Natural Language Processing (2026), **32**, pp. 184–203
doi:10.1017/nlp.2026.10019

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

An experimental study on data augmentation techniques for named entity recognition on low-resource domains

Arthur Elwing Torres[1], Edleno Silva de Moura[1,2], Altigran Soares da Silva[1], Mario A. Nascimento[3] and Filipe Mesquita[4]

[1]Universidade Federal do Amazonas, Manaus, Amazonas, Brazil, [2]Jusbrasil, Salvador, Bahia, Brazil, [3]University of Alberta, Edmonton, Alberta, Canada, and [4]Diffbot Technologies Corp., Menlo Park, California, USA
Corresponding authors: Arthur Elwing Torres; Email: arthur.torres@icomp.ufam.edu.br, Mario A. Nascimento; Email: mario.nascimento@ualberta.ca

(Received 9 December 2024; revised 30 December 2025; accepted 11 February 2026)

Abstract

Named Entity Recognition (NER) is a natural language processing task that traditionally relies on supervised learning and annotated data. Acquiring such data is often a challenge, particularly in specialized fields like medical, legal, and financial sectors. Those are commonly referred to as *low-resource* domains, which comprise *long-tail* entities, due to the scarcity of available data. To address this, data augmentation techniques are increasingly being employed to generate additional training instances from the original dataset. In this study, we evaluate the effectiveness of two prominent text augmentation techniques, *Mention Replacement* and *Contextual Word Replacement*, on two widely used NER models, Bi-LSTM + CRF and BERT. We conduct experiments on three datasets from low-resource domains, and we explore the impact of various combinations of training subset sizes and the number of augmented examples. We not only confirm that data augmentation is particularly beneficial for smaller datasets, but we also demonstrate that there is no universally optimal number of augmented examples, i.e., NER practitioners must experiment with different quantities in order to fine-tune their projects.

Keywords: named entity recognition; data augmentation; low-resource domains; long-tail entities

1. Introduction

Named Entity Recognition (NER) is a specialized subtask of Information Extraction that focuses on identifying and categorizing mentions of real-world entities in text (Jurafsky and Martin 2000; Li, Shang, and Chen 2021; Li *et al.* 2022a). Common entity types such as *Person*, *Location*, and *Organization* are frequently encountered in abundant online sources such as news, websites, and social media, and thus form what are called *generic* domains. These domains have been extensively studied, and numerous datasets are readily available for research (Li *et al.* 2020; Wang *et al.* 2021a). In contrast, entities in specialized application domains, also called *low-resource* domains (Dai and Adel 2020; Liu, Chen, and Xu 2022), like medical, legal, and financial sectors, are termed as *long-tail* entities (Cao *et al.* 2020; Liu *et al.* 2020). These domains are less commonly found in typical online text corpora. Additionally, the annotation of these specialized entities often demands technical expertise, making the generation of annotated datasets a challenging task (Hou *et al.* 2020; Friedrich *et al.* 2020). Given the critical role of annotated data in the development, training, and testing of NER models (Li 2022; Au, Lampos, and Cox 2022), there is a growing interest in

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

methodologies for building NER models in scenarios where annotated data are scarce or difficult to generate, as is the case with these low-resource domains (Mohammed, Khan, and Bashier 2016).

One such methodology is the application of data augmentation, which is a technique that improves the construction of effective NER models when annotated training data is limited or costly to obtain. This approach automatically generates new data samples by applying transformations to existing data (Wong *et al.* 2016; Krizhevsky, Sutskever, and Hinton 2017; Shorten and Khoshgoftaar 2019). Originally popularized in the field of computer vision, data augmentation is gaining traction in Natural Language Processing (NLP) tasks, including NER (Xie *et al.* 2018; Dai and Adel 2020; Liu *et al.* 2020, 2022; Tikhomirov *et al.* 2020). For example, a sentence like *I have a cat named Serena* could be transformed into *You have a dog named Beethoven* or *I own a cat called Serena*, thereby enriching the training set. These augmented sentences maintain the core structure of the original, while introducing variations that enhance the diversity of the training set.

Despite the growing interest, existing research has not sufficiently explored how the volume of augmented examples influences the performance of NER models. This is a crucial parameter that needs to be determined prior to model training. Additionally, many applications of NER require domain-specific entities; therefore, determining the best ways to create and augment the training data is important in the practical field of NLP. In this study, we examine the effectiveness of two specific data augmentation techniques, *Mention Replacement* (MR) and *Contextual Word Replacement* (CWR), across three datasets from various low-resource domains.

We experiment with different quantities of augmented sentences and diverse samples from these datasets, employing BERT (Devlin *et al.* 2019) and Bi-LSTM + CRF (Huang, Xu, and Yu 2015), two widely used NER architectures (Dai and Adel 2020; Liu *et al.* 2022).

Consistent with prior research, our findings confirm that data augmentation is particularly beneficial for smaller datasets. However, for larger datasets, we observed that models trained with augmented data yielded equivalent or even inferior performance compared to those trained without augmentation. We examine the impact of varying amounts of augmented data on model performance as our key contribution. We demonstrate that there exists a saturation point beyond which additional augmentation may degrade model quality. This underscores the need for NER practitioners to carefully experiment with different volumes of augmented examples, as the optimal quantity cannot be predetermined.

Our experiments also reveal that CWR generally outperforms MR, and that BERT models benefit more from data augmentation than Bi-LSTM + CRF models, at least within the scope of our datasets.

The remainder of this paper is structured as follows: Section 2 provides an overview of NER models and text data augmentation techniques. Section 3 reviews prior works that have applied data augmentation to NER. Section 4 details the datasets, augmentation techniques, and experimental parameters. Finally, Section 5 presents and discusses our findings, and Section 6 concludes the paper.

2. Background

In this section, we present a review of NER techniques in the literature, as well as a summary of the best-known text augmentation methods, focusing mainly on the ones applied alongside those techniques.

2.1 NER techniques

In the realm of NER, various architectures have been proposed in the literature (Yadav and Bethard 2018; Li *et al.* 2022b). Classical approaches include the *Hidden Markov Model* (HMM) (Baum and Petrie 1966), the *Maximum-entropy Markov Model* (MEMM) (McCallum, Freitag, and Pereira 2000), and the *Conditional Random Fields* (CRF) (Lafferty, McCallum, and Pereira 2001). In addition to these, Deep Learning-based models, such as *Long short-term Memory* (LSTM)

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

(Hochreiter and Schmidhuber 1997) and *Gated Recurrent Unit* (GRU) (Cho *et al.* 2014), have also emerged as effective architectures for NER tasks. While this diverse range of models offers a rich set of techniques for developing robust and accurate NER systems, two prominent architectures have shown exceptional performance within the context of NER, particularly in low-resource settings (Dai and Adel 2020; Liu *et al.* 2022).

The first, *Bi-LSTM + CRF* (Huang *et al.* 2015), served as the standard for NER before the rise of Transformers. This architecture uses a bidirectional LSTM to process text in both directions, creating a rich representation of each word based on its full context. The critical component is the final CRF layer, which performs global label decoding. Instead of predicting each tag independently, the CRF considers the relationships between neighbor tags (e.g., ensuring an "I-ORG" tag always follows a "B-ORG"), which significantly reduces illegal tag sequences and improves overall consistency.

The second, *BERT* (Bidirectional Encoder Representations from Transformers) (Devlin *et al.* 2019), is an evolution of the *Transformer* architecture. Unlike previous models that process sequences linearly, BERT uses an attention mechanism to weigh the importance of every word in a sentence simultaneously. The model is first pre-trained on massive unlabeled corpora to learn general language patterns and is then fine-tuned on a specific NLP-task dataset. For the NER task, this process of fine-tuning leverages the model's understanding derived from pre-training, but also tunes it to the specific dataset, allowing it to learn the unique context and the specific labels it needs to recognize (Dai and Adel 2020).

