# Data Augmentation for Named Entity Recognition in Domain-Specific Scenarios in Portuguese

**Higor Moreira[1], Patricia Ferreira da Silva [2], Luciana Bencke[1] and Viviane Moreira[1]**
[1]Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil
[2]Petrobras Research and Development Center (CENPES), Brazil
{hmoreira, lrbencke, viviane}@inf.ufrgs.br, patricia.fs@petrobras.com.br

## Abstract

Named Entity Recognition (NER) is an important task of Natural Language Processing. Achieving good results in this task usually requires a large amount of labeled data to train models. This is especially difficult in domain-specific datasets and low-resourced languages. To mitigate the high cost of human-annotated data, data augmentation can be used. In this work, we evaluate Data Augmentation techniques for NER, focusing on domain-specific datasets in Portuguese. We employed data augmentation techniques based on rules, back-translation, and large language models on four datasets of varying sizes to train Transformer-based NER models. The results showed that most techniques improved over the baseline, with the best results achieved using PP-LLM, SR, and MR.

## 1 Introduction

Named Entity Recognition (NER) is a Natural Language Processing (NLP) task that focuses on identifying entities of interest in a sentence and classifying them into predefined categories (*e.g.,* person, organization, location) (Jurafsky and Martin, 2026; Keraghel et al., 2024). NER can be applied to a wide range of downstream tasks such as question answering (Kim et al., 2022), machine translation (Nowakowski et al., 2022), and information retrieval (Perera et al., 2020). The scope of NER expanded over time to accommodate concepts of increased complexity, including specialized domains, such as legal, biomedical, oil & gas, *etc.* (Amaral, 2017; Luz de Araujo et al., 2018; Schneider et al., 2020; Costa et al., 2022). For example, in the oil & gas domain, NER techniques can be useful to identify entities such as rock types, basins, fields, and geochronological units (Moreira et al., 2025).

Most work on NER has focused on general-domain corpora in English (Tjong Kim Sang and De Meulder, 2003; Hovy et al., 2006; Jiang et al., 2022). In specialized domains, the limited available work is also predominantly in English, resulting in a scarcity of NLP resources, such as trained models and annotated datasets, in many languages. This is particularly true for the Portuguese, especially when specific domains are considered. Building NER models for specific languages is important because each language has its own unique linguistic structures, morphology, syntax, naming conventions, and cultural entities, which generic or cross-lingual models often fail to capture accurately (Bayer et al., 2022).

When a specific domain and language are considered, the scenario is likely to involve limited data. In this case, data augmentation techniques can potentially reduce the need for human-annotated data. However, unlike text classification tasks such as polarity prediction and hate speech detection, applying data augmentation to a sequence labeling task, such as NER, is more challenging. NER is sensitive to word manipulation; a simple word substitution can lead to a mismatch between the token and its original label. While data augmentation is often used to mitigate class imbalance in classification tasks, in NER, this is not guaranteed, as many classes may co-occur in a single sentence. The side effect of augmenting an infrequent category is that it also augments a very frequent one. Our focus was on comparing different techniques rather than solving the class imbalance problem.

This work focuses on data augmentation for NER in domain-specific datasets in Portuguese. Our evaluation considered four datasets on various domains, seven data augmentation techniques, and varying sample sizes of original instances. Our main findings showed that data augmentation can improve NER results across most scenarios relative to the baseline. Among the techniques, PP-LLM, SR, and MR were the top performers.

We make our augmented datasets publicly available at https://huggingface.co/datasets/hm



*Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)*
*Vol. 1*, pages 250–259

oreira/da-datasets-propor2026.

## 2 Background

This section provides the necessary background on NER and data augmentation.

### 2.1 NER

NER is a sequence labeling task in NLP that identifies and categorizes named entities in a text corpus. These entities, defined as words or expressions, refer to objects from the real world, like persons, organizations, locations, numeric values for generic domains; rock types, basins, and fields identifiers in the oil & gas industry.

Even with the introduction of Large Language Models (LLMs) into NLP tasks, the application in NER has revealed some limitations, as they are originally designed for text generation. Transformer-based models currently lead the field in NER. This architecture used self-attention mechanisms to effectively capture and integrate contextual information. Empirical studies indicate that employing BERT as a classifier consistently outperforms traditional BiLSTM-CRF (Keraghel et al., 2024; Nunes et al., 2024a). The standard approach to sequence labeling for NER is the BIO (beginning-inside-outside) tagging. This method treats NER as a word-by-word sequence-labeling task, using tags that capture both the boundary and the named entity type.

### 2.2 Data Augmentation

Data augmentation refers to methods that increase the amount of data by modifying the original data or creating synthetic data from it. A number of data augmentation methods have been proposed or adapted to be applied for NER (Dai and Adel, 2020; Yaseen and Langer, 2021; Zhou et al., 2022; Sharma et al., 2023). Existing methods can be divided into the following categories.

**Rule-based techniques** consist of randomly replacing a token with its synonym, retrieved from a Thesaurus (such as *WordNet*), a language model, or a target vocabulary. Alternatively, a specific instance of an entity can be replaced with another randomly selected instance of the same entity class. The selected entity for replacement can be retrieved from the training set, a knowledge graph, or another source. This technique may also include shuffling within segments, splitting the labeled sequence into segments, and reordering the labeled sequence. Label-wise token replacement uses a binomial distribution derived from the original training set to randomly replace tokens in the labeled sequence.

**Back-Translation:** Given an input sequence of tokens, the sequence is split into segments. These segments are the context around the named entity. The segments are translated to intermediate pivot languages and then back to the source language. Back-Translation can be used with language models, API, or LLMs (Yaseen and Langer, 2021).

