NerCo: A Contrastive Learning Based Two-Stage Chinese NER Method

Zai Zhang[1,2], **Bin Shi**[1,2*], **Haokun Zhang**[1,2], **Huang Xu**[3,4], **Yaodong Zhang**[2], **Yuefei Wu**[1,2], **Bo Dong**[2,5] and **Qinghua Zheng**[1,2]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, China
[2]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, China
[3]Servyou Software Group Co., Ltd
[4]Hangzhou City University, China
[5]School of Distance Education, Xi'an Jiaotong University, China
{zhg.zai, yuelfei.wu}@gmail.com, {shibin, dong.bo, qhzheng}@xjtu.edu.cn, {zhanghaokun, zhangyaodong}@stu.xjtu.edu.cn, xhg@servyou.com.cn

Abstract

Sequence labeling serves as the most commonly used scheme for Chinese named entity recognition(NER). However, traditional sequence labeling methods classify tokens within an entity into different classes according to their positions. As a result, different tokens in the same entity may be learned with representations that are isolated and unrelated in target representation space, which could finally negatively affect the subsequent performance of token classification. In this paper, we point out and define this problem as *Entity Representation Segmentation in Label-semantics*. And then we present **NerCo**: **N**amed **e**ntity **r**ecognition with **Co**ntrastive learning, a novel NER framework which can better exploit labeled data and avoid the above problem. Following the pretrain-finetune paradigm, NerCo firstly guides the encoder to learn powerful label-semantics based representations by gathering the encoded token representations of the same Semantic Class while pushing apart that of different. Subsequently, NerCo finetunes the learned encoder for final entity prediction. Extensive experiments on several datasets demonstrate that our framework can consistently improve the baseline and achieve state-of-the-art performance.

1 Introduction

Named Entity Recognition benefits a wide range of downstream tasks in natural language processing(NLP). Due to their simplicity and effectiveness, sequence labeling methods have long been the most common solution tackling this task[Huang *et al.*, 2015; Ma and Hovy, 2016; Lample *et al.*, 2016]. Recent success of Transformer-based [Vaswani *et al.*, 2017] large-scale pre-trained language models(PLMs) such as BERT [Devlin *et al.*, 2018] and RoBERTa [Liu *et al.*, 2019] have greatly boosted these sequence labeling methods to get better contextualized token representations. As a result, although Chinese NER is more difficult due to the language's natural lack of word boundary information, impressive performance improvements have been achieved with the combination of additional techniques such as lexicon integration and so on[Li *et al.*, 2020; Liu *et al.*, 2021a].

Current sequence labeling methods rely on token labels in the form of *'Position-Type'* as their supervision signals [Sang and De Meulder, 2003] to perform token classification. Each entity label is composed of a chunk tag and a type tag, respectively indicating its position within the entity and its category. Taking an `PER` (Person) entity in *BIO* tagging format for an example, its first token is annotated `B-PER` and the remaining tokens are annotated `I-PER` , where the chunk tag `B-` facilitates recognizing consecutive entities of the same type [Ramshaw and Marcus, 1995]. Though it is simple and effective, we find this straightforward scheme may lead to a serious problem that severely hinders the generalization performance of the model. Specifically, one entity usually contains different positional tokens. Those tokens naturally share the same label semantics since they belongs to the same entity type. However, the supervising scheme adopted by sequence labeling assigns independent labels and enforces them to be treated as distinct and isolated classes (e.g. B-Type, I-Type). As a result, it makes their representations dispersed into several separate and unrelated clusters, even though they are in the same entity. We call this problem *Entity Representation Segmentation in Label-semantics*.

As shown in Figure 1, taking a labeled entity '张家川县' (Zhangjiachuan County) which is a Geo-Political Entity(`GPE`) labelled in *BIO* tagging format as an example, traditional methods will still guide models to label the first token "张(Zhang)" `B-GPE` and subsequent tokens `I-GPE`, leading to the aforementioned problem of *Entity Representation Segmentation in Label-semantics*. This training scheme results in a contradiction that tokens of an entity sharing the same label semantics would be embedded far and unrelated in Euclidean space and the integral entity representation would be semantically segmented, only due to the slightly different inner-entity position indicated by their chunk tags. Such representation deficiency is not trivial, because it could severely damage subsequent classification by confusing models in determining entity boundaries and naturally influence their type decision.