Comparative studies often reveal that BERT models outperform Bi-LSTM + CRF models in NER tasks (Li *et al.* 2022b). However, it is important to consider the higher computational overhead associated with BERT, which typically involves the estimation of approximately 110 million parameters, compared to a typically lower parameter count for Bi-LSTM + CRF models. In our experiments, this count ranged from approximately 0.8 million to 2 million parameters, depending on the dataset.

2.2 Text augmentation methods

Data augmentation techniques for NER are generally categorized into two groups: substitution-based and language model-based (Zhang *et al.* 2024). The former relies on exchanging tokens or entities with equivalents of the same type, drawing from the training corpus or external knowledge bases. In contrast, language model-based techniques utilize the predictive capabilities of discriminative or generative architectures. While discriminative models typically employ masked language modeling for token replacement, generative models exploit next-token prediction to synthesize entirely new sentences.

We summarize some of the best-known text augmentation techniques in the literature below, grouped into the two aforementioned categories. Table 1 shows examples of applying these techniques. We adopt the well-known IOB (Ramshaw and Marcus 1995) tagging scheme. In this scheme, each target entity is tagged with its corresponding entity type, prefixed by "B-" (beginning). Tokens that do not belong to any relevant entity type are labeled as "O" (outside). For sequences of consecutive tokens that share the same entity type, the initial token is prefixed with "B-", while subsequent tokens receive an "I-" (inside) prefix to denote a contiguous "chunk" of tokens.

- **Substitution-based methods**
  - *Easy Data Augmentation (EDA)* (Wei and Zou 2019): EDA encompasses a suite of straightforward text manipulations, which include:
    - Synonym Replacement (SR) — Replacing random words in a sentence with their corresponding synonyms from a thesaurus.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 1. Text augmentation techniques with examples

Technique	Sentence										Obs.
None	I	took	a	medicine	today	for	acute	colitis	.		
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		
EDA -SR	I	took	a	**remedy**	today	for	acute	colitis	.		
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		
EDA -RI	I	took	a	medicine	today	**and**	for	acute	colitis	.	
	O	O	O	B-treatment	O	O	O	B-disease	I-disease	O	
EDA -RS	I	**today**	a	medicine	**took**	for	acute	colitis	.		
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		
EDA-RD	I	took	medicine	today	for	acute	colitis	.			Removed token “a”
	O	O	B-treatment	O	O	B-disease	I-disease	O			
LwTR	I	**he**	a	**surgery**	today	for	acute	**headache**	.		New tokens came from training corpus
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		
MR	I	took	a	medicine	today	for	**cancer**	.			New tokens came from training corpus
	O	O	O	B-treatment	O	O	B-disease	O			
SWS	**took**	**a**	**I**	medicine	**for**	**today**	**colitis**	**acute**	.		
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		
CWR	I	**got**	a	medicine	**now**	for	acute	colitis	.		New tokens came from BERT
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		
LcWR	I	took	a	**surgery**	today	for	acute	**headache**			New tokens came from BERT
	O	O	O	B-treatment	O	O	B-disease	I-disease	O		

    * Random Insertion (RI) — Inserting random synonyms of words at arbitrary positions in the sentence.
    * Random Swap (RS) — Interchanging the positions of two randomly-selected words in the sentence.
    * Random Deletion (RD) — Removing a randomly-chosen word from the sentence.

    Notably, some techniques discussed later in this section can be viewed as nuanced variations of EDA methods.
  - *Label-wise Token Replacement (LwTR)* (Dai and Adel 2020): This method involves replacing random words in a sentence with other words from the original training set that share the same label.
  - *Mention Replacement (MR)* (Dai and Adel 2020): An extension of LwTR, MR replaces randomly-selected mentions, that is, one or more contiguous tokens with a uniform label type, with other mentions from the original training set that have the same label type. The number of tokens in the replacement mention may differ from the original.
  - *Shuffle Within Segments (SWS)* (Dai and Adel 2020): In this approach, a sentence is segmented based on mentions (as defined in MR). Tokens within each segment are then randomly reordered.
- **Language model-based methods**
  - *Contextual Word Replacement (CWR)* (Jiao *et al.* 2020): This method employs BERT models to replace words contextually within a sentence.
  - *Label-conditioned Word Replacement (LcWR)* (Liu *et al.* 2022): This technique utilizes a pre-trained BERT model that is fine-tuned to capture word-label dependencies for word replacement tasks.

In the field of NER, Bi-LSTM + CRF and BERT have been identified as particularly effective architectures for low-resource domains (Dai and Adel 2020; Liu *et al.* 2022). To both adhere to established research and examine the cost-effectiveness of data augmentation across different model types, we opted to utilize both architectures. For data augmentation, we focused on methods that are well regarded in the literature and have demonstrated improvements in the performance of the NER model (Dai and Adel 2020; Liu *et al.* 2020, 2022; Jiao *et al.* 2020), while also being relatively simple to implement. Hence, we chose MR for augmenting tokens tagged with specific entity types, excluding those labeled "O," and CWR for tokens specifically tagged as "O." Detailed information on the implementation of these augmentation techniques is available in Section 4.

3. Related work

Next, we discuss previous work on data augmentation for NER models in this section. One particular issue with current data augmentation methods in the literature is that most of the work done on NLP tasks considers annotations at the sentence level rather than at the token level, as it should be the case with NER (Dai and Adel 2020).

In a closely related study, Dai and Adel (2020) investigated the effectiveness of four text augmentation techniques, LwTR, SR, MR, and SWS, described in Section 2.2, on two specific low-resource domains: biomedical and materials science. These authors assessed the impact of text augmentation on NER models employing both Bi-LSTM + CRF and BERT architectures. Their findings indicate that none of the augmentation methods consistently outperforms the others, and that augmentation is more effective when applied to smaller (and random) subsets rather than to the entire dataset. Another relevant contribution comes from Liu *et al.* (2022), who introduced a novel text augmentation technique, LcWR, also described in Section 2.2, as well as an

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

automatic sentence annotation approach using BERT, termed *prompting with question answering*. Additionally, they proposed a noise-reduction system to counteract the potential introduction of noisy data by the augmentation techniques. Their experiments, conducted on three datasets, including one in the low-resource domain of materials science, utilized both Bi-LSTM + CRF and BERT architectures. They found that their techniques not only improved the overall model quality but were particularly effective on smaller subsets of the datasets, while also generating more label-consistent augmented data.

Our work aligns with these studies in terms of the experimental protocol, as we too explore various text augmentation techniques across multiple low-resource domains, employing the same NER architectures. However, a key distinction lies in our exploration of different quantities of augmented instances, enabling us to evaluate the impact of these varying amounts on the quality of the NER models.

Another similar work conducted a comprehensive examination of text augmentation techniques on a dataset encompassing long-tail entities within academic papers (Liu *et al.* 2020). The techniques used were Entity Replacement (ER), which is a variant of MR (Section 2.2) that replaces entities with alternative entities from sources other than the original training set, and Entity Mask (EM), which is similar to CWR (Section 2.2), but focuses exclusively on replacing entity tokens. The authors evaluated the impact of varying data augmentation scales: 20%, 50%, and 100% of the original dataset. This evaluation was performed using a model that seamlessly integrated BERT, Bi-LSTM, and CRF architectures. Their findings demonstrated an enhancement in model quality attributed to both ER and EM augmentation techniques. However, a noteworthy observation is the non-linear relationship between the volume of augmented examples and model improvement. This insight underscores the necessity for further exploration in this domain, as increasing the quantity of augmented data does not invariably bolster the model's performance. Remarkably, this study stands as a pioneering contribution to the literature, being the sole research to rigorously assess the influence of diverse data augmentation magnitudes on NER models to our knowledge.

