Self-Adapted Entity-Centric Data Augmentation for Discontinuous Named Entity Recognition

Anonymous ACL submission

Abstract

Named Entity Recognition (NER) is a critical task in natural language processing, particularly challenging in identifying discontinuous entities. This study is the first to explore the application of image data augmentation techniques in the preprocessing phase for discontinuous entity recognition, aiming to overcome the limitations of traditional text segmentation methods. Through experiments, we found that traditional sentence segmentation methods might lead to incorrect segmentation of cross-sentence discontinuous entities, affecting the accuracy of model training and entity recognition. To address this, we introduced a new preprocessing strategy that combines graphic cropping, scaling, and padding techniques to improve the model's ability to recognize discontinuous entities. Experiments on three benchmark datasets, CADEC, ShARe13, and ShARe14, demonstrated that our preprocessing method increased the F1 scores of two state-of-the-art grid models by approximately 1% to 2.5%, proving the effectiveness of this method. [1]

1 Introduction

Named Entity Recognition (NER) is a crucial task in the field of natural language processing, aiming to locate and classify named entities into predefined categories from text. In recent years, NER research has been subdivided into various task types, including flat (Lample et al., 2016; Strubell et al., 2017), overlapping (Yu et al., 2020; Shen et al., 2021), and discontinuous (Dai et al., 2020a; Li et al., 2021) NER tasks, with discontinuous NER seen as the most challenging among them. As shown in Figure 1, entities in the sentence are discontinuous; their representation might be nested, overlapping, or even span multiple sentences. This diversity significantly increases the difficulty of the recognition task.

[1]The code is publicly available at https://github.com/fang1204/SEDA.git

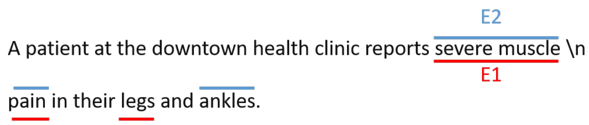

Figure 1: Example showing two discontinuous entities

Therefore, preprocessing methods for discontinuous entities are particularly tricky, requiring special consideration to maintain entity integrity after text processing. A literature review reveals that most studies focus on enhancing model architectures or developing related auxiliary loss functions (Tang et al., 2018; Lu and Roth, 2015; Katiyar and Cardie, 2018; Wang and Lu, 2018; Yan et al., 2021; Fei et al., 2021; Wang et al., 2020a; Yu et al., 2020; Shen et al., 2021), with few discussing preprocessing methods. Among these, we notice that many research teams adopted the preprocessing script proposed by Dai et al. (2020b) to segment datasets, including specific character retention and tokenization of sentences. In terms of sentence breaking, a simple newline character is used as the division standard, for example in Figure 1, the sentence is divided into "Sentence one: A patient at the downtown health clinic reports severe muscle" and "Sentence two: pain in their legs and ankles.", where the two discontinuous entities will be separated. Thus, when cross-sentence discontinuous entities occur, this method will fail to correctly identify the entity, thereby affecting the model's training performance.

Recently, with various model architecture innovations, Li et al. (2022) introduced a unified model to address different NER tasks, called the Unified Word-Word Framework (Word2NER), transforming discontinuous named entity recognition into a problem of relationships between words, and calculating lexical relations through a grid structure to identify entity boundaries and words. This framework has shown excellent performance in NER tasks. This method starts from a graphical perspective, incorporating CNN-related technologies, prompting us to consider whether we could adopt graphical preprocessing methods for the domain of discontinuous entity recognition. Thus, this paper proposes a preprocessing method combining image data enhancement techniques and a self-learning strategy to further enhance model performance. Our adopted image data enhancement techniques (Mikołajczyk and Grochowski, 2018; Connor and Khoshgoftaar, 2019) include graphical cropping, scaling, and filling intervals, to improve model predictions.

Visual explanation of image data augmentation, as shown in Figure 2: assuming the main task is to identify a cat, the target object inside the red frame can be recognized, and the blue area fills out the supplemental intervals. We apply this concept to the grid model, as shown below Figure 2, in the sentence "I do experience stomach pain from time to time," where the entity "stomach pain" corresponds to the key part in the red area, while the surrounding blue blocks serve as supplemental intervals, extending the key area, with other areas serving as background information. Overall, our research contributions can be summarized in three points:

- To our knowledge, we are the first to propose the application of image enhancement methods to the field of discontinuous entity recognition.
- Our proposed method uses various image enhancement techniques to overcome the limitations of grid models.
- Finally, we demonstrate the generalizability and effectiveness of our method through different datasets and grid models.