**PLM-based Conditional Generation:** uses pretrained Masked Language Models (MLMs), to generate diverse tokens within a context. A prominent example is Masked Entity Language Modeling (MELM), which fine-tunes an MLM to explicitly condition the prediction of masked entity tokens on their labels (Zhou et al., 2022).

**LLM-based Paraphrasing** leverages the world knowledge and rewriting capabilities of LLMs to create new labeled data from original sources. The augmented data should convey information very similar to that of the original data.

## 3 Related Work

Data augmentation has been explored to address data scarcity in low-resource NER. Early approaches primarily relied on rule-based word-level modifications. Dai and Adel (2020) investigated synonym replacement and mention replacement, while Torres et al. (2024) examined similar strategies for Portuguese, noting that heuristic rules can sometimes compromise semantic integrity.

To overcome the limitations of rule-based methods, generative neural approaches have been introduced to synthesize entirely new, labeled sentences. Ding et al. (2020) proposed DAGA, employing an LSTM-based language model trained on linearized sentences. Building on masked language modeling, Zhou et al. (2022) introduced MELM to generate diverse entities within specific contexts. However, these methods often rely heavily on the sequential dependency of entities seen in training. Addressing this, Hu et al. (2022) developed EnTDA, an entity-to-text framework that decouples entity dependencies. By performing operations such as adding, deleting, or swapping entities in a list before generating the corresponding text, EnTDA effectively acts as an inverse NER task, improving generalization across flat, nested, and discontinuous NER.

More recently, the focus has shifted towards prompting an LLM to enhance robustness and

diversity. Song et al. (2024) proposed RoPDA, a prompt-based method using continuous soft prompts on Pre-trained Language Models (PLMs). RoPDA introduces specific operations for label flipping and context augmentation, coupled with a self-consistency filtering mechanism to eliminate low-quality samples and a mixup strategy to mitigate overfitting to adversarial examples. Leveraging the advanced rewriting capabilities of LLMs, Ye et al. (2024) introduced LLM-DA, which augments data at both the contextual and entity levels. Their approach employs multiple rewriting strategies, altering sentence length or presentation style, to maintain semantic coherence while injecting noise to bolster model robustness. Focusing on the trade-off between diversity and adherence to data distribution, Evuru et al. (2024) proposed CoDa. Unlike unconstrained LLM generation, CoDa extracts specific lexical, syntactic, and semantic constraints from the source data and constructs natural language instructions to guide the LLM, ensuring that the generated augmentations remain linguistically consistent with the target domain.

Despite these significant advancements, most research remains concentrated on English benchmarks (Hu et al., 2022; Song et al., 2024; Ye et al., 2024; Evuru et al., 2024). Research on generative data augmentation for Portuguese NER is still limited. In this work, we aim to bridge this gap by evaluating and combining these advanced techniques within the context of Portuguese domains.

## 4 Materials and Methods

This section describes the methodology we followed to evaluate data augmentation methods for Portuguese NER. Figure 1 shows the pipeline used. The input (Step 1) is a NER dataset. From the training split of the dataset, in Step 2, we select a number of instances. In Step 3, we apply data augmentation methods to the selected instances. The augmented data is combined with the original instances (Step 4) and used to train a NER model (Step 5). Finally, we evaluate the generated models on the test set (Step 6).

### 4.1 Data Augmentation Techniques

This section describes the configuration and techniques used for data augmentation mentioned in Section 2. We selected a diverse set of strategies, ranging from heuristic rule-based methods to approaches that leverage pre-trained language models (MLMs and LLMs). These techniques encompass token replacement, translation, and conditional generation paradigms.

**Synonym Replacement (SR)**: With a threshold of 20% for the tokens to be replaced by a synonym. While this method works more effectively for English in general domains, it may not perform as well for other languages, such as Portuguese, which have a smaller *WordNet*. For this reason, we only consider the context surrounding the named entity for SR when selecting replacement tokens that are not part of the named entity.

**Mention Replacement (MR)**: Each entity in the sentence is replaced by another entity belonging to the same class, randomly selected from a gazetteer created from the original training set. The length of the replacement entity can differ from the original entity length; therefore, the BIO-label sequence is adjusted to match the new length.

**Context-Aware Mention Replacement (CAMR)**: Similar to MR, this technique uses a gazetteer derived from the training set to source replacement candidates. However, to mitigate the risk of generating semantically incoherent sentences, a common limitation of random replacement is that it relies on a pre-trained MLM as a scoring mechanism. For each instance, the original entity is masked, and the MLM calculates the log-probability of each candidate from the gazetteer fitting into that specific context. The final replacement is sampled from the top-scoring candidates, ensuring that the substituted entity aligns semantically with the surrounding text.

**MELM**: Creates augmented data by fine-tuning a pre-trained MLM to generate diverse entities. To address the issue where a replaced token might not match the original entity type (token-label misalignment), MELM employs a labeled sequence linearization strategy (Zhou et al., 2022). This involves explicitly inserting label tokens before and after entity mentions in the input sequence. During the generation phase, the model masks entity tokens and predicts replacements by conditioning on both the sentence context and these injected label markers, ensuring the generated entities align with the correct class.

**Back-Translation (BT)**: Used three steps. First, the input sentence is split into token sequences (segments) that represent the context surrounding a named entity. Only the context surrounding the named entity is considered for the back-translation, as we need to preserve the BIO format in the output.

