Named Entity Recognition using context similarity data augmentation

Ilaria Bartolini[1,†], Angelo Chianese[2,†], Vincenzo Moscato[2,3,†], Marco Postiglione[2,†], Giancarlo Sperlí[2,3,*,†] and Andrea Vignali[2,†]

[1] *Alma Mater Studiorum, University of Bologna, Via Zamboni 33, 40126, Bologna, Italy*

[2] *University of Naples Federico II, Dept. of Electrical Engineering and Information Technology (DIETI), Via Claudio 21, 80125, Naples, Italy*

[3] *CINI - ITEM National Lab, Complesso Universitario Monte S.Angelo, Naples, Italy*

Abstract

This paper is an extended abstract of a recent work, in which we introduce COSINER, a novel approach to enhancing Named Entity Recognition (NER) tasks through data augmentation. Unlike traditional methods that risk introducing noise, COSINER leverages context similarity to substitute entity mentions with more contextually appropriate ones, yielding superior performance in limited-data scenarios. Experimental results demonstrate COSINER's effectiveness over existing baselines, with computational times comparable to basic augmentation methods and superior to pre-trained model-based approaches.

Keywords

Named Entity Recognition, Data Augmentation, Similarity Learning, Few Shot Learning.

1. Introduction

Named Entity Recognition (NER) is a crucial component of natural language processing (NLP), which aims to understand and process natural language for various tasks like sentiment analysis, text classification, and machine translation. NER's objective is to identify and classify entity mentions (e.g., person, organization, disease) in unstructured text. It serves as a foundational step in several applications (e.g., machine translation or information discovery). NER identifies and extracts relevant items from unstructured text, like diseases or genes in medical records, serving as a crucial initial step for various applications like knowledge graphs and Q/A bots.

NER model training typically requires vast annotated data, but obtaining quality annotations, especially in specialized domains, is time-consuming and costly. Few-shot learning, exploring unique strategies for constrained datasets, addresses this challenge, particularly in fields lacking readily available domain specialists.

Data augmentation, a method to enhance dataset size by generating additional samples, is commonly used to address data scarcity. In Natural Language Processing (NLP), techniques

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding author.

[†] These authors contributed equally.

ilaria.bartolini@unibo.it (I. Bartolini); angelo.chianese@unina.it (A. Chianese); vincenzo.moscato@unina.it (V. Moscato); marco.postiglione@unina.it (M. Postiglione); giancarlo.sperli@unina.it (G. Sperlí); andrea.vignali@unina.it (A. Vignali)

like word replacement [1], random deletion [2], word position swap [3] and generative models [4] are popular. However, token-level classification in NER becomes more and more complex using traditional augmentation, requiring an increasing effort in analyzing possible approaches in this area [5]. Recent efforts explore transfer learning [6] and Masked Language Models (MLM) [7] to alleviate label misalignment and augment datasets effectively. Moreover, while data augmentation holds promise, the current manipulation methods often generate noisy and misclassified samples. The added data may contain syntactic or semantic errors, leading to inaccuracies in classification.

To address this challenge, we present our method, *COntext SImilarity-based data augmentation for NER (COSINER)* [8], which utilizes similarity metrics to generate augmented examples that closely resemble real context. Our approach introduces a context-based mention replacement technique, substituting mentions in input data with entities from an Entity Lexicon that are contextually appropriate. In this paper, which is an extended abstract of our previous work [9], our contribution consists of the development of COSINER and an extensive evaluation across three prominent biomedical benchmark datasets that demonstrate COSINER's superiority over existing methods, highlighting its general applicability beyond the biomedical domain. Notably, COSINER's effectiveness is attributed to its ability to improve performance primarily through top-ranked samples, reducing reliance on large augmented datasets and enhancing computational efficiency.

2. Methodology

COSINER utilizes mention replacement to expand the initial training set, a technique previously explored by Dai et al. [5]. While their method randomly substitutes entities within sentences using a binomial distribution, we introduce a systematic approach centered on similarity, where entity mentions are replaced with counterparts closely matching in syntax, semantics, and context. Despite the quadratic time complexity of our methodology, equal to $O(mn^2)$ for computing cosine similarity between n embeddings (with a size of m), the time spent generating new examples remains insignificant. Figure 1 provides an overview of our methodical flow, elaborated further in subsequent sections.