2 Related work

2.1 Grid-tagging method

Recently, grid-tagging methods have shown promising performance in the domain of discontinuous Named Entity Recognition (NER). These methods use architectures such as CLN(Liu et al., 2021) and convolutional neural networks(Yu and Koltun, 2016) to predict the relationships between words. The earliest related research can be traced back to Wang et al. (2020b), who proposed converting the boundaries of entities into a token pair linking problem and designed it on a grid. The following year, Wang et al. (2021) further proposed

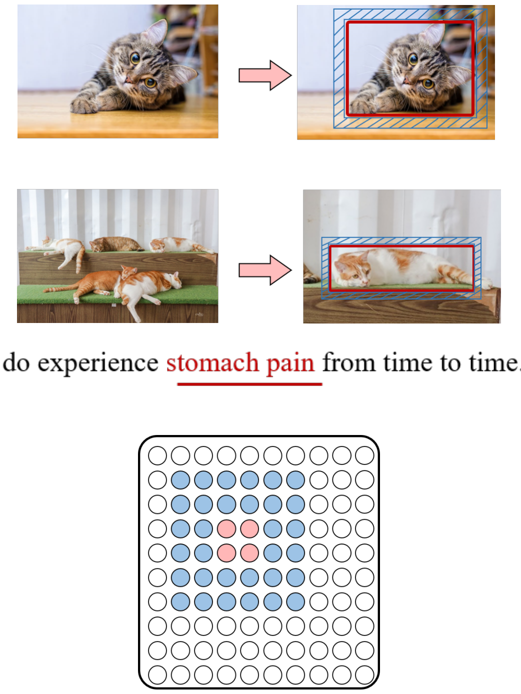

Figure 2: The object boundaries in the image correspond to the entity boundaries in the grid

using two grids to predict entity boundaries and entity word relationships, then decoding the complete entities from the entity fragment graph through maximal clique discovery (Dutta and Lauri, 2019). Subsequently, Li et al. (2022) also proposed a grid-labeling based approach. It transforms discontinuous Named Entity Recognition (NER) into a problem of identifying relationships between words, setting up two textual relations: Next-Neighboring-Word (NNW) and Tail-Head-Word (THW), and using a grid to contain all relationships between words. Lastly, Liu et al. (2022) proposed extending the word-to-word relationships introduced by Li et al., by further expanding with two additional labels: Previous-Neighboring-Word (PNW) and Head-Tail-Word (HTW), to enrich the representation of relationships between words.

2.2 Image Data Augmentation

Image data augmentation is a technique used to enhance the performance of machine learning models, particularly in the fields of image processing and computer vision (Wang et al., 2017; Kumar et al., 2023). This technique involves applying a series of transformations to original images to generate new training samples, thereby increasing the diversity and size of the dataset. The primary goal of data augmentation is to improve the model's generalizability to new, unseen data, enhancing its accuracy