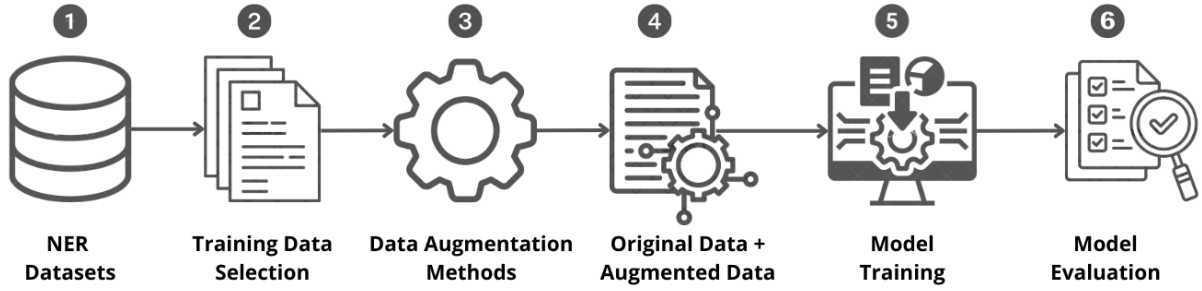



Figure 1: Pipeline for data augmentation in NER

The entity is excluded from the context because it may be translated into a different word, its position could change, and we might lose the annotation as a result. To prevent breaking the back-translation loop, a segment is valid for back-translation if it has three or more tokens and fewer than 70% non-alphabetical characters. Segments are translated from Portuguese to English and then back to the source language, with back-translated tokens replacing the originals and the BIO-label sequence adjusted accordingly.

**LLM-based Paraphrasing (PP-LLM)**: Each sequence of tokens in the training set, each named entity surrounded in the sentence with the prefix <ENTITY> to indicate the beginning and end of the entity. The prompt explicitly instructs the LLM not to remove or alter the named entity between the markers, but only to paraphrase the context surrounding the entity, without altering the overall meaning. In our experiments, we tested with GPT4.1 (Hurst et al., 2024).

**Retrieval-Augmented Generation for Data Augmentation (RAG-DA)**: Draws inspiration from the LLM-DA framework. While LLM-DA demonstrates the potential of generative models to synthesize training samples, our approach adapts this concept to the specialized requirements of the oil & gas domain. By integrating an external knowledge retrieval component, we aim to direct the LLM's generative capabilities towards producing content that reflects authentic technical contexts found in domain documentation. The proposed pipeline consists of three stages:

*1. Domain Context Retrieval*: We use the source instance as a query to retrieve potentially relevant contexts from the REGIS collection (Lima de Oliveira et al., 2021), a corpus on the oil & gas domain. The collection segments were encoded using BGE embeddings (Chen et al., 2024) and indexed. For each input instance, the system retrieves the top-$k = 50$ most semantically similar chunks to serve as a knowledge base for augmentation.

*2. Entity-Constrained Generation*: Adapting the generative process, we employ a prompt[1] that contains the named entities from the source instance and the top-$k = 10$ retrieved contexts. The LLM is instructed to generate a new sentence that preserves the original entities but integrates them into the retrieved technical contexts. This approach leverages the LLM's linguistic capabilities to weave the labeled entities into the retrieved domain knowledge.

*3. Quality Inspection and Re-annotation*: To ensure the consistency of the augmented data, we introduce a verification step inspired by human annotation principles, but using LLM. A second prompt analyzes the newly generated instance to identify potentially unlabeled entities. The model receives the full definitions of all entity classes. Furthermore, to provide a representative context without exceeding the context window, we include a stratified random sample of up to 15 existing mentions for each entity class using a gazetteer (e.g., classes with fewer examples are fully included, while frequent classes are sampled). This allows the model to inspect and re-annotate the instance, thereby minimizing the likelihood of missing labels.

### 4.2 Datasets

We evaluated the data augmentation methods using four Portuguese datasets from different domains oil & gas, legal, beverage, and legislative consultations. All datasets are publicly available to ensure reproducibility. We utilized the original training, development, and testing splits provided by the authors. Table 1 summarizes the entity class, the total entity occurrences, and unique occurrences for each dataset.

**PetroGeoNER** (Moreira et al., 2025) is a dataset

[1] https://huggingface.co/datasets/hmoreira/da-datasets-propor2026

focused on the oil & gas domain. The sentences are from two main sources: The Brazilian Sedimentary Basins texts, comprising theses, dissertations, and Geosciences Bulletins from Petrobras. The second source is the Technical Bulletins and Reports from Petrobras. PetroGeoNER combines two datasets: GeoCorpus (Amaral, 2017) and PetroNER (Freitas et al., 2023). The combination of the datasets involved four steps: preprocessing, entity class mapping, entity matching, and domain expert verification.

**Ulysses-NER** (Costa et al., 2022; Nunes et al., 2024b) is a dataset of from legislative consultations. We used the *ST-Corpus*, which contains 790 sentences (*solicitações de trabalho*).

**CachacaNER** (Silva et al., 2023) is a dataset on the beverage domain, specifically the Brazilian popular drink known as 'Cachaça'. The dataset contains seventeen entity classes. Six of these classes are generic and traditional entities; the remaining eleven are related to the beverage domain.