*Corresponding Author

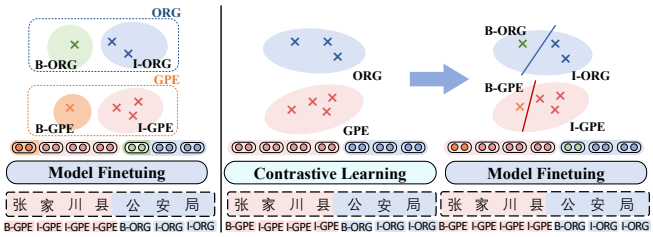


Figure 1: A comparison between traditional methods and our proposed method. Unlike traditional methods on the left, our approach takes a two-stage learning strategy. In the first stage, we conduct contrastive learning for label-semantics based representations. Then we finetune the learned model in the second stage, equipping it with inner-entity position discrimination for chunk tags and linear mapping to type tags for each token.

In fact, to facilitate model's classification head to correctly recognize an entity's boundary, the representations in this sequence should be semantically **coherent and consistent inside** and **distinctive and discriminative outside**. So, in this paper, we propose a novel framework named **NerCo**: **N**amed **e**ntity **r**ecognition with **Co**ntrastive learning to address *Entity Representation Segmentation in Label-semantics* problem. To satisfy the demand of intra-entity coherence and inter-entity distinctiveness, NerCo naturally leverages contrastive loss to pull semantically close neighbors of entity tokens together and push non-neighbors apart to learn a label-semantics based representation learning model in the first stage. In the second stage, we finetune the model with traditional token supervision in *BIO* format to capture inner-entity positional information for final prediction. Although NerCo is simple and easy to use, requiring neither data augmentation nor base encoder modification, it can fully exploit model's representation potential after injecting label-semantic signals entailed in raw data, thus effectively alleviating the *Entity Representation Segmentation in Label-semantics* problem and achieving better generalization performances.

Experiments on four datasets show that NerCo achieves consistent improvements over only finetuning the base encoder[1]. Our method improves F1-score by 4.11% on Weibo dataset specifically and achieves new state-of-the-art performances on all of the Chinese NER datasets in our experiments. The results demonstrate that sequence labeling methods can exhibit much stronger power in entity representation and recognition after properly solving the *Entity Representation Segmentation in Label-semantics* problem.

2 Related Work

Our work is related to existing sequence labeling methods for Chinese NER and contrastive learning in NLP.

2.1 Chinese NER as Sequence-Labeling

NER has long been formulated as a sequence labeling task and the current state-of-the-art results for sequence labeling have been achieved by neural network approaches[Huang *et al.*, 2015; Ma and Hovy, 2016; Chiu and Nichols, 2016]. Compared with NER in English, Chinese NER is more difficult due to the absence of explicit word delimiter in Chinese sentences. Previous studies have empirically proved the superiority of character-based methods over word-based ones[He and Wang, 2008; Liu *et al.*, 2010; Li *et al.*, 2014].

Since lexicon features can provide rich word boundary information, integrating them into a character-based sequence encoder has attracted research attention. [Zhang and Yang, 2018] first proposed the lattice structure to encode all potential words matched in a sentence. [Gui *et al.*, 2019a] introduced a rethinking mechanism to tackle the conflicts between potential words. To better capture distance and direction information, TENER [Yan *et al.*, 2019] customized the transformer encoder to incorporate both character and word features using relative position encoding. While LGN[Gui *et al.*, 2019b] and CGN[Sui *et al.*, 2019] utilized graph neural network to model the interaction within a character and word sequence, [Li *et al.*, 2020] converted the lattice structure into a flat sequence consisting of spans and achieved excellent and stable performances.

Considering the significant NER improvement brought by pre-trained models [Devlin *et al.*, 2018; Liu *et al.*, 2019], many researchers turned to incorporating lexicon knowledge into pre-trained models to combine both advantages. ERNIE [Sun *et al.*, 2019] leveraged entity-level and word-level masking strategies to implicitly integrate external knowledge into BERT. [Diao *et al.*, 2019] proposed a BERT-based Chinese text encoder to explicitly consider potential word boundaries while pre-training and fine-tuning. Different from previous pre-training methods, [Liu *et al.*, 2021b] proposed to integrate lexicon information using a lexicon adapter between transformer blocks, which encouraged more sufficient interactions between lexicon features and BERT.