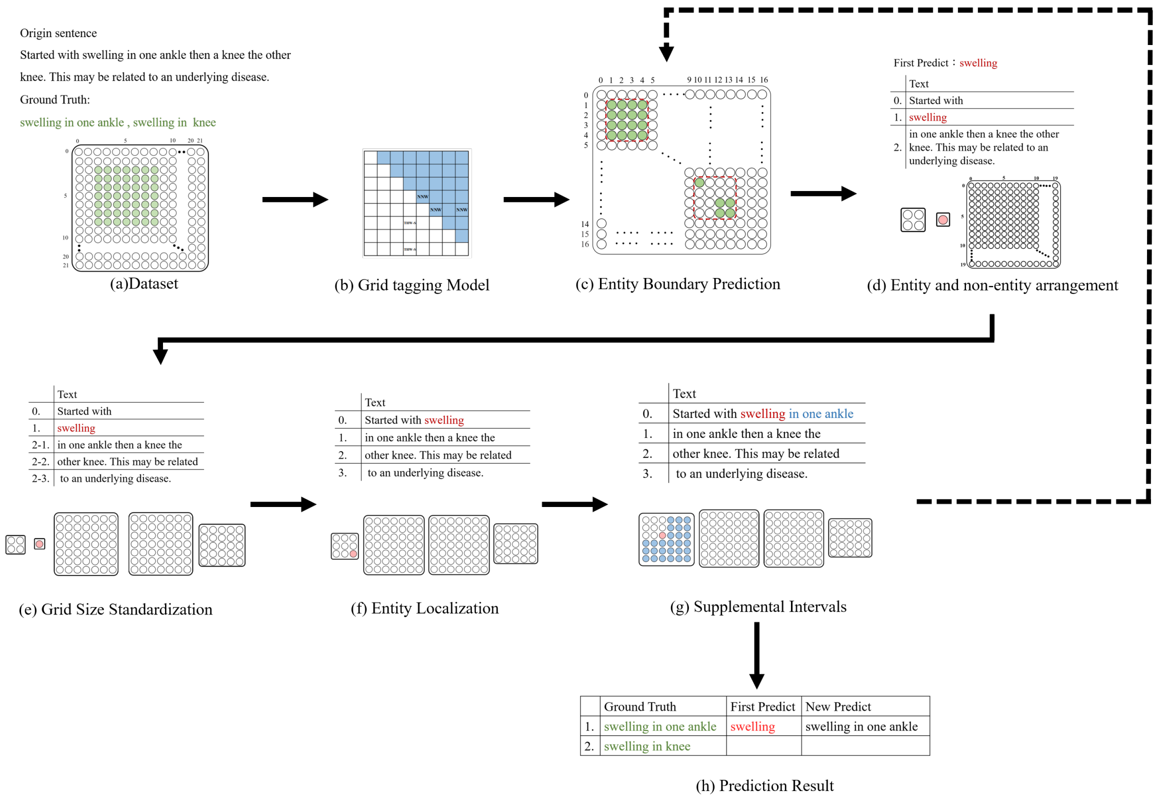

Figure 3: Flowchart of the method

and robustness (Fawzi et al., 2016; Mikołajczyk and Grochowski, 2018). In this study, we have chosen to adjust scaling and cropping as our methods of augmentation, which are applied within a grid framework. We will detail our methods in subsequent chapters.

3 Methodology

This paper proposes a preprocessing method that enhances grid training by combining image data augmentation techniques, leveraging a self-adapted learning approach to progressively train the model to be entity-centric. This method is named "SEDA". Initially, the method optimizes the entity boundaries predicted by the model and re-predicts through an augmentation process. The process comprises four stages: entity boundary prediction, grid size normalization, entity localization, and supplementary interval. The detailed methodologies for each stage will be explained in Sections 3.1 to 3.4, with Figure 3 illustrating the process architecture. Subsequently, we can further optimize the predictions iteratively. This concept is based on using the results after a single augmentation as a baseline to develop this stage, repeating the process outlined in Sections 3.1 to 3.4. In each iteration, we intersect the current predictions with the previous predictions and continue with the subsequent augmentation steps, allowing the entities to progressively approach the correct answers. The subsequent experimental analysis will separately demonstrate the effects of single and multiple augmentations. Next, we will introduce each stage in detail.

3.1 Entity Boundary Prediction

Firstly, as shown in Figures 3(a) to (c), we use a grid tagging model to generate initial target entities and obtain better entity boundary positions through entity boundary prediction. To evaluate these results, we designed a unique set of scoring metrics—Entity Boundary F1 score (EBF), Entity Boundary Precision (EBP), and Entity Boundary Recall (EBR), where EBF is the primary scoring metric used in the initial boundary prediction phase as well as in the subsequent section 3.4. Unlike the traditional F1 score, EBF employs a more lenient strategy, allowing for the prediction of more potential entities. Specifically, we only consider whether the head and tail tokens of the predicted entity match those of the correct entity.

Assuming there are $|\mathcal{D}|$ number of documents, each with correct entities $E = \{e_1, ..., e_n\}$, where n is the number of answer entities, and predicted entities $E' = \{e'_1, ..., e'_m\}$, where m is the number of predicted answer entities. Extract the last word of each entity from texts E and E', resulting in $G = \{g_1, ..., g_n\}$, the last words of the correct entities, and their corresponding last words of predicted entities $P = \{p_1, ..., p_m\}$, with a specific formula design as follows:

$$EBP = \frac{1}{|\mathcal{D}|m} \sum_{d}^{|\mathcal{D}|} \sum_{i}^{n} \sum_{j}^{m} S(p_i, g_j) \quad (1)$$

$$EBR = \frac{1}{|\mathcal{D}|n} \sum_{d}^{|\mathcal{D}|} \sum_{i}^{n} \sum_{j}^{m} S(p_i, g_j) \quad (2)$$

$$S(p_i, g_j) = \begin{cases} 1 & p_i = g_j \\ 0 & p_i \neq g_j \end{cases} \quad (3)$$

$$EBF = \frac{2 \times EBP \times EBR}{EBP + EBR} \quad (4)$$

Document length	Grid Size
∼200	7
200∼350	9
350∼500	11
500∼1000	13
1000∼1350	15
1350∼1500	16
1500∼2000	17
2000∼	19

Table 1: Setting grid sizes corresponding to document sizes

3.2 Grid Size Standardization

The purpose of grid size standardization is to keep the grid size within a specific range to address the issue of reduced model prediction capability when the input is too long. Entity Boundary Prediction results are arranged in an odd-even pattern: sentences at odd positions represent predicted entities[2], and even-positioned sentences contain the remaining text and are longer. As shown in Figure 3(d), we extract these predicted entities and use them as a basis for segmentation. To limit the grid size, we segment sentences at even positions into block sizes according to the size of different texts, as detailed in Table 1. For instance, the size specification for the text in the example is 7, and the sentence "In one ankle then a knee the other knee. This may be related to an underlying disease." exceeds the size specification, so it is split into three segments: 2-1, 2-2, and 2-3, as illustrated in Figure 3(e). This approach effectively maintains the grid size within a certain range.