In another closely related work, Tikhomirov *et al.* (2020) introduce two innovative text augmentation techniques, termed "inner" and "outer" descriptor replacement. These techniques involve replacing non-specific entity mentions, termed as *descriptors*, with proper names to enhance specificity (e.g., replacing "medicine" with "Paracetamol"). The *Inner* variant applies this replacement to sentences in the original training set, whereas the *Outer* variant extends this to automatically annotate unlabeled sentences that contain descriptors and are outside the original dataset. These methods are applied to a dataset within the cybersecurity domain, a field characterized by its low-resource nature. However, the application is confined to a limited set of entity types, specifically *virus* and *hacker*. The authors carried out a comprehensive evaluation, assessing the impact of these augmentation techniques on models constructed using diverse architectures, including CRF, BERT, and two specially fine-tuned variations of BERT, denominated as *RuBERT* and *RuCYBERT*. These tailored models are explicitly designed to address the problem at hand. The empirical results indicate that while the introduced augmentation techniques enhance the model recall, a concomitant decline in precision is observed. Notably, substantial improvements are manifested in the CRF and BERT models. In contrast, the RuBERT and RuCYBERT models exhibit negligible enhancement, thereby alluding to the potential significance of fine-tuning prior to the application of data augmentation techniques. This observation propels further inquiry into the intricate interplay between fine-tuning and data augmentation, and their collective impact on model performance.

Finally, in Longpre *et al.* (2020), the authors assessed the efficacy of backtranslation alongside a suite of easy data augmentation (EDA) techniques, described in Section 2.2, across various NLP tasks, employing six diverse datasets and three distinct transformer architectures: BERT, RoBERTa (Liu *et al.* 2019), and XLNet (Yang *et al.* 2019). Despite not directly addressing NER, the research holds significance. Their findings reveal an inconsistent enhancement in quality by the examined text augmentation techniques on transformer models. This inconsistency underscores the critical

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

role of the scale of pre-training in models constructed using transformers when contemplating the application of data augmentation. Furthermore, Longpre *et al.* (2020) propose a potential preference for task-related augmentation over task-agnostic methods, exemplified by the use of backtranslation for machine translation tasks.

In conclusion, there has been limited advancement in low-resource NER settings. This work aims to address this gap by conducting further experiments and presenting additional results. Utilizing established NER architectures and augmentation techniques, such as those outlined in Section 2.2, the experiments are designed in accordance with methods used in current research. In particular, there is a notable lack of research concerning the impact of the number of augmented sentences on NER models. Although one study has touched upon this issue (Liu *et al.* 2020), it is of a smaller scale compared to the present work. The amount of augmentation is an important *hyperparameter* that must be set before training the model, as it can introduce noise and affect the quality of the dataset. By investigating the effect of different numbers of augmented sentences, as described in Section 4, this work aims to provide insights into how this parameter influences NER model quality.

4. Experimental setup

In this section, we present the details of the experiments we carried out on the effectiveness of data augmentation in different low-resource domain datasets.

4.1 Datasets

In our experiments, we used three different datasets, all of which are in the English language. Two are from the biomedical domain: BioCreative V CDR Task (Li *et al.* 2016), which contains 'Chemical' and 'Disease' entities, and i2b2-2010 (Uzuner *et al.* 2011), which includes 'treatment', 'problem', and 'test' types. The third dataset, MaSciP (Mysore *et al.* 2019), is from the materials science domain, featuring a wide range of specialized entities related to synthesis protocols, including categories such as 'Material', 'Operation', 'Apparatus-Descriptor', 'Condition-Type', among others. Other articles on NER models (Dai and Adel 2020; Liu *et al.* 2022) previously used i2b2-2010 and MaSciP.

For all datasets, we used the train-dev-test split provided by the authors. For MaSciP, we the used *NLTK* default sentence tokenizer (*PunktSentenceTokenizer*[a]). Table 2 summarizes the number of sentences, tokens, entities, and entity types in each dataset. We also present the average and median entity length in tokens in each dataset, as well as the average and median number of different unique strings per entity type. Finally, we also present the percentage of unique words in the testing set that were not seen in the training and validation sets.

To assess the impact of data augmentation in various low-resource scenarios, we selected different sentence subsets from each training dataset: 50, 150, and 500 sentences, aligning with prior research (Dai and Adel 2020; Liu *et al.* 2022). Additionally, subsets amounting to 25%, 50%, and 75% of the original dataset size are chosen, provided these subsets exceed 500 sentences. The validation and testing datasets remain unaltered throughout the evaluation.

4.2 Data augmentation

We used two techniques in our experiments, *MR* and *CWR*, both described in Section 2. We applied two small changes to MR. First, we replaced the mentions in the sentences with new mentions only seen in the training subsets. This does not apply to the full dataset. This was done because we did not want the technique to use additional information from external sources, which

[a] https://www.nltk.org/api/nltk.tokenize.PunktSentenceTokenizer.html

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 2. Datasets used in the experiment

Dataset	Training set					
	Sentences	Tokens	Entity types	# of entities	Entity length (tokens)	# of unique strings per entity type
i2b2-2010	13,867	127,151	3	14,376	2.1 (avg); 2.0 (median)	5,137.67 (avg); 4,628.0 (median)
MaSciP	2,253	53,718	21	18,874	1.29 (avg); 1.0 (median)	29.9 (avg); 12.0 (median)
BioCreative V CDR	1,000	116,913	2	9,385	1.27 (avg); 1.0 (median)	1,268.5 (avg); 1,268.5 (median)
Dataset	Validation set					
	Sentences	Tokens	Entity Types	# of Entities	Entity Length (tokens)	# of unique strings per Entity Type
i2b2-2010	2,448	22,390	3	2,143	2.16 (avg); 2.0 (median)	671.33 (avg); 650.0 (median)
MaSciP	138	3,548	20	1,190	1.26 (avg); 1.0 (median)	30.55 (avg); 16.5 (median)
BioCreative V CDR	1,000	115,965	2	9,591	1.29 (avg); 1.0 (median)	1,209.0 (avg); 1,209.0 (median)
Dataset	Testing Set					
	Sentences	Tokens	Entity Types	# of Entities	Entity Length (tokens)	# of unique strings per Entity Type
i2b2-2010	27,625	267,249	3	31,161	2.12 (avg); 2.0 (median)	2,658.67 (avg); 2,280.0 (median)
MaSciP	182	4,066	21	1,259	1.26 (avg); 1.0 (median)	246.19 (avg); 96.0 (median)
BioCreative V CDR	1,000	122,838	2	9,809	1.32 (avg); 1.0 (median)	1,258.5 (avg); 1,258.5 (median)
Dataset			Language			Unseen words in testing set (%)
i2b2-2010			English			38.01
MaSciP			English			79.6
BioCreative V CDR			English			32.19

would be the discarded part of the dataset for each subset. Second, we selected a random number of tokens of the new mention to produce more variability in the dataset. For augmented sentences with CWR, we applied the technique only to words whose tags are not of the target entity types of the dataset, that is, the annotated tag is "O". In this way, we assume that BERT is not highly specialized in the low-resource language domains of the datasets, replacing only the words that require less technical expertise. By applying these two augmentation techniques separately and with the aforementioned changes, we can generate augmented sentences by applying data augmentation only to the tokens of target entity types (MR), and sentences where augmentation was applied only to tokens of tag "O" (CWR).