**LeNER-BR** (Luz de Araujo et al., 2018; Nunes et al., 2024b) It is a dataset composed of 70 legal documents collected from the *Lei Maria da Penha* and several Brazilian Courts, including the *Supremo Tribunal Federal*, *Superior Tribunal de Justiça*, *Tribunal de Justica de Minas Gerais*, and *Tribunal de Contas da União*.

| Dataset | Entity | # Occurrences | # Unique |
|---|---|---|---|
| PetroGeoNER | BACIA | 2,889 | 378 |
| | CAMPO | 501 | 196 |
| | ESTRUTURA_FISICA | 1,900 | 198 |
| | FLUIDODATERRA | 1,621 | 56 |
| | FOSSEIS | 1,496 | 286 |
| | MINERAIS | 889 | 99 |
| | NAO_CONSOLID | 555 | 74 |
| | PALEOAMBIENTE | 2,121 | 190 |
| | POÇO | 508 | 258 |
| | ROCHA | 4,438 | 400 |
| | TEXTURA | 138 | 41 |
| | UNIDADE_CRONO | 5,753 | 921 |
| | UNIDADE_LITO | 2,216 | 555 |
| Ulysses-NER | DATA | 538 | 94 |
| | EVENTO | 19 | 9 |
| | FUNDAMENTO | 869 | 334 |
| | LOCAL | 731 | 202 |
| | ORGANIZACAO | 765 | 211 |
| | PESSOA | 1,086 | 258 |
| | PRODUTODELEI | 399 | 117 |
| CachacaNER | CARAC_SENSORIAL_AROMA | 935 | 265 |
| | CARAC_SENSORIAL_CONSISTÊNCIA | 278 | 50 |
| | CARAC_SENSORIAL_COR | 562 | 136 |
| | CARAC_SENSORIAL_SABOR | 906 | 260 |
| | CLASSIFICACAO_BEBIDA | 1,325 | 53 |
| | EQUIPAMENTO_DESTILACAO | 292 | 23 |
| | GRADUACAO_ALCOOLICA | 1,144 | 69 |
| | NOME_BEBIDA | 3,171 | 639 |
| | NOME_LOCAL | 4,232 | 441 |
| | NOME_ORGANIZACAO | 974 | 322 |
| | NOME_PESSOA | 743 | 296 |
| | PRECO | 885 | 409 |
| | RECIPIENTE_ARMAZENAMENTO | 991 | 90 |
| | TEMPO | 1,302 | 200 |
| | TEMPO_ARMAZENAMENTO | 1,210 | 107 |
| | TIPO_MADEIRA | 2,557 | 176 |
| | VOLUME | 2,532 | 170 |
| LeNER-BR | JURISPRUDENCIA | 1,392 | 942 |
| | LEGISLACAO | 2,368 | 1,526 |
| | LOCAL | 699 | 210 |
| | ORGANIZACAO | 3,049 | 716 |
| | PESSOA | 1,805 | 765 |
| | TEMPO | 1,568 | 981 |

Table 1: Statistics of the datasets

### 4.3 NER Model

We used XLM-RoBERTa-large (561M parameters) (Conneau et al., 2019) for training and evaluating the NER models with augmented data. The choice of the RoBERTa model over a BERT model (which has a Portuguese version) was motivated by the tokenizer. While RoBERTa relies on byte-level BPE, BERT uses WordPiece, which can result in poor tokenization of some named entities, particularly those with special characters between words.

We used the following hyperparameters: evaluation_strategy = epochs, learning_rate = 3e-5, num_train_epochs = 10, weight_decay = 0.01, train and eval batch_size = 16, optimizer = Adam. The remaining non-specified parameters follow default parameters suggested by HuggingFace[2]. For MELM, we used XLM-RoBERTa-base with num_train_epochs = 20, top-k sampling = 5. The remaining parameters followed the authors' (Zhou et al., 2022).

[2]https://huggingface.co/docs/transformers/en/tasks/token_classification

### 4.4 Evaluation Metrics

We computed the F1-score micro and macro to account for giving equal weights to instances and classes, respectively. To compute the F1 metrics, we used the SeqEval [3] library. The library is supported by HuggingFace and calculates the results based on the sequence of tags to each entity ('B-Tag' following 'I-Tag').

### 4.5 Experimental Setup

In all cases, the number of instances used for training was 1000 – what varied was the proportion of the *original* and *synthetic* (*i.e.,* produced by a data augmentation method) instances used for training. The development set and test set remained unchanged. Four training sets were generated for each dataset: 50/950, 100/900, 250/750, and 500/500. The naming convention is original/synthetic instances. For example, 100/900 indicates that the augmentation methods were applied to 100 original instances and generated 900 synthetic instances.

[3]https://github.com/chakki-works/seqeval

The seven augmentation techniques described in Section 4.1 were applied to each dataset. The only exception was RAG-DA, which we applied only to PetroGeoNER because it is the only dataset for which we have access to a collection of in-domain documents, the entity class definition used by the authors to annotate entities, and a gazetteer of possible entities. Additionally, a baseline run using only the original sample of instances (without augmentation) was trained for comparison. Finally, to serve as an upper-bound, we trained an "Original-Only" model on 1,000 original instances.

The experiments were conducted on the Atena supercomputer, which is available for machine learning and high-performance computing projects at CENPES Petrobras. The Atena has 224 CPU nodes, each equipped with 2 Intel Xeon 6248 processors and 384GB of RAM, and 24 GPU nodes, each with 8 Nvidia Tesla V100 GPUs with 32GB of RAM and 754GB of RAM. We used Python 3.12 due to its extensive library of NLP tools. For the LLM experiments, we used Azure AI with GPT4.1 (version 2024-08-06). For Back-translation, we used the deep-translator[4] library with Google Translate.

## 5 Results

This section presents our results, organized into quantitative and qualitative analyses.

### 5.1 Quantitative Analysis

We report the performance of various augmentation techniques on the test sets in Tables 2 and 3 with the results in terms of F1-micro and F1-macro. The tables summarize results for all combinations of dataset, augmentation method, and proportion of synthetic vs augmented instances. To mitigate the effects of randomness and provide a stable, scientifically valid estimate of how well the models perform, we repeated each experimental run three times and averaged the scores. This required 348 experimental runs in total.