[2]When entities are predicted to be close to each other (char<=10), we merge them.

3.3 Entity Localization

Next, we design the placement of predicted entities at predetermined locations, primarily positioning the predicted entities later, detailed location choices can be seen in Appendix A. The purpose of this stage is to allow the model to learn the regularity of the data, enhancing its training efficiency. We merge the odd and even-positioned sentences, arranging them in the order of even sentences first, followed by odd sentences, ensuring that each predicted entity is positioned at the end of the sentence, as the change from (e) to (f) in Figure 3 shows, sequentially connecting even sentence "0" to odd sentence "1". For cases like "2-1", "2-2", "2-3" where the sentences are even-numbered and do not contain predicted entities, and no predicted entities are subsequently found, no action is taken in this stage. Simultaneously, we also define two types of sentences: ES (sentences containing predicted entities) and NES (sentences not containing predicted entities), for subsequent use.

3.4 Supplemental Intervals

Subsequently, to prevent the unintended truncation of entities during text segmentation and to ensure the integrity of the predicted entities, we draw inspiration from the concept of magnification in graphic data enhancement to adjust grid size, as shown in (g) of Figure 3. The blue blocks in the example represent the supplemental intervals. Specifically, we designed strategies for pre-supplemental and post-supplemental intervals. For sentences containing predicted entities (ES), one can choose to apply either a pre-supplemental or a post-supplemental; for sentences that do not contain predicted entities (NES), it's possible to apply either or both types of supplements. In the figure, the setting is to apply a post-supplemental of 3 to ES. This method effectively compensates for the parts of predicted entities that might be missed or lacking.

	CADEC				ShARe13				ShARe14			
	All	Train	Dev	Test	All	Train	Dev	Test	All	Train	Dev	Test
#Entities	6,318	4,430	898	990	11,161	5,146	675	5,340	19,157	10,354	810	7,993
#Discontinuous	675	489	93	93	1,090	560	93	437	1,710	992	93	635
%Discontinuous	10.7	11	10.4	9.4	9.8	10.9	13.8	8.2	8.9	9.6	11.5	7.8

Table 2: Complete statistics of three datasets.

Finally, in the testing phase, the process is identical to the aforementioned stages: starting with a preliminary prediction of the target document, followed by steps such as grid size normalization, entity localization, and supplemental intervals, and then prediction. Next, through experiments, we will demonstrate the effectiveness using the current state-of-the-art (SOTA) grid model.

4 Experiments and Results Analysis

4.1 Dataset Introduction

To validate the effectiveness of our method, we selected three Named Entity Recognition (NER) datasets from the biomedical or clinical domain, each featuring discontinuous, nested, and flat entities, namely CADEC, Share/CLEF 2013 (abbreviated as ShARe13), and Share/CLEF 2014 (abbreviated as ShARe14). Additionally, this study continues previous research on Adverse Drug Events (ADEs) entities and Disorder entities. We have tabulated the number of entities in each dataset and the proportion of discontinuous entities as shown in Table 2. Furthermore, we have corrected the test data counts for past datasets, adjusting the number for **ShARe13 from 5333 to 5340, for ShARe14 from 7922 to 7993**, while the data for CADEC remains unchanged. For a detailed comparison of entity counts, please refer to the table 10.

4.2 Backbone Models

We employ the two most state-of-the-art (SOTA) grid models currently used in discontinuous NER: 1) **W2NER** (Li et al., 2022): This model represents the adjacency relationships between entity words as a two-dimensional grid and refines the grid representation through multi-granularity two-dimensional convolution operations to capture complex relationships between entities.

2) **TOE** (Liu et al., 2022): Based on the W2NER model, this model constructs additional textual relations and designs a Tag Representation Embedding Module (TREM) to enhance the model's understanding and representation of entity relationships.

We apply our method to these two models and demonstrate its effectiveness through experimental results.

4.3 Results and Analyses

The results are shown in Table 3, where the first two columns, W2NER and TOE, represent the original scores from the papers, and † denotes our results after multiple validations. The discrepancy between the original paper scores and our execution results was discussed with other researchers and is preliminarily attributed to randomness. Additionally, we discovered a lack of completeness in the original dataset statistics. Therefore, we updated the dataset to reflect the correct number of entities and fixed the seed for subsequent experiments, executing both the original model and our method. Parameter settings are detailed in Table 9. The final scores were calculated based on all test entities in each dataset. "SEDA-Once" and "SEDA-Mul" represent the results of single and multiple enhancements, respectively. It can be observed that applying our method to the grid models significantly improved scores across all datasets.