Lexicon generation In the training set, each entity, referred to as a $concept$, needs to be collected for replacement purposes. A $concept$ can comprise one or a group of words, and we also record the frequency of each word's appearance in the training set within the Lexicon $C_{concept}$. The size of the Lexicon varies depending on the number of mentions in the dataset. It is significant to emphasize that although the size of the Lexicon influences the speed of computing similarity values between entity pairs, this influence is not a constraint, particularly as we conduct experiments in few-shot scenarios.

Embeddings extraction In order to calculate entities similarities, it's imperative to establish a comprehensive representation ($V_{concept}$) for all Lexicon concepts, which serves as viable input for our predictive model. We employ a pre-trained language model as a feature extractor [10, 11] to process each phrase containing a given mention from the Lexicon, mapping each token to

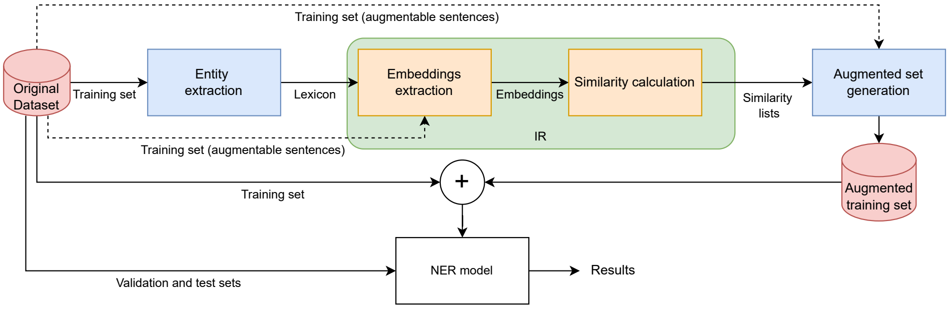

Figure 1: COSINER methodological flow: (1) Original training set is utilized to create a Lexicon of all entities. (2) Entities are embedded into a vector space based on sentences containing at least on mention, (3) Similarity scores between pairs of embeddings are computed to establish connections between each entity and the related ranked list similar entities, (4) An augmented training set is formulated, (5) The model undergoes training employing both the original and augmented training datasets.

its word embedding $V_{context}$ (i.e. an array of numerical features representing the token in its context). In cases where mentions consist of multiple tokens, $V_{context}$ is obtained by averaging the word embeddings of all tokens.

Upon retrieving $V_{context}$, the numerical representation of the concept $V_{concept}$ is updated using the formula:

$$V_{concept} = V_{concept} + lr \cdot (1 - \text{sim}) \cdot V_{context},$$

where lr denotes a regularization term determined by the inverse of the frequency of a mention across the entire dataset, and sim represents the cosine similarity between $V_{concept}$ and $V_{context}$. Initially, $V_{concept}$ is set to the $V_{context}$ value of the first sentence where the mention appears.

Similarity computation We calculate the cosine similarity between the embeddings $V_{concept}$ of every pair of entities in the Lexicon to derive a ranked list of similarity scores $z_{ij} = \text{sim}(V^{i}_{concept}, V^{j}_{concept})$ associated with each Lexicon entry. We define two ranking criteria:

1)Maximum (descending order): Prioritizing concepts with the highest relatedness at the top of the list. This approach facilitates the generation of realistic augmented samples that uphold contextual consistency within sentences.

2)Minimum (ascending order): By initially considering the least similar entities, we encompass samples farthest from the knowledge boundary. This inclusion enables the recognition and accurate classification of extreme cases.

Augmented set generation The augmented set is constructed from all sentences featuring at least one mention. Each sentence is assigned a similarity value s_{m}, , which is computed as the mean of entity similarity scores $z_{ij}$for the additional entities present within the sentence. We employ two strategies:

1) Local Augmentation: Each sentence results in the generation of k new samples, ensuring the contribution of every training instance to the augmented set.