The first conclusion we can draw from Table 2 is that almost all data augmentation techniques can improve over the baseline, where no augmentation is used. Overall, PP-LLM is the technique with the most "wins" (as evidenced by its highest number of scores in bold). SR and MR are drawn in a close second, which is somewhat surprising given their simplicity. SR performed best on PetroGeoNER and Ulysses-NER, while PP-LLM performed better on the LeNER-BR and CachacaNER datasets. RAG-DA outperformed the other techniques in the scenario with the smallest proportion of original instances. This is a good indication of its performance on low-data regimes, but further experiments in other datasets are needed to confirm its adequacy. MELM was the worst performer because this technique requires training an MLM to predict new entities for a specific class. Working with domain-specific and low-resource data makes it difficult for the model to generate quality examples of entities to replace the originals. As a result, the model often produces noisy entities that are not relevant to the domain.

Table 3 presents the results in terms of F1-macro, which gives the same weight to the classes. These scores are generally lower than the micro-averaged results due to the class imbalance present in the datasets. PP-LLM remains a top performer, especially in LeNER-BR and CachacaNER datasets. MR showed promising results (similar to the F1-micro scores), and CAMR demonstrated stronger performance in F1-macro scores compared with the F1-micro scores, suggesting that mention replacement strategies are particularly effective at stabilizing the recall of infrequent classes. RAG-DA achieved the best results in the lowest-data regime for PetroGeoNER, reinforcing its potential in highly specialized domains with scarce data. MELM did not outperform the other augmentation methods in any scenario, likely because it introduces noise when predicting entities for rare classes without sufficient context.

Compared with the Original-Only model trained on 1,000 original instances, results indicate that data augmentation reduces the performance gap in low-resource settings, but the model trained on the full dataset generally has an advantage. In the PetroGeoNER dataset, the best augmentation strategy (using 500 original instances) trails the Original-Only benchmark by approximately 3 percentage points in F1-micro and 2 points in F1-macro. This suggests difficulties in generating synthetic data that captures the specific semantic characteristics of expert-annotated examples in specialized domains. On the other hand, the results for Ulysses-NER and LeNER-BR show a different pattern, in which data augmentation serves as a substitute for manual annotation. For these datasets, models trained on 500 original instances plus augmentation achieved performance comparable to those trained on 1,000

[4] https://github.com/nidhaloff/deep-translator

original instances. Specifically, in Ulysses-NER, Back-Translation combined with 50% of the data surpassed the Original-Only baseline. These findings suggest that for domains with standard linguistic structures, generative augmentation can reduce the annotation volume by approximately half without compromising predictive quality.

| PetroGeoNER | | | | |
|---|---|---|---|---|
| | **50/950** | **100/900** | **250/750** | **500/500** |
| BASELINE | 0.00 ± 0.00 | 0.40 ± 0.01 | 0.62 ± 0.02 | 0.74 ± 0.01 |
| SR | 0.49 ± 0.04 | **0.61 ± 0.02** | **0.71 ± 0.01** | **0.76 ± 0.02** |
| MR | 0.49 ± 0.01 | 0.59 ± 0.02 | 0.70 ± 0.02 | 0.75 ± 0.01 |
| CAMR | 0.47 ± 0.02 | 0.58 ± 0.02 | 0.70 ± 0.01 | 0.75 ± 0.01 |
| MELM | 0.39 ± 0.02 | 0.52 ± 0.03 | 0.63 ± 0.01 | 0.72 ± 0.01 |
| BT | 0.46 ± 0.02 | 0.57 ± 0.02 | 0.70 ± 0.02 | **0.76 ± 0.01** |
| PP-LLM | 0.49 ± 0.04 | 0.60 ± 0.02 | 0.70 ± 0.01 | **0.76 ± 0.02** |
| RAG-DA | **0.54 ± 0.00** | 0.59 ± 0.01 | 0.69 ± 0.00 | 0.73 ± 0.02 |
| Original-Only (1k) | 0.79 ± 0.01 | | | |
| **Ulysses-NER** | | | | |
| | **50/950** | **100/900** | **250/750** | **500/500** |
| Baseline | 0.12 ± 0.03 | 0.41 ± 0.01 | 0.68 ± 0.01 | 0.78 ± 0.03 |
| SR | 0.62 ± 0.02 | **0.71 ± 0.03** | **0.77 ± 0.02** | 0.82 ± 0.03 |
| MR | **0.66 ± 0.02** | **0.71 ± 0.02** | 0.76 ± 0.03 | 0.80 ± 0.03 |
| CAMR | 0.62 ± 0.04 | **0.71 ± 0.01** | 0.75 ± 0.04 | 0.82 ± 0.02 |
| MELM | 0.62 ± 0.02 | 0.67 ± 0.02 | 0.72 ± 0.01 | 0.79 ± 0.01 |
| BT | 0.58 ± 0.03 | 0.70 ± 0.02 | 0.76 ± 0.03 | **0.83 ± 0.01** |
| PP-LLM | 0.58 ± 0.02 | 0.67 ± 0.02 | **0.77 ± 0.03** | 0.78 ± 0.01 |
| Original-Only (1k) | 0.82 ± 0.03 | | | |
| **LeNER-BR** | | | | |
| | **50/950** | **100/900** | **250/750** | **500/500** |
| Baseline | 0.20 ± 0.06 | 0.56 ± 0.08 | 0.77 ± 0.02 | 0.83 ± 0.02 |
| SR | 0.67 ± 0.04 | 0.76 ± 0.04 | 0.87 ± 0.01 | 0.88 ± 0.00 |
| MR | **0.71 ± 0.02** | **0.79 ± 0.02** | 0.87 ± 0.01 | **0.89 ± 0.00** |
| CAMR | 0.67 ± 0.02 | 0.80 ± 0.03 | 0.87 ± 0.01 | 0.88 ± 0.04 |
| MELM | 0.65 ± 0.03 | 0.73 ± 0.02 | 0.82 ± 0.02 | 0.88 ± 0.01 |
| BT | 0.67 ± 0.04 | 0.72 ± 0.05 | 0.86 ± 0.01 | 0.88 ± 0.02 |
| PP-LLM | **0.71 ± 0.01** | 0.76 ± 0.04 | **0.88 ± 0.00** | **0.89 ± 0.01** |
| Original-Only (1k) | 0.91 ± 0.02 | | | |
| **CachacaNER** | | | | |
| | **50/950** | **100/900** | **250/750** | **500/500** |
| BASELINE | 0.00 ± 0.00 | 0.39 ± 0.07 | 0.68 ± 0.10 | 0.83 ± 0.00 |
| SR | 0.63 ± 0.05 | 0.68 ± 0.03 | 0.80 ± 0.01 | **0.85 ± 0.00** |
| MR | 0.63 ± 0.05 | 0.69 ± 0.04 | **0.81 ± 0.01** | 0.84 ± 0.01 |
| CAMR | 0.65 ± 0.02 | 0.67 ± 0.06 | **0.81 ± 0.01** | **0.85 ± 0.01** |
| MELM | 0.58 ± 0.05 | 0.66 ± 0.04 | 0.79 ± 0.02 | 0.84 ± 0.01 |
| BT | 0.62 ± 0.03 | 0.67 ± 0.01 | 0.80 ± 0.01 | 0.84 ± 0.01 |
| PP-LLM | **0.66 ± 0.03** | **0.71 ± 0.01** | 0.80 ± 0.01 | 0.84 ± 0.01 |
| Original-Only (1k) | 0.87 ± 0.01 | | | |