Under single enhancement, the effects on W2NER and TOE are notable. W2NER's F1 scores increased by 1.79% on CADEC, 1.06% on ShARe13, and 0.56% on ShARe14. TOE's scores improved by 0.43%, 0.88%, and 0.23%, respectively. After multiple enhancements, W2NER's scores further increased to 2.48% on CADEC, 1.27% on ShARe13, and 1.12% on ShARe14, while TOE's scores improved by 0.95%, 1.4%, and 0.93%. This demonstrates that our enhancement technique can improve prediction accuracy by approximately 1%.

Comparing W2NER and TOE, we found that TOE's performance is slightly lower than W2NER. This may be because TOE primarily enhances the relationships between words to improve prediction accuracy, whereas our technique adjusts the grid through trimming and supplementation.

Finally, Table 4 shows the upper limits of our method's scores, all based on single enhancement. If the EBF is successfully increased to 1 during

METHOD	CADEC			ShARe13			ShARe14		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
W2NER (Li et al., 2022)	74.09	72.35	73.21	85.57	79.68	82.52	79.88	83.71	81.75
TOE (Liu et al., 2022)	77.77	70.66	**74.04**	85.18	80.12	82.57	82.26	82.57	82.41
W2NER †	71.01	74.24	72.59	81.91	81.16	81.53	79.38	81.63	80.49
W2NER + SEDA-Once	71.34	**75.25**	*73.29	82.73	82.45	*82.59	78.46	**83.81**	*81.05
W2NER + SEDA-Mul	74.44	73.54	**73.98	82.92	**82.67**	****82.80**	80.30	82.90	****81.61**
TOE †	**76.07**	68.66	72.17	82.04	80.52	81.27	78.38	82.52	80.39
TOE + SEDA-Once	71.50	73.74	72.60	82.57	81.74	*82.15	**80.34**	80.91	80.62
TOE + SEDA-Mul	74.14	72.12	*73.12	**86.62**	79.05	**82.67	79.46	83.26	*81.32

Table 3: The table shows the results of different datasets under two grid architectures, where "†" indicates the scores we replicated, and bold numbers represent the highest scores in each column. "*" denotes significance at $p - value < 0.05$ and "**" denotes < 0.01. Our scores presented are based on calculations from all test entities. SEDA-Once refers to single data augmentation, while SEDA-Mul refers to multiple data augmentations.

Model	Dataset	P	R	F1
W2NER	CADEC	80.41	79.60	80.00
	ShARe13	86.95	82.62	84.73
	ShARe14	84.11	86.47	85.27
TOE	CADEC	78.95	79.19	79.07
	ShARe13	84.61	83.11	83.85
	ShARe14	84.31	85.30	84.80

Table 4: Orcale score

	CADEC	ShARe13	ShARe14
ES	1	1	1
NES	1	0	0
look forward	4	3	2
look backward	4	3	2

Table 5: Parameter settings

enhancement, W2NER's F1 scores on CADEC, ShARe13, and ShARe14 could reach 80.00%, 84.73%, and 85.27%, respectively. TOE's scores could reach 79.07%, 83.95%, and 84.80%. This indicates that if we can perfectly predict the head and tail of discontinuous entities, the model's performance will further improve under our framework.

4.4 Experiment Settings

In terms of experimental settings, the configuration of parameters is detailed in Table 5. In the table, the ES/NES settings with 1 and 0 indicate whether supplemental intervals were used, where 1 represents the use of supplemental intervals and 0 represents not using them. "Look forward" and "look backward" correspond to the ratios of forward and backward supplementation, respectively. For other model parameters and learning rates, please refer to Table 9.

4.5 SEDA-Mul

Figure 4 displays the results of multiple enhancements, where the number of enhancements is indicated. When the number of enhancements is 0, it shows the score results of the original model; other numbers represent the results after the corresponding number of enhancements. The table reveals that the performance of the model does not continuously improve with an increase in the number of enhancements. For instance, on the three datasets, the W2NER model's scores decreased after the second enhancement, while the TOE model first showed a slow increase in scores, followed by a downward trend. Consequently, in practice, we have designed a mechanism in the SEDA-Mul process: if the EBF score on the validation set declines, further enhancements are stopped to prevent worsening the predictive performance.

4.6 Ablation Experiments

We conducted ablation experiments to verify the effectiveness of the enhancement technique in different modules. The baseline for comparison was the score of each setting with a single enhancement. According to the results in Section 4.3, the W2NER model combined with the enhancement technique outperformed the TOE model on multiple datasets. Therefore, we chose the W2NER model to further explore its impact.