2.2 Contrastive Learning in NLP

Contrastive Learning tries to learn powerful representations in such a way that similar features are pulled together and dissimilar ones are pushed apart in representation space[Hadsell *et al.*, 2006; Jaiswal *et al.*, 2020]. Recently, it has become a rising domain and achieved great success in computer vision community[Chen *et al.*, 2020; He *et al.*, 2020]. To employ contrastive learning in NLP, a key question to answer is how to construct positive pairs. [Fang *et al.*, 2020] exploited back-translation to create positive instances of original sentences. Similar to visual domain, [Wu *et al.*, 2020] conducted multiple sentence-level data augmentations, such as word substitution, synonym substitution and reordering. While [Giorgi *et al.*, 2021] regarded spans within a document as similar instances, [Gao *et al.*, 2021] applied *dropout* to the same sentence embedding twice to get positive pairs. Another line of positive pair construction in NLP research utilized a similar contrastive learning objective, the only difference is that they used labelled datasets for constructing positive instances[Henderson *et al.*, 2017; Gillick *et al.*, 2019; Gao *et al.*, 2021].

Existing approaches have attempted to apply contrastive learning to NER. SCL-RAI[Si *et al.*, 2022] is proposed to cope with the Unlabeled Entity Problem by decreasing the distance

[1]We release our codes at https://github.com/zhzai/nerco.

Semantic Class	Original Class (BIO)	Original Class (BMES)
ORG	B-ORG, I-ORG	B-ORG, M-ORG,E-ORG,S-ORG
PER	B-PER, I-PER	B-PER, M-PER, E-PER, S-PER
GPE	B-GPE, I-GPE	B-GPE, M-GPE, E-GPE, S-GPE
LOC	B-LOC, I-LOC	B-LOC, M-LOC,E-LOC,S-LOC
Non-Entity	O	O

Table 1: The *Semantic Classes* in Ontonotes 4.0 dataset. We merge original classes in 'BIO' and 'BMES' format into *Semantic Classes* for contrastive pair construction.

of span representations with the same label and increasing it for different ones via span-based contrastive learning. [Das *et al.*, 2022] presents a novel contrastive learning technique called CONTaiNER for few-shot named entity recognition. [Zhang *et al.*, 2023] proposes a bi-encoder framework which applies contrastive learning to map text and entity types into the same vector space. Unlike our work, this paper requires large modifications to the model, which brings a relatively high complexity.

3 Method

Inspired by the idea of self-supervised representation learning and pretrain-finetune paradigm, we propose a two-stage framework for Chinese named entity recognition. In the first stage, we leverage labeled data to construct contrastive pairs and train the encoder for label-semantics based token representations using contrastive loss, which is shown in Figure 2. In the second stage, we maintain the previous sequence labeling convention, in which original labeled data and loss function are utilized for final entity recognition. This framework also complies with our intuition of solving *Entity Representation Segmentation in Label-semantics*, for which we first learn cohesive representations within an entity corresponding to its label semantics and then finetune the encoder, equipping it with the ability of inner-entity position discrimination for chunk tags and linear mapping to type tags for each token, i.e. assigning *'Position-Type'* tags.

3.1 Contrastive Pair Construction

To further clarify our construction of contrastive pairs and better resolve the *Entity Representation Segmentation in Label-semantics* problem, we first present the definition of ***Semantic Class*** here.

Semantic Class represents the semantic category of the entity token, regardless of its positional role within an entity. A *Semantic Class* is signified only by the type tag of the token's *'Position-Type'* label. In other words, tokens with different labels of position but the same of semantics are merged into a single category named *Semantic Class*. For tokens with label `O`, we set them a separate `Non-Entity` class. Taking `B-PER` and `I-PER` in Ontonotes 4.0 dataset using *BIO* tagging format as an example, tokens labelled by them both belong to the *Semantic Class* of `PER`, indicating their 'person entity' label semantics. The mapping relationship is illustrated in Table 1.

Then, we can answer the usually critical question in contrastive learning: how to construct $(x_i; x_i^+)$ and $(x_i; x_i^-)$ pairs. Following a simple implementation in SimCLR[Chen *et al.*, 2020], we construct positive and negative pairs using the token representations within a mini-batch. We consider a query and a key as a positive pair if they belong to the same *Semantic Class* S_t. All token representations from different *Semantic Classes* are considered negative samples for each token.