4.3 Setup

We selected two popular architectures used to build NER models, Bi-LSTM + CRF (Huang *et al.* 2015) and BERT (Devlin *et al.* 2019). Following previous work (Dai and Adel 2020), we used SciBERT (Beltagy, Lo, and Cohan 2019), a fine-tuned version of BERT for scientific texts, for datasets from the biomedical and materials science domains, and regular multilingual BERT for the other datasets. For each dataset and its subsets, we gradually increased the amount of augmented data in the training, ranging from 0% (original data without augmented data) to 500% (original data plus 5 times the amount of original data as augmented data). Finally, we used validation loss as a parameter to interrupt model training if it stopped decreasing for 5 epochs. This setup yields 532 groups of models, considering the combination of all datasets and subsets, architectures, and augmentation techniques and amounts. Each group consists of 10 models trained from scratch, which we evaluated in the testing data, in order to check model variance. We calculated the average F1-score of the models of each group, which we present in Section 5.

While datasets like MaSciP present a significantly higher classification difficulty due to the increased complexity of inter-class disambiguation, it is important to note that we present only the average F1-score for each group (calculated by taking the micro-average F1-score for each model and computing the mean). Although a per-class analysis would provide deeper insights, presenting this breakdown for every entity type would result in an overwhelming volume of data that would overshadow the primary trends of this research, given the scale of our experimental setup. Therefore, we report only the aggregated F1-score to maintain a standardized comparison across all architectures and datasets, consistent with established benchmarks in the field (Dai and Adel 2020; Liu *et al.* 2020).

All codes used in the experiments are publicly available to encourage the carrying out of more experiments.[b]

5. Results and discussion

We now present the results of experiments performed with the datasets described in Section 4 to study the effects of different amounts of data augmentation on the quality of NER models.

5.1 General results

In Figures 1 and 2, we plot a distinct graph for each dataset and architecture pair, using, respectively, MR and CWR as the augmentation technique. For each of these pairs, we plot the average F1-score of each 10-model group as a function of the amount of data augmentation used. Each curve in the graphs corresponds to a subset of the full dataset, both in terms of percentage (25%, 50%, 75%, and 100%) and in terms of the absolute number of sentences (50, 150, and 500).

[b] https://github.com/arttorres0/augmented-ner-model

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

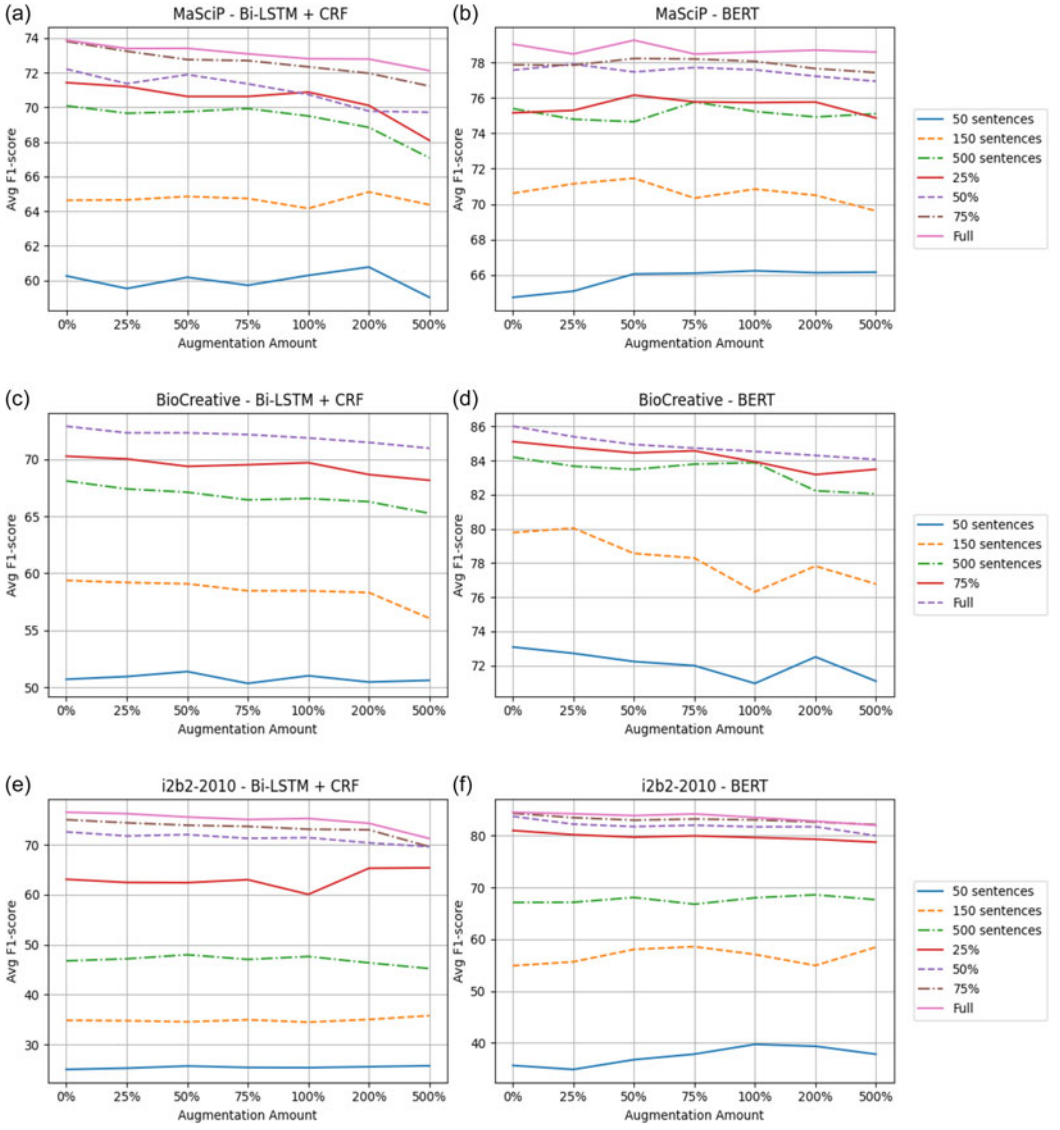

Figure 1. Effects of different amounts of augmented examples (MR) on average F1-score by model and dataset.

As detailed in Section 4, depending on the size of the dataset, not all percentages and absolute numbers of sentences were considered.

Looking at Figures 1 and 2, we notice that the BERT models outperform the Bi-LSTM + CRF models for almost all datasets. This is consistent with previous results in the literature that compare BERT and Bi-LSTM + CRF models (Dai and Adel 2020; Liu *et al.* 2022).

