$	$33.94_{\pm 3.99}$	$41.11_{\pm 2.98}$	$33.46_{\pm 3.47}$	$39.74_{\pm 7.37}$	$35.83_{\pm 3.84}$
	CPDA	$83.55_{\pm 0.84}$	$86.30_{\pm 0.70}$	$\mathbf{84.90}_{\pm \mathbf{0.71}}$	$54.22_{\pm 5.14}$	$38.87_{\pm 4.34}$	$\mathbf{45.10}_{\pm \mathbf{3.61}}$	$39.17_{\pm 1.30}$	$40.31_{\pm 4.31}$	$\mathbf{39.58}_{\pm \mathbf{1.82}}$
500	Baseline	$82.70_{\pm 2.28}$	$84.55_{\pm 3.03}$	$83.62_{\pm 2.65}$	$54.48_{\pm 3.57}$	$35.82_{\pm 4.72}$	$43.06_{\pm 4.08}$	$29.06_{\pm 1.80}$	$34.06_{\pm 9.74}$	$30.79_{\pm 5.02}$
	EDA	$83.50_{\pm 0.77}$	$88.17_{\pm 0.76}$	$85.29_{\pm 0.60}$	$55.18_{\pm 5.08}$	$40.01_{\pm 2.43}$	$44.28_{\pm 2.61}$	$36.17_{\pm 6.58}$	$39.17_{\pm 9.23}$	$37.02_{\pm 6.43}$
	WordNet	$84.87_{\pm 1.23}$	$88.26_{\pm 0.61}$	$85.53_{\pm 0.82}$	$57.75_{\pm 2.51}$	$37.62_{\pm 2.20}$	$44.50_{\pm 1.56}$	$33.72_{\pm 6.12}$	$39.02_{\pm 11.14}$	$35.28_{\pm 6.54}$
	MELM	$83.82_{\pm 1.35}$	$87.12_{\pm 0.54}$	$85.43_{\pm 0.83}$	$53.67_{\pm 3.25}$	$37.20_{\pm 3.68}$	$43.86_{\pm 3.17}$	$41.05_{\pm 3.68}$	$45.30_{\pm 5.52}$	$\mathbf{42.65}_{\pm \mathbf{1.36}}$
	CPDA	$84.70_{\pm 0.40}$	$87.56_{\pm 0.94}$	$\mathbf{86.11}_{\pm \mathbf{0.47}}$	$53.87_{\pm 3.74}$	$39.97_{\pm 3.03}$	$\mathbf{45.76}_{\pm \mathbf{1.97}}$	$41.21_{\pm 2.27}$	$42.92_{\pm 2.95}$	$41.96_{\pm 1.66}$

Table 3: Inference of best performing CPDA model. Texts in orange colour stand for ground truth, whereas the predictions are reported in red. For each dataset considered in this work, we report the output of the best-performing model in generating the predictions. We choose *top* 4 predictions returned by the model.

Dataset	**Method**	**Text**
	Ground Truth	It said the KDP was responsible for breaking the previous ceasefire by refusing to endorse it publicly.
CoNLL-2003	CPDA	It said the UN was responsible for breaking the previous ceasefire by refusing to endorse it publicly. It said the UNHCR was responsible for breaking the previous ceasefire by refusing to endorse it publicly. It said the government was responsible for breaking the previous ceasefire by refusing to endorse it publicly. It said the Taliban was responsible for breaking the previous ceasefire by refusing to endorse it publicly.
	Ground Truth	Cowboy fans remember when Da Bears demolished you guys 44-0 in your own home ...
WNUT-17	CPDA	Cowboy fans remember when da Pats demolished you guys 44-0 in your own home in 1985 Cowboy fans remember when The Lions demolished you guys 44-0 in your own home in 1985 Cowboy fans remember when the Eagles demolished you guys 44-0 in your own home in 1985. Cowboy fans remember when Ohio Spartans demolished you guys 44-0 in your own home in 1985.
	Ground Truth	You'll also define and promote code standardization and automation processes for the organization to help us scale ...
SKILLSPAN	CPDA	You'll also build & manage promote automated and code tools for the organization to help us scale ... You'll also design and execute deploy new data engineering practices for the organization to help us scale ... You'll also code the deploy define automation product and services for the organization to help us scale ... You'll also develop / train about internal Data processing skills for the organization to help us scale ..

Table 4: Quantitative error analysis of soft skills in the SKILLSPAN dataset. Ground truth corresponds to the total number of entities present in the test set. The reported number in the table corresponds to the intersection of entities. Bear in mind that the table is symmetric.