3.2 Semantic Class Guided Contrast

Naturally, the token representations belonging to the same *Semantic Class* should be more similar and related to the their entity's type semantics. Thus we use contrastive learning to pull the positive pairs together and push the negative pairs apart in target space. Following the de facto procedure in contrastive learning in computer vision [Chen *et al.*, 2020; He *et al.*, 2020], we propose to take the contrastive training as the first learning stage of our framework. We adopt InfoNCE loss in SimCLR[Chen *et al.*, 2020] as our contrastive loss function. For a query token x_i and one of its positive key, the contrastive loss for (x_i, x_i^+) is defined as $\mathcal{L}_i^{ctr}$:

$$\mathcal{L}_i^{ctr} = -log\frac{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)/\tau} + \sum_{j=1}^{N}(e^{sim(\mathbf{h}_i,\mathbf{h}_j^-)/\tau})} \quad (1)$$

where $\mathbf{h}_i$ and $\mathbf{h}_i^+$ denote the representations of x_i and x_i^+, $\mathbf{h}_j^-$ is the representation of each x_j^- in the mini-batch, τ is the temperature hyper-parameter, and $sim(\mathbf{h_1}, \mathbf{h_2})$ is the cosine similarity $\frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\| \cdot \|\mathbf{h}_2\|}$. The sum in the denominator is over one positive and N negative tokens.

In the first stage of contrastive learning, we optimize the sum of the above loss function corresponding to all positive pairs in the mini-batch. Considering that tokens with `O` labels are not entities and contain miscellaneous label semantics, pushing their representations together in embedding space as positive pairs does not make a clear sense. Here we filter out the cases where tokens of the *Semantic Class* `Non-Entity` serve as queries.

3.3 Token Label Supervised Fine-tuning

In the second stage, we finetune the above learned model using conventional sequence labeling method. To capture dependencies between consecutive tags, a CRF layer is leveraged on top of the encoder.

Given K labelled data $\{\mathbf{s}_i, \mathbf{y}_i\}|_{i=1}^K$, K is the number of sentences in a mini-batch. $H = \{h_1, h_2, ..., h_n\}$ is the representation sequence of sentence $\mathbf{s}_i$ output by the encoder. We first perform a linear transformation for subsequent classification:

$$O = \mathbf{W_o}H + \mathbf{b_o} \quad (2)$$

where $\mathbf{W_o}$ and $\mathbf{b_o}$ are learnable parameters. After that, we can get the probability of the label sequence $\mathbf{y}_i = \{y_1, y_2, ..., y_n\}$ of $\mathbf{s}_i$:

$$p(\mathbf{y}_i|\mathbf{s}_i) = \frac{\exp(\sum_j (O_{j,y_j} + T_{y_{j-1},y_j}))}{\sum_{\tilde{\mathbf{y}}_i} \exp(\sum_j (O_{j,\tilde{y}_j} + T_{\tilde{y}_{j-1},\tilde{y}_j}))} \quad (3)$$

where T is the transition score matrix and $\tilde{\mathbf{y}}_i$ denotes all possible tagging sequences.

We train the model by minimizing the sum of all sentences' negative log-likelihood losses in the mini-batch as:

$$\mathcal{L}_{finetune} = -\sum_i \log(p(\mathbf{y}_i|\mathbf{s}_i)) \quad (4)$$

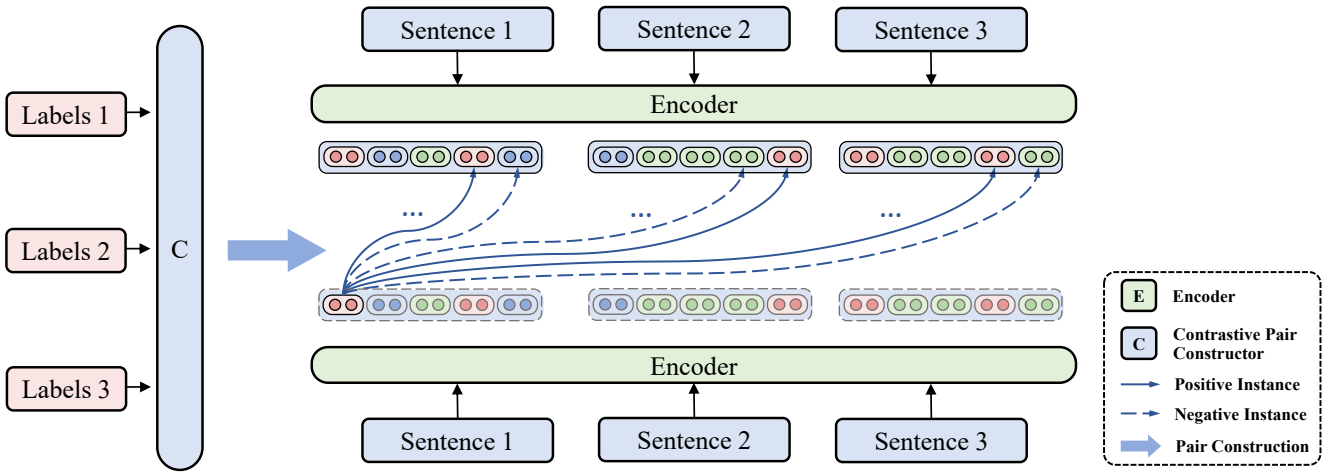


Figure 2: Contrastive representation learning as the first stage of NerCo.