We observe a modest increase in model quality for lower-sized subsets before the model starts to reduce the F1-score value, when analyzing models augmented with MR. However, for the full dataset and other larger subsets, the average F1-score has not increased, maintaining or losing quality. The only exceptions to this are *MaSciP* (BERT), in which every subset benefits from data augmentation. This reduction in F1-score may be a consequence of the introduction of invalid augmented data, which had their original ground-truth labels altered by the augmentation technique. An example of this noise introduction is the following: consider the sentence "Pt had HIDA

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

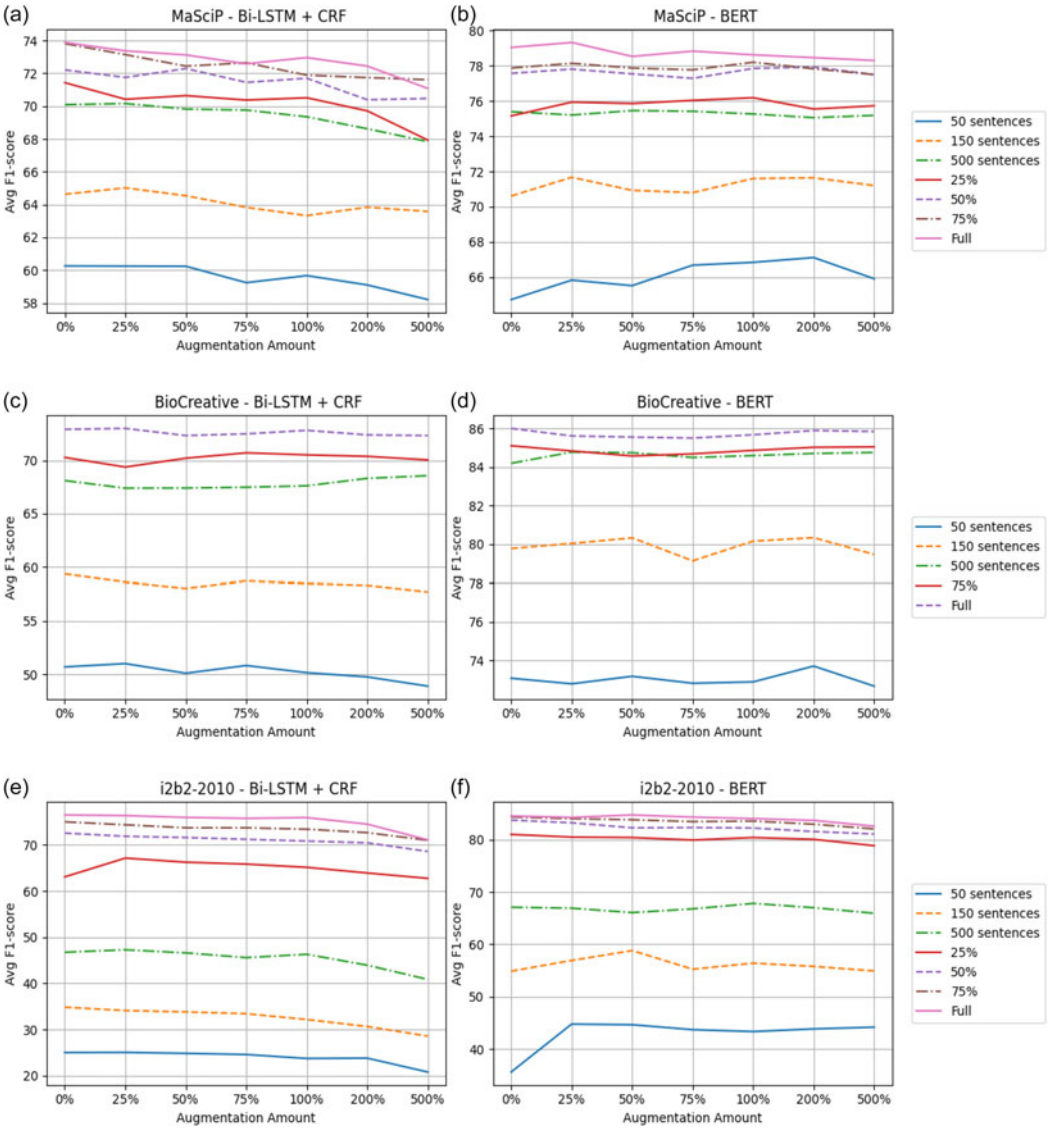

Figure 2. Effects of different amounts of augmented examples (CWR) on average F1-score by model and dataset.

scan which was negative for **obstruction**, and CT was unremarkable", where "obstruction" is a *B-PROBLEM* entity. Searching for other entities of the same class, we may find "increased shortness of breath". As MR selects a random number of tokens, a new augmented sentence could be "Pt had HIDA scan which was negative for **of breath**, and CT was unremarkable.", which is a syntactically incorrect sentence. These kinds of sentences can be introduced in the dataset. This suggests that data augmentation should be used with caution, as the injection of high noise into the training dataset may cause models to overfit these invalid instances. This trade-off between diversity and validity has been reported in previous research (Dai and Adel 2020; Xie *et al.* 2020). These inconsistencies could also be due to other factors, such as the characteristics of the datasets or the parameters of the augmentation techniques. We believe that identifying these factors is an important area for future research.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 3. 0% and best augmentation (MR) average F1-scores and applied paired Student's t-tests for i2b2 dataset

	i2b2-2010 (BERT)					i2b2-2010 (Bi-LSTM + CRF)				
Training size	F1 No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test	F1 No No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test
50 sent.	35.65	39.73	11.44	100	NS	25.01	25.74	2.92	500	S
150 sent.	54.89	58.57	6.70	75	S	34.83	35.77	2.70	500	NS
500 sent.	67.09	68.57	2.21	200	NS	46.75	47.97	2.61	50	NS
25%	80.97	80.97	0	0	–	63.08	65.39	3.66	500	NS
50%	83.7	83.7	0	0	–	72.57	72.57	0	0	–
75%	84.3	84.3	0	0	–	75.01	75.01	0	0	–
Full	84.49	84.49	0	0	–	76.5	76.5	0	0	–

Models augmented with CWR show similar results. However, we notice that more models benefit from data augmentation. This is an indicator that this technique produces more useful data variability than MR in the studied datasets, which makes sense, as BERT can suggest several different tokens. We also have more possible target replacements ("O" tokens) than when applying MR, which targets non-"O" tokens. On the other hand, we observe a more inconsistent pattern of improvement in F1-score across different subset sizes. This is also a consequence of the higher variability introduced by BERT, but on the negative side, where there is more noise introduced than actual valid augmented examples in the datasets.

5.2 Detailed results

Tables 3–8 present a different perspective of the above results. We show the average F1-score for each subset and each model without data augmentation, and indicate which amount of augmentation yields the best models on average. This summarization yields 76 averages. We also determine whether the augmentation technique produces statistically significant differences between the model without augmentation and the best augmented model, determined by paired Student's t-tests with a significance level of 0.05. This is only if the best amount of augmentation is greater than 0%. In these tables, we indicate the training size and, for each model, we present the F1-score of the model with no augmentation (column "F1 No Aug"), the F1-score of the best model (column "F1 Best"), the absolute difference in F1-score between these two (column "Δ"), and which augmentation amount yields the best model (column "Aug Amount"). Column "Student's T-test" indicates the result of a Student's t-test, if "Δ" is nonzero, where "S" indicates that the difference is statistically significant, and "NS" means that it is not.