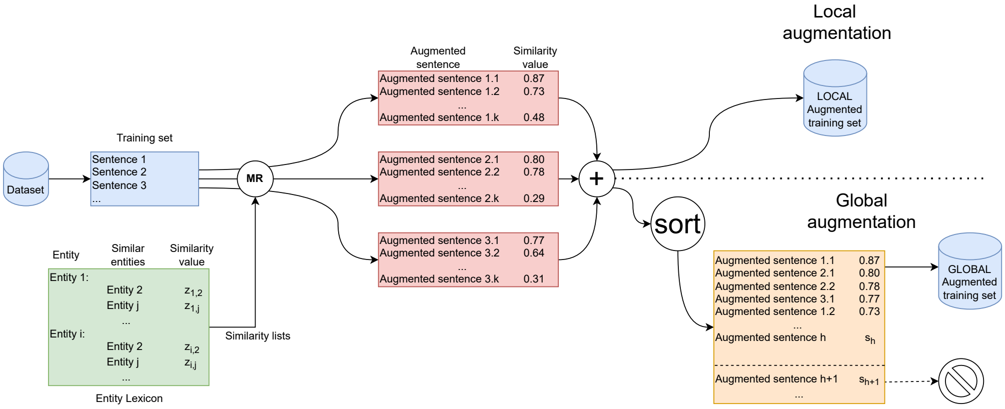
Local
augmentation
Augmented
sentence
Similarity
value
Augmented sentence 1.1 0.87
Augmented sentence 1.2 0.73
...
Augmented sentence 1.k 0.48
Training set
Dataset
Sentence 1
Sentence 2
Sentence 3
...
MR
Augmented sentence 2.1 0.80
Augmented sentence 2.2 0.78
...
Augmented sentence 2.k 0.29
+
LOCAL
Augmented
training set
sort
Global
augmentation
Entity
Similar
entities
Similarity
value
Entity 1:
Entity 2 $z_{1,2}$
Entity j $z_{1,j}$
...
Entity i:
Entity 2 $z_{i,2}$
Entity j $z_{i,j}$
...
Entity Lexicon
Similarity lists
Augmented sentence 3.1 0.77
Augmented sentence 3.2 0.64
...
Augmented sentence 3.k 0.31
Augmented sentence 1.1 0.87
Augmented sentence 2.1 0.80
Augmented sentence 2.2 0.78
Augmented sentence 3.1 0.77
Augmented sentence 1.2 0.73
...
Augmented sentence h s_h
Augmented sentence h+1 s_{h+1}
...
GLOBAL
Augmented
training set

Figure 2: COSINER Augmentation strategies. Both local and global strategies start by generating k augmented sentences per phrase with at least one mention, using Mention Replacement (MR) and similarity lists from the training set. Then, each augmented example is assigned a sentence similarity value s_m. In the local strategy, the new training set comprises all augmented examples. In the global approach, a new list is generated, arranging examples based on their s_m, values, and the top h sentences are selected for the augmented training set.

2) Global Augmentation: Similar to the previous strategy, k new samples are generated for each sentence. Subsequently, we rank all newly generated sentences in a single list based on their similarity value s_m and select the first h elements.

In Figure 2 we emphasize the distinctions between the two strategies.

NER model training We adhere to the IOB2 scheme for the NER token-classification task [12]. The original training dataset and the augmented samples are fed into a Transformer network backbone [10, 11]. Model parameters undergo optimization via cross-entropy minimization.

3. Experimental Analysis

We conduct training and evaluation on three renowned benchmark datasets sourced from biomedical articles: i) NCBI-Disease [13]: Comprising 793 PubMed abstracts, with 6,881 *disease* entities, ii) BC5CDR [14]: Comprising 1,500 PubMed articles, containing 15,935 *chemical* mentions, and BC2GM [15]: Comprising 20,000 sentences extracted from PubMed abstracts, involving 20,702 *gene* entities.

We delineate three distinct few-shot scenarios, each characterized by the percentage of samples drawn from the available corpora employed in implementing our methods: specifically, 2%, 5%, and 10%. Subsequently, we present all experimental findings within these aforementioned few-shot scenarios. Dataset statistics and few-shot scenarios details are summarized in Table 1.

3.1. Hyperparameter tuning

Table 2 presents results achieved using various parameter configurations for similarity computation (Maximum vs Minimum) and augmented set generation (Local vs Global), as discussed

Table 1
Statistics of the dataset used.

			Dataset splits			Few-shot size		
Dataset	**Entity type**	**N. Annotations**	*Train*	*Dev*	*Test*	*2%*	*5%*	*10%*
NCBI-disease	Disease	6881	5425	924	941	108	271	542
BC5CDR	Chemical	15411	4561	4582	4798	91	228	456
BC2GM	Gene	20703	12575	2520	5039	251	628	1257

Table 2
Exploration of optimal strategies for COSINER.