Datasets	Type	Train	Dev	Test
OntoNotes	Sentence	15.7k	4.3k	4.3k
	Char	491.9k	200.5k	208.1k
MSRA	Sentence	46.4k	-	4.4k
	Char	2169.9k	-	172.6k
Weibo	Sentence	1.4k	0.27k	0.27k
	Char	73.8k	14.5k	14.8k
Resume	Sentence	3.8k	0.46k	0.48k
	Char	124.1k	13.9k	15.1k

Table 2: Statistics of datasets.

For inference, we search the label sequence with the highest score using Viterbi algorithm to recognize entities.

4 Experimental Setup

4.1 Datasets

We conducted experiments on four Chinese NER datasets to evaluate our proposed method. (1) **OntoNotes 4.0**[Weischedel *et al.*, 2011] is a multilingual corpus that is manually annotated in the news domain with various text annotations, including Chinese named entity tags. We only utilized Chinese documents, consisting of four entity types, and processed the data in the same manner as [Che *et al.*, 2013]. (2) **MSRA**[Levow, 2006] is also a news domain corpus serving for word segmentation and named entity recognition in Chinese. In our study of Chinese NER, it includes three named entity types: LOC, PER, and ORG. (3) **Weibo NER**[Peng and Dredze, 2016] comprises annotated NER labels from the social media website Sina Weibo. The corpus includes both named entities and nominal mentions for four types of entities: PER, ORG, GPE, and LOC. (4) **Resume NER** annotated by [Zhang and Yang, 2018] consists of resumes of senior executives and is annotated with 8 types of named entities. The statistic information of the datasets is presented in Table 2.

4.2 Baselines

We take FLAT Transformer in [Li *et al.*, 2020] as our baseline model. **FLAT** is a strong baseline that converted the lattice structure into a flat sequence and used relative position encoding for Chinese NER. **BERT**[Devlin *et al.*, 2018] is the mostly used method to finetune a pre-trained Chinese BERT for Chinese NER. **LatticeLSTM**[Zhang and Yang, 2018], **TENER**[Yan *et al.*, 2019], **SoftLexicon**[Ma *et al.*, 2020] and **LEBERT**[Liu *et al.*, 2021a] are other representative models integrating lexicon information into Chinese sequence for better NER performance. **RICON**[Gu *et al.*, 2022] and **MarkBERT**[Li *et al.*, 2022b] are newly proposed competitive models for Chinese named entity recognition and both achieved strong performances. **W^{2}NER**[Li *et al.*, 2022a] treats NER as a word-word relation classification task in a unified formalism, while **BoundarySmoothing**[Zhu and Li, 2022] proposed a boundary smoothing method as a regularization technique for span-based NER model, achieving the sota level on the four Chinese NER datasets.

4.3 Implementation Details

We adopt the FLAT Transformer in [Li *et al.*, 2020] as our base encoder architecture. Following the setting in [Li *et al.*, 2020], We use one layer of encoding layer, and SGD in both stages to optimize the encoder. We leverage BERT for contextual token embedding[Devlin *et al.*, 2018]. During the first stage of contrastive learning, we utilize InfoNCE objective to optimize the parameters in FLAT encoder(with BERT parameters frozen) until achieving the minimal loss. Afterwards, we finetune all the parameters of the learned model, including BERT, in the manner of normal sequence labeling. Considering the relatively small scale of Weibo and Resume, we only tune the classification head as a prior step before updating all parameters to stabilize this second-stage finetuning process. The way to select hyper-parameters is also the same as FLAT.