The pattern observed in the results of Section 5.1 can also be observed in these tables, where we see some smaller subsets benefiting from data augmentation, while larger subsets never improve. While we refer to 'smaller' datasets throughout this paper, it's important to note that the term is relative and depends on the specific context and domain. For instance, in our experiments, we found that data augmentation was particularly beneficial for subsets of the datasets with fewer than 500 sentences, such as the 50-sentence subsets of the BioCreative dataset. However, this threshold may vary for other datasets or domains, and further research is needed to determine the optimal size for applying data augmentation techniques. Another observation is that there is no optimal predetermined amount of data augmentation, probably due to the random introduction of noise by the augmentation techniques. Because of this, we can conclude that NER practitioners should test different amounts of augmented examples in their models.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 4. 0% and best augmentation (MR) average F1-scores and applied paired Student's t-tests for MaSciP dataset

	i2b2-2010 (BERT)					i2b2-2010 (Bi-LSTM + CRF)				
Training size	F1 No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test	F1 No No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test
50 sent.	64.74	66.24	2.32	100	S	60.26	60.77	0.85	200	NS
150 sent.	70.61	71.46	1.20	50	NS	64.63	65.11	0.74	200	NS
500 sent.	75.39	75.76	0.49	75	NS	70.09	70.09	0	0	–
25%	75.15	76.15	1.33	50	S	71.43	71.43	0	0	–
50%	77.56	77.9	0.44	25	NS	72.21	72.21	0	0	–
75%	77.86	78.22	0.46	50	NS	73.81	73.81	0	0	–
Full	79.03	79.25	0.28	50	NS	73.88	73.88	0	0	–

Table 5. 0% and best augmentation (MR) average F1-scores and applied paired Student's t-tests for BioCreative dataset

	i2b2-2010 (BERT)					i2b2-2010 (Bi-LSTM + CRF)				
Training size	F1 No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test	F1 No No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test
50 sent.	73.08	73.08	0	0	–	50.7	51.38	1.34	50	NS
150 sent.	79.78	80.04	0.33	25	NS	59.37	59.37	0	0	–
500 sent.	84.2	84.2	0	0	–	68.1	68.1	0	0	–
75%	85.11	85.11	0	0	–	70.27	70.27	0	0	–
Full	86.01	86.01	0	0	–	72.9	72.9	0	0	–

Table 6. 0% and best augmentation (CWR) average F1-scores and applied paired Student's t-tests for i2b2 dataset

	i2b2-2010 (BERT)					i2b2-2010 (Bi-LSTM + CRF)				
Training size	F1 No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test	F1 No No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test
50 sent.	35.65	44.8	25.67	25	S	25.01	25.04	0.12	25	NS
150 sent.	54.89	58.82	7.16	50	S	34.83	34.83	0	0	–
500 sent.	67.09	67.81	1.07	100	NS	46.75	47.28	1.13	25	NS
25%	80.97	80.97	0	0	–	63.08	67.15	6.45	25	S
50%	83.7	83.7	0	0	–	72.57	72.57	0	0	–
75%	84.3	84.3	0	0	–	75.01	75.01	0	0	–
Full	84.49	84.68	0.22	50	NS	76.5	76.5	0	0	–

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 7. 0% and best augmentation (CWR) average F1-scores and applied paired Student's t-tests for MaSciP dataset

	i2b2-2010 (BERT)					i2b2-2010 (Bi-LSTM + CRF)				
Training size	F1 No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test	F1 No No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test
50 sent.	64.74	67.12	3.68	200	S	60.26	60.26	0	0	–
150 sent.	70.61	71.67	1.50	25	S	64.63	65.02	0.60	25	NS
500 sent.	75.39	75.45	0.08	50	NS	70.09	70.16	0.10	25	NS
25%	75.15	76.18	1.37	100	S	71.43	71.43	0	0	–
50%	77.56	77.92	0.46	200	NS	72.21	72.29	0.11	50	NS
75%	77.86	78.19	0.42	100	NS	73.81	73.81	0	0	–
Full	79.03	79.31	0.35	25	NS	73.88	73.88	0	0	–

Table 8. 0% and best augmentation (CWR) average F1-scores and applied paired Student's t-tests for BioCreative dataset

	i2b2-2010 (BERT)					i2b2-2010 (Bi-LSTM + CRF)				
Training size	F1 No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test	F1 No No Aug	F1 Best	Δ (%)	Aug amount (%)	Student's t-test
50 sent.	73.08	73.71	0.86	200	NS	50.7	51.01	0.61	25	NS
150 sent.	79.78	80.34	0.70	200	NS	59.37	59.37	0	0	–
500 sent.	84.2	84.78	0.69	25	NS	68.1	68.56	0.68	500	NS
75%	85.11	85.11	0	0	–	70.27	70.7	0.61	75	NS
Full	86.01	86.01	0	0	–	72.9	72.99	0.12	25	NS

Table 9 summarizes the content of the previous tables, presenting all models that achieved statistically significant improvements by data augmentation. We sorted these models by improvement, in descending order. We note that 10 of the 76 best models, or 13,16%, were found to present statistically significant improvements in the average F1-score. The conclusions above are also evident here, that is, first, smaller subsets improve with data augmentation, unlike larger subsets, and second, in spite of these promising results, there is not an augmentation amount that yields the best results.

We also observe that, for the datasets studied in this experiment, CWR yields more improved models than MR, that is, 6 versus 4. This can be an indication that CWR should be prioritized when selecting data augmentation techniques for NER models. We can also see that two of the three datasets benefit from data augmentation, where i2b2-2010 and MaSciP have 5 improved models each. Although i2b2-2010 and BioCreative are similar datasets in terms of domain and number of entities, models of the first dataset are improved by data augmentation, unlike models of the second dataset, which never improved. This contrasting effect can be explained by intrinsic differences in text style and annotation characteristics between the two sources. The i2b2-2010 dataset is based on relatively homogeneous clinical narratives where relations are often expressed locally and follow recurrent linguistic patterns; in this setting, data augmentation strategies are more likely to preserve relational semantics while increasing useful variability for training. In contrast, BioCreative relies on scientific abstracts in which chemical–protein relations depend on

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 9. Summary of models improved by augmentation

Dataset	Training size	Model	Improvement (%)	Aug amount (%)	Aug technique
i2b2-2010	50 sentences	BERT	25.67	25	CWR
i2b2-2010	150 sentences	BERT	7.16	50	CWR
i2b2-2010	150 sentences	BERT	6.70	75	MR
i2b2-2010	25%	Bi-LSTM + CRF	6.45	25	CWR
MaSciP	50 sentences	BERT	3.68	200	CWR
i2b2-2010	50 sentences	Bi-LSTM + CRF	2.92	500	MR
MaSciP	50 sentences	BERT	2.32	100	MR
MaSciP	150 sentences	BERT	1.50	25	CWR
MaSciP	25%	BERT	1.37	100	CWR
MaSciP	25%	BERT	1.33	50	MR

denser, highly specialized context and are often distributed across multiple sentences, in addition to involving normalization to controlled vocabularies. Under these more complex conditions, simple synthetic perturbations are more prone to introducing semantic noise or disrupting critical dependencies, which explains why augmentation fails to yield improvements and may even harm performance in the scientific domain.

Finally, BERT models are improved more than Bi-LSTM + CRF models, that is, 8 versus 2. This can be an indication that data augmentation suits BERT-based NER models better than Bi-LSTM + CRF-based ones.

In order to assess whether models are truly worsened by large amounts of data augmentation, we developed Tables 10 and 11, which followed a similar process as Table 9: first, we selected the models with the most significant decrease in average F1-score and the models without data augmentation for each of the 76 aforementioned groups. Second, we performed paired Student's t-tests with a significance level of 0.05 to determine if the difference in average F1-score between the worst model and the model without data augmentation is statistically significant, provided that the worst model is not the model without data augmentation (i.e., when every amount of data augmentation improved that group of models). We identified 44 models that exhibit statistically significant decreases in performance, with 23 identified by MR and 21 by CWR. We observed that the vast majority of these models were affected by the largest amount of data augmentation in our experiments (500%), indicating that the introduction of large amounts of data augmentation may introduce noise into the dataset rather than merely reflecting statistical variance.

The results obtained in these experiments show that there is an unfilled potential for the improvement of low-resource NER models, where all models of the studied datasets have below 90% F1-score. Compared to other datasets of high-resource domains studied in the literature, where the quality of the model is higher (Yadav and Bethard 2018; Li *et al.* 2020; Zhong and Chen 2021; Wang *et al.* 2021a; Zhong and Chen 2021), there is still a margin for improvement of low-resource NER models.