Dataset size	Similarity	Strategy	NCBI Disease	BC5CDR	BC2GM
2%	Maximum	Global	0.688 ± 0.077	0.83 ± 0.023	0.658 ± 0.036
	Minimum	Global	0.683 ± 0.086	0.823 ± 0.032	0.652 ± 0.027
	Maximum	Local	0.689 ± 0.088	$\mathbf{0.832} \pm \mathbf{0.022}$	$\mathbf{0.665} \pm \mathbf{0.038}$
	Minimum	Local	$\mathbf{0.692} \pm \mathbf{0.081}$	0.824 ± 0.015	0.659 ± 0.049
5%	Maximum	Global	$\mathbf{0.765} \pm \mathbf{0.035}$	0.858 ± 0.023	0.717 ± 0.007
	Minimum	Global	0.756 ± 0.028	0.853 ± 0.029	0.713 ± 0.009
	Maximum	Local	0.76 ± 0.031	$\mathbf{0.863} \pm \mathbf{0.042}$	$\mathbf{0.726} \pm \mathbf{0.022}$
	Minimum	Local	0.764 ± 0.041	0.86 ± 0.031	0.714 ± 0.007
10%	Maximum	Global	0.807 ± 0.038	0.88 ± 0.018	0.76 ± 0.02
	Minimum	Global	0.807 ± 0.029	0.873 ± 0.016	0.761 ± 0.012
	Maximum	Local	$\mathbf{0.816} \pm \mathbf{0.066}$	$\mathbf{0.882} \pm \mathbf{0.007}$	$\mathbf{0.767} \pm \mathbf{0.023}$
	Minimum	Local	0.807 ± 0.038	0.876 ± 0.016	0.76 ± 0.009

in Section 2. As anticipated, employing Maximum similarity computation generally yields superior performance, as augmented samples are plausible and closer to the test distribution. Nevertheless, the notable performance achieved with the Minimum configuration suggests that at times, considering "distant" entities may prove beneficial in expanding the NER model's scope. Regarding augmented set generation, the Local criterion typically outperforms, owing to its augmentation of *all* sentences in the original dataset. In summary, it's noteworthy that Maximum local emerges as the most favorable overall strategy.

When creating an augmented dataset, the quantity of augmented samples is a crucial parameter to consider. Therefore, we conducted experiments using three distinct *budgets* for the augmented set: small (100 samples), medium (300 samples), and large (500 samples).

Figure 3 illustrates the results obtained across the three benchmark datasets. Due to the similarity-based approach, which prioritizes the most informative examples in the top-ranked positions, there is minimal discrepancy observed when using higher budgets.

4. Result

We contrast our top-performing results with baselines drawn from current literature [5], as follows:

- No Augmentation: Results obtained using the original training set assessed with a BERT

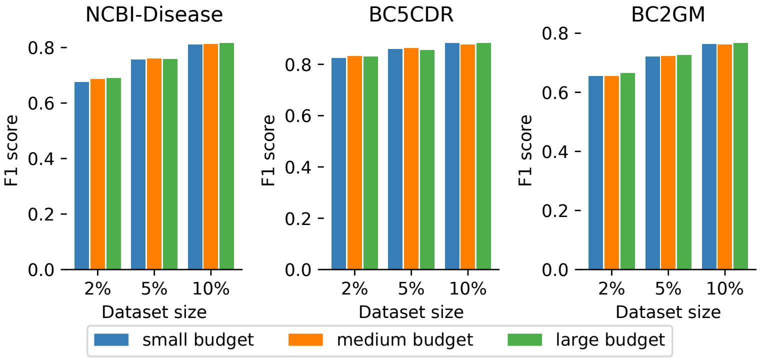

Figure 3: Comparison of outcomes among the small, medium, and large budget allocations for local augmentation strategy using the maximum similarity technique.

or BioBERT pre-trained model.

- Mention Replacement (MR): Random selection of a mention from the original training set with the same entity type for each mention in the instance.
- Label-wise Token Replacement (LwTR): Randomly decide whether to replace each word within a sentence with any other word in the dataset sharing the same label.
- Synonym Replacement (SR): Employ a binomial distribution to determine whether to replace each word within a sentence with a synonym from WordNet [16].
- Masked Entity Language Modeling (MELM): Employ a pre-trained RoBERTa model as MLM to predict masked tokens within the training set. Subsequently, utilize the augmented dataset to train a BERT model.
- Cross-Domain Named Entity Recognition (style_NER): Utilize additional data to transfer knowledge from a source domain to a target domain.

Table 3 compares the precision, recall, and F1 scores of the baselines with our method, which achieved the best outcomes for each dataset and related scenarios. Results indicate that COSINER outperforms most baselines across scenarios and datasets. While it consistently ensures the highest recall scores, signifying the system's ability to identify more entity mentions present in the corpus, COSINER falls short of SR in terms of precision in some scenarios. This suggests that the augmentation process may generate a higher number of false positives.