The results are shown in Table 6. We tested three settings: "only look forward" (forward-only supplementation), "only look backward" (backward-only supplementation), and "look both sides" (forward and backward supplementation), conducting multiple experiments within the range of 2 to 4. The results indicate that as the supplementation ra-

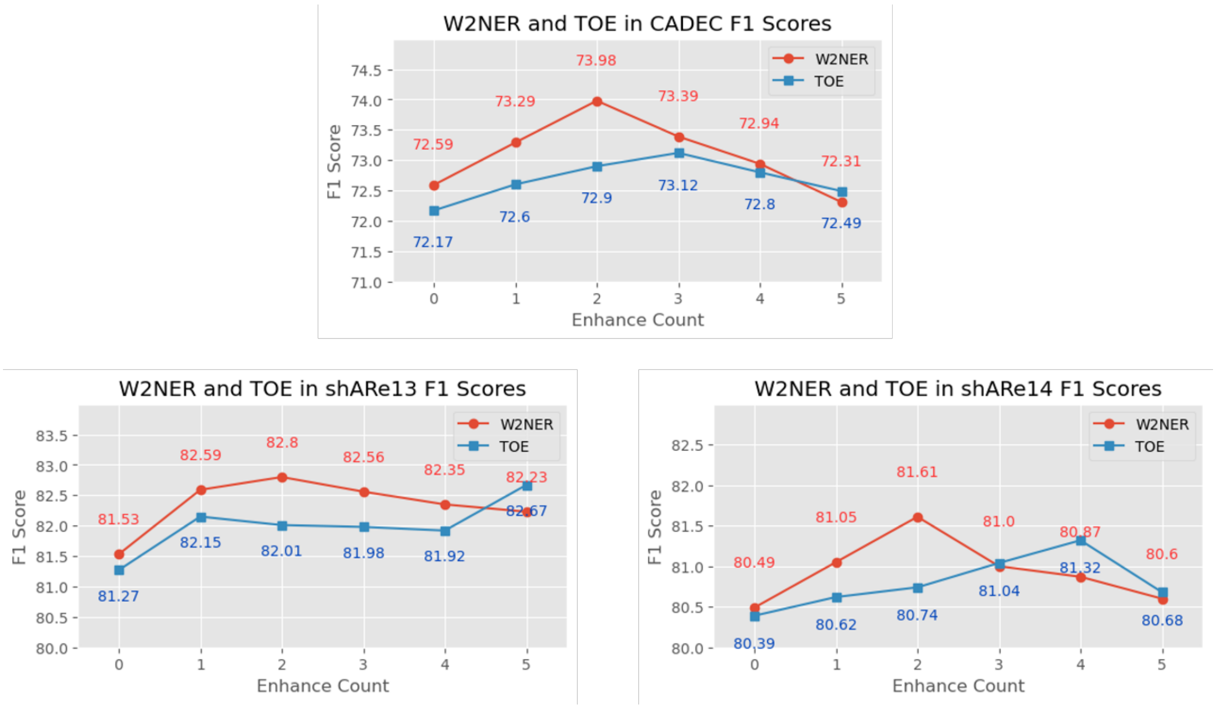

Figure 4: SEDA-Mul score line chart

	CADEC			ShARe13			ShARe14		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Paper Setting	71.34	**75.25**	**73.29**	82.73	**82.45**	**82.59**	78.46	83.81	**81.05**
only look forward =2	**73.71**	70.81	72.23	82.84	81.04	81.93	77.48	83.87	80.55
only look forward =3	71.07	73.94	72.48	82.51	81.85	82.18	79.07	82.27	80.64
only look forward =4	72.11	73.13	72.62	82.86	81.87	82.36	77.93	83.90	80.80
only look backward =2	71.37	74.04	72.68	**83.57**	80.95	82.24	77.03	84.98	80.81
only look backward =3	70.81	74.75	72.73	82.40	82.24	82.32	77.03	**85.07**	80.85
only look backward =4	71.95	74.34	73.12	82.80	82.22	82.51	77.78	84.32	80.92
look both side =2	71.00	73.94	72.44	82.76	82.11	82.44	78.46	83.81	**81.05**
look both side =3	72.04	74.44	73.22	82.73	82.45	**82.59**	79.46	82.61	81.00
look both side =4	71.34	**75.25**	**73.29**	82.45	82.21	82.33	**79.99**	81.92	80.94
w/o EBF	70.45	73.43	71.91	82.74	81.42	82.07	77.52	84.11	80.64

Table 6: Comparison table of scores for ablative experiments on the W2NER grid model

tio increases, the F1 score improves, confirming the effectiveness of the supplemental information. The "only look backward" setting outperformed the "only look forward" setting at supplementation ratios of 2 to 4.

When combining the "look both sides" strategy, performance on all three datasets was better than with single-direction supplementation. Regarding parameter selection, the effect did not always increase with the ratio. The optimal parameter settings for each dataset are shown in Table 9, with the scores based on the best results from multiple experiments.

The ablation analysis of the prediction selection strategy demonstrated its importance in improving the accuracy of entity prediction at the end positions. Without EBF, model performance significantly dropped, nearing the baseline level without enhancement. This is mainly because EBF effectively selects the correct entity positions, avoiding numerous incorrect predictions, thus playing a crucial role in the successful implementation of the enhancement method.