Method	**Ground Truth**	**CPDA**	**MELM**
Ground Truth	795	295	234
CPDA	-	953	279
MELM	-	-	571

and the CPDA augmented dataset against the gold labels. From example 1, we observe that all three models fail to predict the “motivated” token as a soft skill, although the baseline and CPDA models predict more tokens than those in the gold label. Likewise, in example 2 and example 3, CPDA accurately predicts all the soft skill tokens correctly, whereas MELM fails. Example 4 provides an interesting insight into our model and the dataset. According to the gold labels, “develop solution” is not annotated as a soft skill, an issue that we attribute to human error. However, CPDA correctly predicts “develop solution” as a soft skill, although this prediction is then con-

Table 5: Qualitative error analysis of soft skills predictions on randomly sampled instances from the test set. We report the output of the best-performing model from Table 2. The *Baseline* column shows the predictions of the RoBERTa model fine-tuned with the baseline dataset, *MELM* column shows the predictions of RoBERTa model fine-tuned on the dataset augmented with MELM approach. The *CPDA* column shows the results of the RoBERTa model fine-tuned with the dataset augmented by CPDA. Highlighted texts stand for gold labels in the first column, and the corresponding predictions by the models in each column.

№	Gold Labels	Baseline	MELM	CPDA
1.	This opportunity requires a highly motivated candidate to work in a small and talented software development team in order to deliver of a next-generation analytics products for our institutional client base	This opportunity requires a highly motivated candidate to work in a small and talented software development team in order to deliver of a next-generation analytics products for our institutional client base	This opportunity requires a highly motivated candidate to work in a small and talented software development team in order to deliver of a next-generation analytics products for our institutional client base	This opportunity requires a highly motivated candidate to work in a small and talented software development team in order to deliver of a next-generation analytics products for our institutional client base
2.	You'll be required to apply your depth of knowledge and expertise to all aspects of the software development lifecycle as well as partner continuously with your many stakeholders on a daily basis to stay focused on common goals	You'll be required to apply your depth of knowledge and expertise to all aspects of the software development lifecycle as well as partner continuously with your many stakeholders on a daily basis to stay focused on common goals	You'll be required to apply your depth of knowledge and expertise to all aspects of the software development lifecycle as well as partner continuously with your many stakeholders on a daily basis to stay focused on common goals	You'll be required to apply your depth of knowledge and expertise to all aspects of the software development lifecycle as well as partner continuously with your many stakeholders on a daily basis to stay focused on common goals
3.	Open for continuous change with passion for automation and proven experience with Azure PaaS microservices	Open for continuous change with passion for automation and proven experience with Azure PaaS microservices	Open for continuous change with passion for automation and proven experience with Azure PaaS microservices	Open for continuous change with passion for automation and proven experience with Azure PaaS microservices
4.	You'll develop solutions for a bank entrusted with holding 18 trillion of assets and 393 billion in deposits	You'll develop solutions for a bank entrusted with holding 18 trillion of assets and 393 billion in deposits	You'll develop solutions for a bank entrusted with holding 18 trillion of assets and 393 billion in deposits	You'll develop solutions for a bank entrusted with holding 18 trillion of assets and 393 billion in deposits
5.	Company provides strategic advice raises capital manages risk and extends liquidity in markets spanning over 100 countries around the world	Company provides strategic advice raises capital manages risk and extends liquidity in markets spanning over 100 countries around the world	Company provides strategic advice raises capital manages risk and extends liquidity in markets spanning over 100 countries around the world	Company provides strategic advice raises capital manages risk and extends liquidity in markets spanning over 100 countries around the world

sidered a false positive in the overall evaluation. A similar pattern is observed in Example 5.

7 Conclusion

This paper proposes CPDA (Contextual Prompt Data Augmentation), a novel DA framework for Named Entity Recognition (NER) tasks using contrastive learning and prompt tuning. We use the category labels as a prompt and enable the CPDA to explicitly condition the prompt when predicting the masked entity tokens. The proposed method is able to solve the token-label misalignment problem meanwhile generating diverse entities. When leveraging RoBERTa [22] as the base model, our approach significantly outperforms four established baselines: the baseline RoBERTa model without augmentation, EDA, WordNet, and MELM, across various scenarios when the annotated dataset is scarce.