5.3 Practical recommendations for data augmentation in low-resource NER

Based on the empirical evidence presented in Sections 5.1 and 5.2, we summarize below practical recommendations for applying data augmentation in low-resource NER settings.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 10. Summary of models worsened by augmentation (MR)

Dataset	Training size	Model	Decrease (%)	Aug amount (%)
MaSciP	500 sentences	Bi-LSTM + CRF	−4.49	500
MaSciP	25%	Bi-LSTM + CRF	−4.91	500
MaSciP	50%	Bi-LSTM + CRF	−3.57	500
MaSciP	75%	Bi-LSTM + CRF	−3.62	500
MaSciP	Full	Bi-LSTM + CRF	−2.44	500
MaSciP	50%	BERT	−0.82	500
BioCreative	150 sentences	Bi-LSTM + CRF	−5.92	500
BioCreative	500 sentences	Bi-LSTM + CRF	−4.35	500
BioCreative	75%	Bi-LSTM + CRF	−3.10	500
BioCreative	Full	Bi-LSTM + CRF	−2.72	500
BioCreative	50 sentences	BERT	−3.00	100
BioCreative	150 sentences	BERT	−4.55	100
BioCreative	500 sentences	BERT	−2.62	500
BioCreative	75%	BERT	−2.32	200
BioCreative	Full	BERT	−2.32	500
i2b2-2010	25%	Bi-LSTM + CRF	−5.03	100
i2b2-2010	50%	Bi-LSTM + CRF	−4.27	500
i2b2-2010	75%	Bi-LSTM + CRF	−7.77	500
i2b2-2010	Full	Bi-LSTM + CRF	−7.37	500
i2b2-2010	25%	BERT	−2.86	500
i2b2-2010	50%	BERT	−4.66	500
i2b2-2010	75%	BERT	−2.65	500
i2b2-2010	Full	BERT	−3.06	500

First, data augmentation should be considered primarily in genuinely low-resource regimes. In our experiments, statistically significant improvements were observed almost exclusively for small training subsets (typically fewer than 500 sentences). For larger subsets, augmentation rarely improved performance and often degraded it.

Second, moderate amounts of augmentation should be preferred over aggressive augmentation. The largest augmentation rate evaluated (500%) was responsible for the majority of statistically significant performance degradations, indicating that excessive augmentation tends to introduce noise rather than useful variability.

Third, CWR is generally a safer default than MR. Across datasets and models, CWR produced more statistically significant improvements and fewer severe degradations, suggesting that contextual augmentation of non-entity tokens better preserves sentence validity.

Finally, BERT-based NER models benefit more consistently from data augmentation than Bi-LSTM + CRF models. Most improvements were observed for BERT, whereas Bi-LSTM + CRF models were more frequently harmed, particularly under large augmentation rates.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Table 11. Summary of models worsened by augmentation (CWR)

Dataset	Training size	Model	Decrease (%)	Aug amount (%)
MaSciP	50 sentences	Bi-LSTM + CRF	−3.52	500
MaSciP	150 sentences	Bi-LSTM + CRF	−2.05	100
MaSciP	500 sentences	Bi-LSTM + CRF	−3.32	500
MaSciP	25%	Bi-LSTM + CRF	−5.15	500
MaSciP	50%	Bi-LSTM + CRF	−2.59	200
MaSciP	75%	Bi-LSTM + CRF	−3.07	500
MaSciP	Full	Bi-LSTM + CRF	−3.94	500
MaSciP	Full	BERT	−0.95	500
BioCreative	50 sentences	Bi-LSTM + CRF	−3.68	500
BioCreative	150 sentences	Bi-LSTM + CRF	−2.88	500
BioCreative	75%	BERT	−0.63	50
i2b2-2010	50 sentences	Bi-LSTM + CRF	−20.36	500
i2b2-2010	150 sentences	Bi-LSTM + CRF	−21.91	500
i2b2-2010	500 sentences	Bi-LSTM + CRF	−14.42	500
i2b2-2010	50%	Bi-LSTM + CRF	−5.79	500
i2b2-2010	75%	Bi-LSTM + CRF	−5.65	500
i2b2-2010	Full	Bi-LSTM + CRF	−7.59	500
i2b2-2010	25%	BERT	−2.73	500
i2b2-2010	50%	BERT	−3.27	500
i2b2-2010	75%	BERT	−2.78	500
i2b2-2010	Full	BERT	−2.36	500

Overall, while no single augmentation configuration can be recommended universally, these results delineate clear boundaries within which data augmentation is more likely to be beneficial and less likely to be harmful.

6. Conclusion and future work

In this paper, we investigate the effectiveness of two representative data augmentation techniques, MR (Dai and Adel 2020) and CWR (Jiao *et al.* 2020), when used for NER tasks on low-resource domain datasets, which contain long-tail entities. In our study, we use two popular neural architectures for NER, Bi-LSTM + CRF and BERT, to generate several different NER models with distinct configurations. For training, these configurations used subsets of varying sizes of the original datasets.

Specifically, we generated 532 groups of 10 models, from which we measure the average F1-score and pick the best one for each subset. This yields 76 averages, 10 of which are considered statistically significant improvements after using text augmentation. As these improved models were those trained using smaller subsets, our study further supports the notion that data augmentation for NER models is more effective for smaller training sets than for larger ones. We also

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

verify that there is no predetermined optimal number of augmented instances that will certainly improve the model. Not only that, but also augmentation also reduces model quality in several other cases. Thus, NER practitioners should test different amounts, especially if the augmentation technique is prone to introducing spurious instances in the training data. We also show that, for the datasets studied in this paper, CWR yields models that are better than MR, and that data augmentation suits BERT NER models better than Bi-LSTM + CRF.

Our work on augmentation for NER in low-resource domains indicates that this landscape can still be improved in both the case of top-notch but high-demanding architectures, such as BERT, or in the case of competitive but less-demanding architectures, such as Bi-LSTM + CRF. While our study provides valuable insights into the effectiveness of data augmentation techniques for NER on low-resource domains, it also highlights several limitations that warrant further investigation.

Firstly, our findings show that data augmentation can introduce noise into the training data, which can potentially degrade the quality of the NER models. This is particularly evident in our experiments with larger subsets of the datasets, where data augmentation often resulted in lower average F1-scores compared to models trained without augmented data. This suggests that while data augmentation can help to increase the quantity of training data, it may also compromise its quality, leading to overfitting or poor generalization to unseen data.

To mitigate this issue, future research could explore more sophisticated or controlled augmentation techniques that minimize the introduction of noise. For example, one could consider techniques that take into account the semantic consistency between the original and augmented sentences, or that adjust the degree of augmentation based on the complexity or diversity of the dataset.

Secondly, our study assumes that all entities are equally important and should be augmented at the same rate. However, in many real-world applications, some entities may be more critical or rare than others, and therefore might benefit more from augmentation. Future work could investigate differential augmentation strategies that prioritize certain entity types based on their importance or rarity.

Lastly, our study focuses on two specific NER architectures (Bi-LSTM + CRF and BERT) and two text augmentation techniques (MR and CWR). While these methods are widely used and have shown promising results in many NLP tasks, they may not be optimal or applicable for all types of NER problems or domains. Future research should consider exploring other architectures and techniques (e.g., knowledge graphs to suggest new entities), as well as hybrid or ensemble methods that combine the strengths of multiple approaches. Also, as future work, we intend to explore the use of generative models, particularly large language models (LLMs), due to recent experiments adopting them in NER. We aim to implement LLMs not only as advanced NER models but also as innovative data augmentation techniques. By leveraging the generative capabilities of these models, we anticipate enhancing the robustness and accuracy of our NER systems while simultaneously enriching our datasets. This dual approach could lead to improved performance in diverse NER tasks and better adaptability to various domains.