Table 2: F1-micro Results for the data augmentation techniques under different proportions of original/synthetic instances

| PetroGeoNER | | | | |
|---|---|---|---|---|
| | **50/950** | **100/900** | **250/750** | **500/500** |
| BASELINE | 0.02 ± 0.04 | 0.18 ± 0.00 | 0.45 ± 0.03 | 0.64 ± 0.01 |
| SR | 0.35 ± 0.03 | 0.48 ± 0.05 | **0.62 ± 0.01** | **0.68 ± 0.03** |
| MR | 0.36 ± 0.02 | 0.47 ± 0.05 | 0.59 ± 0.02 | **0.68 ± 0.03** |
| CAMR | 0.35 ± 0.04 | 0.46 ± 0.04 | **0.62 ± 0.02** | 0.66 ± 0.01 |
| MELM | 0.26 ± 0.03 | 0.39 ± 0.03 | 0.53 ± 0.01 | 0.63 ± 0.01 |
| BT | 0.32 ± 0.02 | 0.46 ± 0.04 | 0.60 ± 0.02 | **0.68 ± 0.03** |
| PP-LLM | 0.34 ± 0.05 | 0.47 ± 0.04 | 0.60 ± 0.02 | 0.66 ± 0.01 |
| RAG-DA | **0.44 ± 0.02** | **0.52 ± 0.02** | 0.60 ± 0.01 | 0.65 ± 0.00 |
| Original-Only (1k) | 0.70 ± 0.01 | | | |
| **Ulysses-NER** | | | | |
| | **50/950** | **100/900** | **250/750** | **500/500** |
| Baseline | 0.1 ± 0.02 | 0.20 ± 0.14 | 0.57 ± 0.01 | 0.68 ± 0.04 |
| SR | 0.51 ± 0.02 | 0.60 ± 0.02 | **0.65 ± 0.02** | 0.72 ± 0.05 |
| MR | **0.55 ± 0.02** | **0.61 ± 0.02** | 0.64 ± 0.02 | 0.71 ± 0.01 |
| CAMR | 0.51 ± 0.02 | 0.60 ± 0.01 | 0.64 ± 0.04 | 0.71 ± 0.02 |
| MELM | 0.50 ± 0.02 | 0.57 ± 0.02 | 0.60 ± 0.01 | 0.70 ± 0.03 |
| BT | 0.47 ± 0.04 | 0.59 ± 0.01 | 0.64 ± 0.02 | **0.73 ± 0.03** |
| PP-LLM | 0.47 ± 0.01 | 0.57 ± 0.01 | **0.65 ± 0.03** | 0.69 ± 0.03 |
| Original-Only (1k) | 0.73 ± 0.03 | | | |
| **LeNER-BR** | | | | |
| | **50/950** | **100/900** | **250/750** | **500/500** |
| Baseline | 0.18 ± 0.09 | 0.55 ± 0.08 | 0.78 ± 0.02 | 0.84 ± 0.01 |
| SR | 0.66 ± 0.03 | 0.76 ± 0.03 | 0.86 ± 0.01 | 0.88 ± 0.00 |
| MR | **0.70 ± 0.04** | 0.79 ± 0.01 | 0.87 ± 0.01 | **0.89 ± 0.00** |
| CAMR | 0.67 ± 0.02 | **0.80 ± 0.03** | 0.86 ± 0.01 | 0.87 ± 0.03 |
| MELM | 0.65 ± 0.05 | 0.74 ± 0.02 | 0.82 ± 0.02 | 0.88 ± 0.01 |
| BT | 0.66 ± 0.07 | 0.73 ± 0.03 | 0.86 ± 0.02 | 0.88 ± 0.02 |
| PP-LLM | **0.70 ± 0.03** | 0.76 ± 0.04 | **0.88 ± 0.00** | **0.89 ± 0.01** |
| Original-Only (1k) | 0.91 ± 0.01 | | | |
| **CachacaNER** | | | | |
| | **50/950** | **100/900** | **250/750** | **500/500** |
| BASELINE | 0.00 ± 0.00 | 0.27 ± 0.07 | 0.59 ± 0.12 | 0.78 ± 0.00 |
| SR | 0.53 ± 0.07 | 0.62 ± 0.02 | 0.75 ± 0.02 | 0.81 ± 0.01 |
| MR | 0.53 ± 0.07 | 0.63 ± 0.03 | 0.76 ± 0.02 | 0.81 ± 0.01 |
| CAMR | **0.55 ± 0.05** | 0.61 ± 0.05 | **0.77 ± 0.01** | 0.81 ± 0.00 |
| MELM | 0.50 ± 0.07 | 0.59 ± 0.03 | 0.73 ± 0.02 | 0.80 ± 0.02 |
| BT | 0.53 ± 0.05 | 0.61 ± 0.01 | 0.75 ± 0.00 | 0.81 ± 0.02 |
| PP-LLM | **0.55 ± 0.05** | **0.65 ± 0.02** | 0.75 ± 0.00 | **0.82 ± 0.01** |
| Original-Only (1k) | 0.85 ± 0.02 | | | |