5. Conclusion

In this study, we have employed a *context similarity*-based approach to generate augmented data, aiming to enhance the performance of NER tasks while mitigating the adverse effects of noisy and mislabeled data commonly encountered with existing techniques.

Our experiments conducted in the medical domain, where data augmentation is particularly crucial, underscore the efficacy of our method. We have demonstrated its superiority over several state-of-the-art baselines, achieving comparable or improved execution times.

Looking ahead, our approach holds promise for integration with complementary techniques beyond Mention Replacement. Future investigations will explore its applicability across diverse

Table 3
Comparative results between baselines and our best strategy.

Size	Method	NCBI-Disease			BC5CDR			BC2GM		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
2%	No augmentation	0.430 ±0.193	0.403 ±0.169	0.461 ±0.225	0.628 ±0.179	0.625 ±0.185	0.634 ±0.215	0.510 ±0.036	0.448 ±0.015	0.592 ±0.082
	No augmentation (BioBERT)	0.651 ±0.122	0.619 ±0.100	0.688 ±0.162	0.792 ±0.067	0.799 ±0.058	0.786 ±0.110	0.644 ±0.031	0.600 ±0.057	0.695 ±0.022
	MR	0.666 ±0.084	0.626 ±0.1	0.710 ±0.067	0.813 ±0.032	0.806 ±0.06	0.822 ±0.071	0.640 ±0.02	0.593 ±0.062	0.696 ±0.049
	LwTR	0.677 ±0.101	0.637 ±0.125	0.723 ±0.08	0.828 ±0.019	0.808 ±0.052	0.850 ±0.075	0.642 ±0.037	0.591 ±0.059	0.704 ±0.019
	SR	**0.692** ±0.103	**0.649** ±0.132	0.742 ±0.084	0.813 ±0.032	0.811 ±0.085	0.835 ±0.064	0.662 ±0.033	**0.619** ±0.058	0.710 ±0.029
	MELM	0.578 ±0.038	0.545 ±0.046	0.615 ±0.041	0.754 ±0.019	0.719 ±0.047	0.795 ±0.036	0.566 ±0.011	0.504 ±0.006	0.647 ±0.027
	style_NER	0.581 ±0.061	0.537 ±0.076	0.636 ±0.067	0.752 ±0.018	0.713 ±0.041	0.796 ±0.016	0.581 ±0.003	0.540 ±0.018	0.631 ±0.025
	COSINER (ours)	0.689 ±0.088	0.629 ±0.078	**0.764** ±0.11	**0.832** ±0.022	**0.814** ±0.08	**0.853** ±0.066	**0.665** ±0.038	0.614 ±0.065	**0.724** ±0.025
5%	No augmentation	0.621 ±0.055	0.572 ±0.088	0.68 ±0.054	0.757 ±0.039	0.73 ±0.062	0.788 ±0.121	0.612 ±0.022	0.563 ±0.03	0.671 ±0.077
	No augmentation (BioBERT)	0.735 ±0.041	0.706 ±0.051	0.767 ±0.062	0.850 ±0.02	0.836 ±0.01	0.865 ±0.048	0.711 ±0.012	0.680 ±0.028	0.744 ±0.019
	MR	0.743 ±0.048	0.712 ±0.045	0.776 ±0.059	0.849 ±0.021	0.834 ±0.03	0.865 ±0.026	0.713 ±0.006	0.675 ±0.02	0.755 ±0.024
	LwTR	0.743 ±0.072	0.710 ±0.066	0.780 ±0.086	0.860 ±0.039	**0.846** ±0.017	0.876 ±0.067	0.699 ±0.012	0.660 ±0.024	0.742 ±0.029
	SR	0.758 ±0.044	0.719 ±0.049	0.800 ±0.049	0.858 ±0.03	0.841 ±0.033	0.875 ±0.067	0.719 ±0.011	0.684 ±0.023	0.758 ±0.019
	MELM	0.678 ±0.034	0.647 ±0.037	0.713 ±0.035	0.800 ±0.020	0.769 ±0.043	0.835 ±0.030	0.629 ±0.010	0.587 ±0.010	0.677 ±0.021
	style_NER	0.687 ±0.040	0.662 ±0.038	0.714 ±0.042	0.805 ±0.015	0.793 ±0.020	0.818 ±0.020	0.640 ±0.005	0.594 ±0.018	0.695 ±0.017
	COSINER (ours)	**0.76** ±0.031	**0.721** ±0.029	**0.805** ±0.057	**0.863** ±0.042	0.839 ±0.04	**0.892** ±0.058	**0.726** ±0.022	**0.692** ±0.013	**0.767** ±0.03
10%	No augmentation	0.712 ±0.056	0.670 ±0.065	0.76 ±0.046	0.804 ±0.032	0.781 ±0.046	0.829 ±0.054	0.669 ±0.019	0.626 ±0.026	0.720 ±0.045
	No augmentation (BioBERT)	0.791 ±0.028	0.760 ±0.024	0.825 ±0.036	0.875 ±0.013	0.858 ±0.02	0.892 ±0.028	0.759 ±0.017	0.734 ±0.019	0.786 ±0.016
	MR	0.794 ±0.018	0.761 ±0.025	0.831 ±0.019	0.874 ±0.034	0.859 ±0.038	0.889 ±0.04	0.754 ±0.01	0.724 ±0.013	0.787 ±0.032
	LwTR	0.789 ±0.023	0.756 ±0.034	0.825 ±0.036	0.882 ±0.017	**0.870** ±0.021	0.893 ±0.022	0.741 ±0.012	0.712 ±0.023	0.772 ±0.025
	SR	0.803 ±0.033	0.776 ±0.033	0.832 ±0.053	**0.883** ±0.018	0.862 ±0.016	0.904 ±0.021	0.763 ±0.012	0.738 ±0.019	0.788 ±0.02
	MELM	0.740 ±0.017	0.712 ±0.019	0.770 ±0.016	0.841 ±0.010	0.824 ±0.013	0.858 ±0.019	0.685 ±0.006	0.647 ±0.008	0.728 ±0.010
	style_NER	0.745 ±0.014	0.738 ±0.018	0.752 ±0.014	0.838 ±0.012	0.829 ±0.025	0.847 ±0.021	0.694 ±0.004	0.660 ±0.009	0.732 ±0.010
	COSINER (ours)	**0.816** ±0.066	**0.780** ±0.014	**0.856** ±0.068	0.882 ±0.007	0.861 ±0.022	**0.914** ±0.02	**0.767** ±0.023	**0.738** ±0.026	**0.798** ±0.015