These findings highlight the robustness of our proposed CPDA methodology in improving NER performance, particularly in scenarios with limited training data. By leveraging prompts and data augmentation techniques, CPDA effectively addresses challenges posed by diverse datasets and complex entity types. Moreover, our analysis reveals insights into the performance disparities among baseline approaches across different datasets. While simpler datasets like CoNLL-2003 favor baseline approaches, more complex datasets like WNUT-17 and SKILLSPAN showcase the superiority of CPDA. This suggests that CPDA excels in scenarios where entities exhibit diverse semantic characteristics and may pose challenges for traditional approaches.

8 Limitations and Future Work

The proposed study shows promising results under lack of annotated data, especially when the annotated samples are extremely limited such as 100, 200 or 300. However, from Table 2 we see that, as we train CPDA on a higher number of training examples such as 400 and 500, the performance gain over the baseline starts to saturate.

In the future, we will improve the model performance further for a higher number of dataset sizes. Moreover, we plan to expand the proposed approach for different NLP tasks such as text classification, machine translation and question answering.

Acknowledgements

We gratefully acknowledge the support of the AMAJOR SRL SB. for sponsoring the research and providing the feedback. We also thank the University of Padova for providing the necessary computational resources to carry out the research. We also acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. We finally thank the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support.

References

[1] C. Chen and K. Shu. PromptDA: Label-guided data augmentation for prompt-based few shot learners. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 562–574, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.41.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[3] R. Cotterell and K. Duh. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing.

[4] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.161.

[5] X. Dai and H. Adel. 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), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.343.

[6] S. S. S. Das, A. Katiyar, R. J. Passonneau, and R. Zhang. Container: Few-shot named entity recognition via contrastive learning. In *ACL*, 2022.

[7] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4418.

[8] B. Ding, L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si, and C. Miao. 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, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.488.

[9] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. H. Hovy. A survey of data augmentation approaches for nlp. *ArXiv*, abs/2105.03075, 2021.

[10] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. 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 3816–3830, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295.

[11] T. Gao, X. Yao, and D. Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552.

[12] B. Heinzerling and K. Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.153.

[13] Y. Huang, K. He, Y. Wang, X. Zhang, T. Gong, R. Mao, and C. Li. 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, Oct. 2022. International Committee on Computational Linguistics.

[14] Hugging Face. Transformers APIs. https://huggingface.co/docs/transformers/index, 2023. Accessed: 2023-01-21.

[15] D. Iter, K. Guu, L. Lansing, and D. Jurafsky. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.439.

[16] P. Jiang and S. Saripalli. Contrastive learning of features between images and lidar. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, page 411–417. IEEE Press, 2022. doi: 10.1109/CASE49997.2022.9926589.

[17] A. Layegh, A. H. Payberah, A. Soylu, D. Roman, and M. Matskin. Contrastner: Contrastive-based prompt tuning for few-shot ner. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 241–249, 06 2023. doi: 10.1109/COMPSAC57700.2023.00038.

[18] J. Li, B. Li, Q. Zhang, X. Chen, X. Huang, L. Guo, and Y.-G. Fu. Graph contrastive representation learning with nbsp;input-aware and nbsp;cluster-aware regularization. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part II*, page 666–682, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-43414-3. doi: 10.1007/978-3-031-43415-0_39.

[19] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353.

[20] P. Liu, X. Wang, C. Xiang, and W. Meng. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195, 2020. doi: 10.1109/CCNS50731.2020.00049.

[21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3560815.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[23] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2024.

[24] G. A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

[25] R. Mirzaee, H. Rajaby Faghihi, Q. Ning, and P. Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.364.

[26] OpenAI. Gpt-4 technical report, 2023.

[27] D. Rigoni. Understanding multimedia content with prior knowledge. 2023.

[28] D. Rigoni, L. Serafini, and A. Sperduti. A better loss for visual-textual grounding. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 49–57, 2022.

[29] D. Rigoni, L. Parolari, L. Serafini, A. Sperduti, and L. Ballan. Weakly-supervised visual-textual grounding with semantic prior refinement. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, page 229. BMVA Press, 2023.

[30] E. F. Tjong Kim Sang and F. De Meulder. 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, 2003.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1670.

[33] D. Yang, Z. Zhang, and H. Zhao. Learning better masking for better language model pre-training, 2023.

[34] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[35] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

[36] M. Zhang, K. N. Jensen, S. D. Sonniks, and B. Plank. Skillspan: Hard and soft skill extraction from english job postings. In *North American Chapter of the Association for Computational Linguistics*, 2022.

[37] R. Zhou, X. Li, R. He, L. Bing, E. Cambria, L. Si, and C. Miao. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.160.