Table 3: F1-macro Results for the data augmentation techniques under different proportions of original/synthetic instances

## 5.2 Qualitative Analysis

To gain a clearer insight into the results produced by the augmentation methods, a qualitative analysis was conducted. A single specialist in the oil & gas domain analyzed a sample of 350 synthetic instances produced by the augmentation methods (50 instances per method), using PetroGeoNER. This specialist was involved in the guideline and annotation process for both PetroNER (Freitas et al., 2023) and PetroGeoNER (Moreira et al., 2025) and possesses substantial knowledge of the datasets. A sample of original instances (*i.e.,* that were not generated by the augmentation methods) was also shown to the human to serve as a baseline. The specialist was unaware of which methods generated which sentences.

Two aspects were analyzed: *fluency* and *entity span*. For fluency, the analysis consisted of rating the instances using a 4-point scale, in which 0) represents incoherent or unintelligible text; 1) indicates poor coherence or grammar; 2) denotes mostly cohesive text with minor errors; 3) represents fluent and natural text. For entity span quality, the assessment labels are defined as follows: 0) means all spans are incorrect; 1) indicates the span contains at least one relevant error; 2) represents correct spans with minor acceptable inaccuracies; 3) means all spans are correct and complete.

Table 4 presents the evaluation of fluency and entity span quality, performed by the domain specialist. MELM produced the poorest results in entity span quality, with only 26% of the sentences

generated achieving a span score of 3. Most instances (64%) fell into the low-quality categories of 0 and 1. This indicates that the masking strategy often destroys the semantic integrity of the named entities, as exemplified in Figure 2a.

RAG-DA presents a case of hallucination. Notably, RAG-DA achieved the highest fluency score among all methods, with 96% rated as score 3. It even surpassed the Baseline. In terms of span quality, 48% of the samples were rated at score 1, because they contain at least one entity in the sentence with a missing/erroneous class. This suggests that RAG-DA generates highly natural text, but frequently mislabels entities, introducing noise in the NER model as shown in Figure 2b, even when using a second prompt that includes the entity definition and examples.

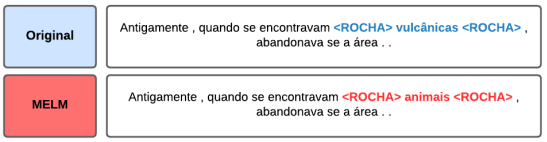
Conversely, the top-performing techniques in micro-F1, SR, and PP-LLM maintained high span consistency as shown in Figure 2c. SR achieved an 88% success rate with a score of 3 in span preservation, despite having lower fluency scores due to synonym replacement breaking sentence flow. This shows that label correctness is often more critical than perfect grammatical fluency for NER.

Table 4: Human evaluation: Fluency and Entity Span Quality (%)

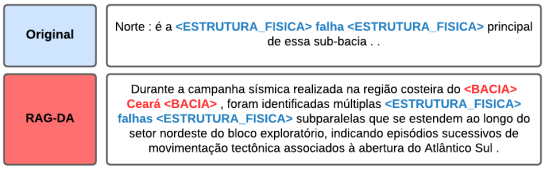
| | Fluency | | | | Entity Span | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **3** | **2** | **1** | **0** | **3** | **2** | **1** | **0** |
| Baseline | 90 | 10 | 0 | 0 | 86 | 2 | 12 | 0 |
| SR | 16 | 16 | 46 | 22 | 88 | 2 | 10 | 0 |
| MR | 70 | 28 | 2 | 0 | 78 | 10 | 12 | 0 |
| CAMR | 76 | 22 | 2 | 0 | 78 | 10 | 12 | 0 |
| MELM | 50 | 40 | 8 | 2 | 26 | 10 | 36 | 28 |
| BT | 50 | 42 | 6 | 2 | 84 | 2 | 14 | 0 |
| PP-LLM | 78 | 20 | 2 | 0 | 86 | 6 | 8 | 0 |
| RAG-DA | 96 | 4 | 0 | 0 | 46 | 6 | 48 | 0 |

## 6 Conclusion

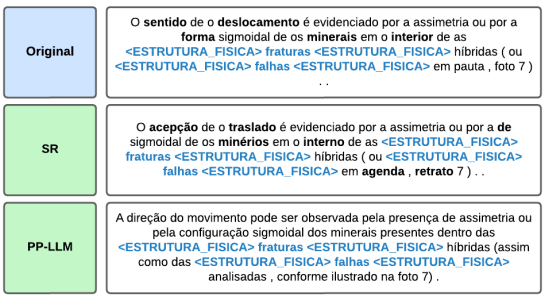
This work evaluated data augmentation techniques for NER in domain-specific scenarios, using four datasets in Portuguese. We augmented four subsets of each training set using rule-based methods, back-translation, pre-trained MLMs, and LLM-based paraphrasing. Our results show that almost all data augmentation techniques can improve compared to the baseline with few instances. The best results were achieved using PP-LLM, SR, and MR, respectively. This suggests that even simple and traditional techniques, such as SR and MR, can yield significant results and are relatively low-cost to implement.