4.7 Case Study

Table 7 shows the comparison before and after SEDA, with the experimental setup using a block size of 7 and the supplementation interval being only look backward 2. During the enhancement process, we were able to remove some nonessential information, making the sentence transformation grid more streamlined. As shown in Examples 1 and 2, this resolves the issue of incomplete entity counts caused by discontinuous sentence segmentation. Example 3 originally contained two sentences, one of which included the discontinuous entity "diffuse...pain". Due to the segmentation of the original text, some entities could not be fully predicted. After SEDA, we were effectively able to predict new entities and gain additional information through the supplemental intervals (marked in blue font). These three examples showed significant improvement after re-prediction, further proving the efficacy of our method.

5 Conclusion

In this paper, we propose a novel approach that applies the concept of image enhancement techniques to neural network models for Named Entity Recognition (NER), which has not been previously introduced in earlier works. This method results in more comprehensive prediction outcomes. After conducting evaluations and analyses across three different datasets, it was found that the integration of enhancement techniques significantly improved the prediction results. Ablation experiments further validated the effectiveness of these enhancements. Further experimental analysis indicates that our proposed model is better at identifying discontinuous entities.

6 Limitations

In the Limitations section, we explore the limitations of our research method from three perspectives. First, our method applies image enhancement techniques to discontinuous Named Entity Recognition (NER), and the enhancement process might disrupt the semantic structure, thus results could be limited when using architectures other than convolutional networks like CNNs. Second, our method relies on initial prediction results to proceed with subsequent SEDA processes, which might lead to higher time complexity in practical operations or computations, but this is acceptable if applied only in preprocessing. Lastly, we observe that the effectiveness of enhancements gradually diminishes after multiple iterations, possibly due to the multi-round intersect strategy still generating erroneous predicted entities. Therefore, in practical applications, one or two enhancements are usually sufficient. Future research will focus on how to

Origin: **1.** Lower legs starting to hurt after three weeks of very painful feet. Can hardly walk without higher pain level. Ground Truth: Lower legs. . . hurt/ very painful feet / hardly walk / pain First Predict: hurt / very painful feet /walk / pain SEDA: **1.**Lower legs starting to hurt after three **2.**after three weeks of very painful feet . Can **3.**. Can hardly walk without higher **4.**without higher pain level. **5.**level. SEDA Predict: very Painful feet/ Legs. . . hurt / hardly walk / pain
Origin: **1.** Constant gas and constipation. It has reduced my cholesterol level. Ground Truth: gas/ constipation First Predict: Constant gas / constipation SEDA: **1.**Constant gas and constipation. It **2.** . It has reduced my cholesterol level. **3.** . SEDA Predict: gas/ constipation
Origin: **1.** Investigation of the transfusion reaction revealed diffuse trunk **2.**and arm pain during transfusion of compatible red blood cells. Ground Truth: transfusion reaction / diffuse...pain First Predict: transfusion reaction / pain SEDA: **1.** Investigation of the transfusion reaction revealed diffuse **2.** revealed diffuse trunk and arm pain during transfusion **3.** during transfusion of compatible red blood cells . **4.** . SEDA Predict: transfusion reaction / diffuse...pain

Table 7: Case Study

effectively remove these erroneous entities.

References

Shorten Connor and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020a. An effective transition-based model for discontinuous NER. In *Proceedings of the ACL*, pages 5860–5870.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020b. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

Sourav Dutta and Juho Lauri. 2019. Finding a maximum clique in dense graphs via 2 statistics. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2421–2424, New York, NY, USA. Association for Computing Machinery.

Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. 2016. Adaptive data augmentation for image classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3688–3692. Ieee.

Hao Fei, Donghong Ji, Bobo Li, Yijiang Liu, Yafeng Ren, and Fei Li. 2021. Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12785–12793.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the NAACL*, pages 861–871.

Teerath Kumar, Alessandra Mileo, Rob Brennan, and Malika Bendechache. 2023. Image data augmentation approaches: A comprehensive survey and future directions. *arXiv preprint arXiv:2301.02830*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the NAACL*, pages 260–270.

Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. A span-based model for joint overlapped and discontinuous named entity recognition. In *Proceedings of the ACL-IJCNLP*, pages 4814–4828.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Jiang Liu, Donghong Ji, Jingye Li, Dongdong Xie, Chong Teng, Liang Zhao, and Fei Li. 2022. Toe: A grid-tagging discontinuous ner model enhanced by embedding tag/word relations and more fine-grained tags. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:177–187.

Ruibo Liu, Jason Wei, Chenyan Jia, and Soroush Vosoughi. 2021. Modulating language models with emotions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4332–4339, Online. Association for Computational Linguistics.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the EMNLP*, pages 857–867.

Agnieszka Mikołajczyk and Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE.

Agnieszka Mikołajczyk and Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. In *Proceedings of the ACL-IJCNLP*, pages 2782–2794.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the EMNLP*, pages 2670–2680.

Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen. 2018. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using lstm-crf. *Wireless Communications & Mobile Computing (Online)*, 2018.

Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *Proceedings of the EMNLP*, pages 204–214.

Jason Wang, Luis Perez, et al. 2017. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8.

Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020a. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the ACL*, pages 5918–5928.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020b. TPLinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1572–1582, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021. Discontinuous named entity recognition as maximal clique discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 764–774, Online. Association for Computational Linguistics.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. *arXiv preprint arXiv:2106.01223*.

Fisher Yu and Vladlen Koltun. 2016. Multi-scale context aggregation by dilated convolutions.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the ACL*, pages 6470–6476.

A Impact of Entity Boundaries

To enhance the efficiency of the model in learning key features, at the beginning of our research, we explored the impact of entity boundaries on model predictions. We designed a series of experiments using the CADEC dataset, where parts of the given text unrelated to entities were masked. The experiments included masking the content before the first entity in the text, after the last entity, and both ends simultaneously, to observe the impact of boundaries on model predictions. The results, as shown in Table 8, indicate that each masking effect significantly influences entity prediction. It is observed that the model performs best when both ends are masked, with an F1 score of 0.7919; followed by masking only after the last entity with a score of 0.7615, and finally masking only before the first entity with a score of 0.7456. These findings motivate us to further explore: if we choose a relatively ample approach in predictions by positioning the prediction of entity locations posteriorly, simulating the situation through boundary confirmation by answers, whether it can enhance model performance.

	P	R	F1
Origin	75.65	67.78	71.50
Masking before the first entity	76.71	72.53	74.56
Masking after the last entity	78.92	73.74	76.15
Masking on both sides	78.99	79.39	79.19

Table 8: The influence of CADEC boundaries on entities.

B Model parameter settings

	CADEC	ShARe13	ShARe14
d_h	768	768	768
d_{Ed}	20	20	20
d_{Et}	20	20	20
d_c	80	80	80
Dropout	0.5	0.5	0.5
Learning rate(BERT)	$5e-6$	$5e-6$	$5e-6$
Learning rate(other)	$1e-3$	$1e-3$	$1e-3$
Batch size	16	20	20
warm factor	0	0	0.1
weight decay	0	0.4	0.4
epoch	10	20	10
W2NER seed	123	123	123
TOE seed	1898	1898	1898

Table 9: Model parameter settings

System	Statistics	CADEC			ShARe13			ShARe14		
		Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Our	#Entities	4,430	898	990	5,146	675	5,340	10,354	810	7,993
	#Discontinuous	489	93	93	560	93	437	992	93	625
	%Discontinuous	11	10.4	9.4	10.9	13.8	8.2	9.6	11.5	7.8
W2NER/TOE	#Entities	4,430	898	990	5,146	669	5,333	10,354	771	7,922
	#Discontinuous	491	94	94	581	71	436	1,004	80	566
	%Discontinuous	11.1	10.5	9.5	11.3	10.6	8.2	9.7	10.4	7.1

	CADEC	**ShARe 13**	**ShARe 14**
Text type	online posts	clinical notes	clinical notes
Entity type	ADE	Disorder	Disorder
# Documents	1,250	298	433
# Tokens	121K	264K	494K
# Sentences	7,597	18,767	34,618
# Mentions	6,318	11,161	19,131
# Disc.M	675 (10.6)	1,090 (9.7)	1,710 (8.9)
Avg mention L.	2.7	1.8	1.7
Avg Disc.M L.	3.5	2.6	2.5
Avg interval L.	3.3	3.0	3.2
Discontinuous Mentions			
2 components	650 (95.7)	1,026 (94.3)	1,574 (95.3)
3 components	27 (3.9)	62 (5.6)	76 (4.6)
4 components	2 (0.2)	0 (0.0)	0 (0.0)
No overlap	82 (12.0)	582 (53.4)	820 (49.6)
Overlap at left	351 (51.6)	376 (34.5)	616 (37.3)
Overlap at right	152 (22.3)	102 (9.3)	170 (10.3)
Multiple overlaps	94 (13.8)	28 (2.5)	44 (2.6)
Continuous Mentions			
Overlap	326 (5.7)	157 (1.5)	228 (1.3)

Table 10: Statistics for three datasets are presented in each paper. The descriptive statistics of the data sets. ADE: adverse drug events; Disc.M: discontinuous mentions; Disc.M L.: discontinuous mention length, where intervals are not counted. Numbers in parentheses are the percentage of each category. This table is referenced by Dai et al. (2020b). Note: We have discovered discrepancies in the total counts of discontinuous entities in the CADEC database as reported by the W2NER and TOE systems compared to the totals described in the Dai et al. (2020b) paper. This may be due to statistical errors or an increase in data volume. To avoid controversy, the experiments in this paper are conducted using the entity counts we have recalculated.