contexts and with various entity types, fostering a deeper understanding of its potential and versatility.

Acknowledgments

This work has been funded by the project NextGenerationEU via PNRR - DM 352 (CUP: E66G22000400009). We acknowledge financial support from the PNRR project "Future Artificial Intelligence Research (FAIR)" – CUP E63C22002150007

References

[1] H. Cai, H. Chen, Y. Song, C. Zhang, X. Zhao, D. Yin, Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6334–6343.

[2] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 6382–6388.

[3] J. Min, R. T. McCoy, D. Das, E. Pitler, T. Linzen, Syntactic data augmentation increases robustness to inference heuristics, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2339–2352.

[4] K. M. Yoo, Y. Shin, S.-g. Lee, Data augmentation for spoken language understanding via joint variational generation, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 7402–7409.

[5] X. Dai, H. Adel, An analysis of simple data augmentation for named entity recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3861–3867.

[6] S. Chen, G. Aguilar, L. Neves, T. Solorio, Data augmentation for cross-domain named entity recognition, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5346–5356.

[7] R. Zhou, X. Li, R. He, L. Bing, E. Cambria, L. Si, C. Miao, Melm: Data augmentation with masked entity language modeling for low-resource ner, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2251–2262.

[8] I. Bartolini, V. Moscato, M. Postiglione, G. Sperlì, A. Vignali, Cosiner: Context similarity data augmentation for named entity recognition, in: International Conference on Similarity Search and Applications, Springer, 2022, pp. 11–24.

[9] I. Bartolini, V. Moscato, M. Postiglione, G. Sperlì, A. Vignali, Data augmentation via context similarity: An application to biomedical named entity recognition, Information Systems 119 (2023) 102291.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[12] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, in: Natural language processing using very large corpora, Springer, 1999, pp. 157–176.

[13] R. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, 2014.

[14] J. Li, Y. Sun, R. Johnson, D. Sciaky, C. Wei, R. Leaman, A. Davis, C. Mattingly, T. Wiegers, Z. Lu, BioCreative V CDR task corpus: a resource for chemical disease relation extraction, 2016.

[15] L. Smith, L. Tanabe, R. Ando et al., The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge, 2008.

[16] G. A. Miller, Wordnet: A lexical database for english, Commun. ACM 38 (1995) 39–41.