(a) MELM error: Semantic violation within the <ROCHA> entity tag '(shown in red).




(b) RAG-DA limitation: Hallucination of new entities (<BACIA>) despite the fluent text.




(c) Successful generation: SR performs lexical substitution (in **bold**) and PP-LLM paraphrases the syntax while preserving entity integrity.

Figure 2: Qualitative comparison of generated instances.

## Limitations

Despite the positive results, our study has some limitations. We did not explore additional prompt variations in PP-LLM, and RAG-DA within PetroGeoNER. In RAG-DA, we only used the top-10 retrieved instances for augmentation. Additionally, we did not vary the number of entities in the second prompt, responsible for searching and annotating new entities. Also, RAG-DA was applied only to PetroGeoNER, as it is the only dataset for which we have access to a segmented collection of documents in the domain. The qualitative analysis in the PetroGeoNER was conducted by a single specialist in the oil % gas domain, which introduces a potential for subjective bias. To mitigate this, a single-blind protocol was implemented, ensuring the expert remained unaware of the origin of each

instance. While a multi-evaluator setup is preferred for measuring agreement, the limited availability of experts in the oil & gas domain made this approach unfeasible within the current scope. Nevertheless, the expert is familiar with the dataset, ensuring a consistent evaluation.

## Future Work

As future work, we plan to conduct additional experiments with RAG-DA. We aim to vary the prompts used, expand our list of entities and definitions. While the text generation has shown good performance, we have identified entity mislabeling as a major weakness. To address this issue, we intend to explore a filtering mechanism to enhance the quality of the generated output, such as implementing TARS presented in the FLAIR library. Additionally, we plan to explore other data augmentation techniques, such as CoDa (Evuru et al., 2024), EnTDA (Hu et al., 2022), and LLM-DA (Ye et al., 2024).

## Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq, and Cenpes/Petrobras.

## References

Daniela Oliveira Ferreira Amaral. 2017. *Reconhecimento de entidades nomeadas na Área da geologia: bacias sedimentares brasileiras*. Ph.D. thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pages 2318–2335.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Rosimeire Costa, Hidelberg Oliveira Albuquerque, Gabriel Silvestre, Nádia Félix F. Silva, Ellen Souza, Douglas Vitório, Augusto Nunes, Felipe Siqueira, João Pedro Tarrega, João Vitor Beinotti, Márcio de Souza Dias, Fabíola S. F. Pereira, Matheus Silva, Miguel Gardini, Vinicius Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2022. Expanding UlyssesNER-Br named entity recognition corpus with informal user-generated text. In *Progress in Artificial Intelligence*, pages 767–779.

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

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057.

Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, Utkarsh Tyagi, Dinesh Manocha, and 1 others. 2024. Coda: Constrained generation based data augmentation for low-resource nlp. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3754–3769.

Cláudia Freitas, Elvis Souza, Maria Clara Castro, Tatiana Cavalcanti, Patricia Ferreira da Silva, and Fábio Corrêa Cordeiro. 2023. Recursos linguísticos para o pln específico de domínio: o petrolês. *Linguamática*, 15(2):51–68.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and Philip S Yu. 2022. Entda: Entity-to-text based data augmentation approach for named entity recognition tasks. *arXiv preprint arXiv:2210.10343*.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and 1 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7199–7208.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study. Working paper or preprint.

Hyunjae Kim, Jaehyo Yoo, Seunghyun Yoon, Jinhyuk Lee, and Jaewoo Kang. 2022. Simple questions generate named entity recognition datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6220–6236.

Lucas Lima de Oliveira, Regis Kruel Romeu, and Viviane Pereira Moreira. 2021. Regis: A test collection for geoscientific documents in portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2363–2368.

Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323.

Higor Moreira, Patricia Ferreira da Silva, Renata Vieira, and Viviane Moreira. 2025. Petrogeoner: A refined and unified dataset for ner in the oil & gas domain. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 259–271. SBC.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam mickiewicz university at wmt 2022: Ner-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326–334.

Rafael Oleques Nunes, Dennis Giovani Balreira, André Suslik Spritzer, and Carla Maria Dal Sasso Freitas. 2024a. A named entity recognition approach for Portuguese legislative texts using self-learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 290–300.

Rafael Oleques Nunes, André Susliz Spritzer, Carla Maria Dal Sasso Freitas, and Dennis Giovani Balreira. 2024b. Reconhecimento de entidades nomeadas e vazamento de dados em textos legislativos: Uma reavaliação da literatura. *Linguamática*, 16.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 65–72.

Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. 2023. When and how to paraphrase for named entity recognition? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7052–7087.

Priscilla Silva, Arthur Franco, Thiago Santos, Mozar Brito, and Denilson Pereira. 2023. Cachacaner: a dataset for named entity recognition in texts about the cachaça beverage. *Lang. Resour. Eval.*, 58(4):1315–1333.

Sihan Song, Furao Shen, and Jian Zhao. 2024. Ropda: robust prompt-based data augmentation for low-resource named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19017–19025.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Arthur Elwing Torres, Edleno Silva de Moura, Altigran Soares da Silva, Mario A Nascimento, and Filipe Mesquita. 2024. An experimental study on data augmentation techniques for named entity recognition on low-resource domains. *arXiv preprint arXiv:2411.14551*.

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

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

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