Data availability statement. All datasets and subsets used in this project and all notebooks generated during the experiments are available in the following repository: https://github.com/arttorres0/augmented-ner-model.

Funding statement. This research is partially supported by Jusbrasil under the UFAM/Jusbrasil Doctoral Fellowship granted to Arthur Torres; by Diffbot Inc.; by CNPq under INCTs IAIA (406417/2022-9) and TILD-IAR (408490/2024-1), Project Rule2 DB (UNIVERSAL Proc. 400936/2025-9), and individual grants to Altigran da Silva (301925/2025-9 and Edleno Moura (310573/2023-8); by the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES) financial code 001; and by FAPEAM under the POSGRAD 2025 Program and the NeuralBond Project (UNIVERSAL 2023 Proc. 01.02.016301.04300/2023-04).

Competing interests. The authors have no conflicts of interest to declare that are relevant to the content of this article.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

References

Au T.W.T., **Lampos V.** and **Cox I.** (2022). E-NER — an annotated named entity recognition corpus of legal text. In Aletras, N., Chalkidis, I., Barrett, L., Goanță, C. and Preoțiuc-Pietro, D. (eds), *Proceedings of the Natural Legal Language Processing Workshop 2022*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics, pp. 246–255.

Baum L.E. and **Petrie T.** (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics* **37**(6), 1554–1563.

Beltagy I., **Lo K.** and **Cohan A.** (2019). SciBERT: A pretrained language model for scientific text. 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)*, pp. 3615–3620.

Cao E., **Wang D.**, **Huang J.** and **Hu W.** (2020). Open knowledge enrichment for long-tail entities. In *Proceedings of The Web Conference 2020*, pp. 384–394.

Cho K., **van Merriënboer B.**, **Gulcehre C.**, **Bahdanau D.**, **Bougares F.**, **Schwenk H.** and **Bengio Y.** (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734.

Dai X. and **Adel H.** (2020). An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3861–3867.

Devlin J., **Chang M.-W.**, **Lee K.** and **Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

Friedrich A., **Adel H.**, **Tomazic F.**, **Hingerl J.**, **Benteau R.**, **Marusczyk A.** and **Lange L.** (2020). The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1255–1268.

Hochreiter S. and **Schmidhuber J.** (1997). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Hou Y., **Che W.**, **Lai Y.**, **Zhou Z.**, **Liu Y.**, **Liu H.** and **Liu T.** (2020). Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1381–1393.

Huang Z., **Xu W.** and **Yu K.** (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Jiao X., **Yin Y.**, **Shang L.**, **Jiang X.**, **Chen X.**, **Li L.**, **Wang F.** and **Liu Q.** (2020). TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4163–4174.

Jurafsky D. and **Martin J.H.** (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st edn. USA: Prentice Hall PTR.

Krizhevsky A., **Sutskever I.** and **Hinton G.E.** (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM* **60**(6), 84–90.

Lafferty J.D., **McCallum A.** and **Pereira F.C.N.** (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289.

Li J., **Chiu B.**, **Feng S.** and **Wang H.** (2022a). Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering* **34**(9), 4245–4256.

Li J., **Shang S.** and **Chen L.** (2021). Domain generalization for named entity boundary detection via metalearning. *IEEE Transactions on Neural Networks and Learning Systems* **32**(9), 3819–3830.

Li J., **Sun A.**, **Han J.** and **Li C.** (2022b). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* **34**(1), 50–70.

Li J., **Sun Y.**, **Johnson R.J.**, **Sciaky D.**, **Wei C.-H.**, **Leaman R.**, **Davis A.P.**, **Mattingly C.J.**, **Wiegers T.C.** and **Lu Z.** (2016). BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database* **2016**, 1–10.

Li W. (2022). The advance of deep learning based named entity recognition. *Highlights in Science, Engineering and Technology* **12**, 68–73.

Li X., **Sun X.**, **Meng Y.**, **Liang J.**, **Wu F.** and **Li J.** (2020). Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 465–476.

Liu J., **Chen Y.** and **Xu J.** (2022). Low-resource ner by data augmentation with prompting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 4252–4258.

Liu Q., **Li P.**, **Lu W.** and **Cheng Q.** (2020). Long-tail dataset entity recognition based on data augmentation. In *Extraction and Evaluation of Knowledge Entities from Scientific Documents at Joint Conference on Digital Library*, pp. 79–80.

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

Longpre S., **Wang Y.** and **DuBois C.** (2020). How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4401–4411.

McCallum A., **Freitag D.** and **Pereira F.C.N.** (2000). Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 591–598.

Mohammed M., **Khan M.** and **Bashier E.** (2016). *Machine Learning: Algorithms and Applications*. 1st edn. Boca Raton, FL: CRC Press.

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press

Mysore S., **Jensen Z.**, **Kim E.**, **Huang K.**, **Chang H.-S.**, **Strubell E.**, **Flanigan J.**, **McCallum A.** and **Olivetti E.** (2019). The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pp. 56–64.

Ramshaw L. and **Marcus M.** (1995). Text chunking using transformation-based learning. In *Third Workshop On Very Large Corpora*, pp. 82–94.

Shorten C. and **Khoshgoftaar T.** (2019). A survey on image data augmentation for deep learning. *Journal of Big Data* **6**, 1–48.

Tikhomirov M., **Loukachevitch N.**, **Sirotina A.** and **Dobrov B.** (2020). Using BERT and augmentation in named entity recognition for cybersecurity domain. In Métais, E., Meziane, F., Horacek, H., and Cimiano, P.(eds), *Natural Language Processing and Information Systems*, pp. 16–24.

Uzuner O., **South B.R.**, **Shen S.** and **DuVall S.L.** (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* **18**(5), 552–556.

Wang X., **Jiang Y.**, **Bách N.**, **Wang T.**, **Huang Z.**, **Huang F.** and **Tu K.** (2021a). Automated concatenation of embeddings for structured prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 2643–2660.

Wang X., **Jiang Y.**, **Bách N.**, **Wang T.**, **Huang Z.**, **Huang F.** and **Tu K.** (2021b). Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 1800–1812.

Wei J. and **Zou K.** (2019). 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 InternationalJoint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 6382–6388.

Wong S.C., **Gatt A.**, **Stamatescu V.** and **McDonnell M.D.** (2016). Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6.

Xie J., **Yang Z.**, **Neubig G.**, **Smith N.** and **Carbonell J.** (2018). Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 369–379.

Xie Q., **Dai Z.**, **Hovy E.**, **Luong M.-T.** and **Le Q.V.** (2020). Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 6256–6268.

Yadav V. and **Bethard S.** (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2145–2158.

Yang Z., **Dai Z.**, **Yang Y.**, **Carbonell J.**, **Salakhutdinov R.** and **Le Q.V.** (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., pp. 5753–5763.

Zhang X., **Chen G.**, **Cui S.**, **Sheng J.**, **Liu T.** and **Xu H.** (2024). Exogenous and endogenous data augmentation for low-resource complex named entity recognition. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*. New York, NY, USA: Association for Computing Machinery, pp. 630–640.

Zhong Z. and **Chen D.** (2021). A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 50–61.

Cite this article: Elwing Torres A, Silva de Moura E, Soares da Silva A, A. Nascimento M and Mesquita F (2026). An experimental study on data augmentation techniques for named entity recognition on low-resource domains. *Natural Language Processing* **32**, 184–203. https://doi.org/10.1017/nlp.2026.10019

https://doi.org/10.1017/nlp.2026.10019 Published online by Cambridge University Press