5 Experimental Results

5.1 Performance

Overall Results

Table 3 shows the experimental results on Chinese NER datasets. As shown in the table, our model outperforms our baseline model and other methods consistently on four Chinese NER datasets[2]. As a strong baseline method, we observe FLAT[Li *et al.*, 2020] already surpass BERT tagger by a large margin and could give us a relatively strong and stable performance in Chinese NER. Thanks to our additional contrastive learning procedure, NerCo can effectively improve the performance of baseline FLAT[Li *et al.*, 2020] by an average F1 score of 1.5% on the datatsets. Especially, our model brings a large relative improvement over the baseline of up to 6% on Weibo, in terms of the F1 score. As for the current state-of-the-art models, taking W^{2}NER[Li *et al.*, 2022a] as an example, it leveraged multiple types of embeddings and proposed complex architectures including multi-granularity dilated convolutions to capture word-word interaction information. Though it shows impressive performance and a large margin over other lexicon-based methods, our model outperforms W^{2}NER on all Chinese NER datasets. In general, our method could surpass all the current top-performing methods, pushing the state-of-the-art performances of Chinese NER[3].

Span F & Type Acc

Compared with FLAT, NerCo performs the additional label-semantics guided contrastive learning as the first stage. We evaluate these two models in terms of Span F and Type Acc to further investigate our performance gains. **Span F** measures the F1 score of recognized entity spans over the gold spans, regardless of the correctness of their types. **Type Acc** is the proportion of full-correct predictions to span-correct predictions.

Table 4 shows these two metrics of FLAT and NerCo. We can find our model performs better than the baseline in both metrics, which demonstrates that the contrastive learning stage benefits to both span boundary detection and span classification. Specifically, the performance gain on Weibo is very obvious in both metrics, indicating that solving the problem of *Entity Representation Segmentation in Label-semantics* can greatly promote boundary detection and also help a lot in type prediction especially for small-scale NER datasets. For Ontonotes and Resume, the improvements on Span F are more significant than that on Type Acc and the situation is reversed on MSRA, showing that NerCo can respectively promote boundary detection and type decision in the two cases.

Performance Against Sentence Length

Figure 3 shows the F1 score trends of the mostly used BERT, our baseline FLAT and our method NerCo against sentence length. Here we show two representative curves of the test

[2]In Table 3, the results of FLAT are implemented by ourselves, while scores of other methods are copied from their original papers. '-' means that the paper didn't report this item.

[3]We refer readers to Appendix for the experimental results on the English setting at https://github.com/zhzai/nerco/blob/master/Appendix.pdf.

Model	Ontonotes		
	Pr.	Rec.	F1
BERT	76.01	79.96	79.96
LatticeLSTM	76.35	71.56	73.88
TENER	76.13	73.68	74.89
SoftLexicon	83.41	82.21	82.81
LEBERT	-	-	82.08
MarkBERT	81.70	83.70	82.70
RICON	81.95	**84.78**	83.33
W^{2}NER	82.31	83.36	83.08
BoundarySmoothing	81.65	84.03	82.83
FLAT	83.64	82.08	82.85
Ours	**84.43**	82.82	**83.62**

Model	Resume		
	Pr.	Rec.	F1
BERT	94.87	96.50	95.68
LatticeLSTM	94.81	94.11	94.46
TENER	95.28	95.46	95.37
SoftLexicon	96.08	96.13	96.11
LEBERT	-	-	96.08
W^{2}NER	**96.96**	96.35	96.65
BoundarySmoothing	96.63	96.69	96.66
FLAT	95.57	96.63	96.10
Ours	96.94	**97.12**	**96.82**

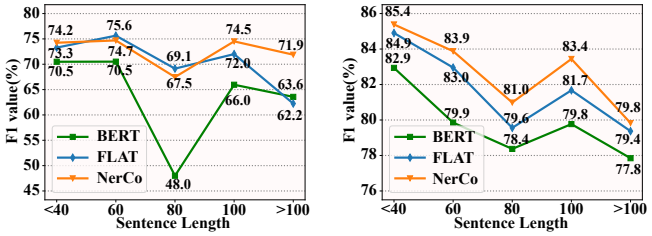
Model	MSRA		
	Pr.	Rec.	F1
BERT	93.40	94.12	93.76
LatticeLSTM	93.57	92.79	93.18
TENER	94.19	92.73	93.46
SoftLexicon	95.75	95.10	95.42
LEBERT	-	-	95.70
MarkBERT	96.10	96.00	96.10
RICON	95.94	**96.33**	96.14
W^{2}NER	96.12	96.08	96.10
BoundarySmoothing	**96.37**	96.15	96.26
FLAT	95.75	95.97	95.86
Ours	96.36	96.23	**96.29**

Model	Weibo		
	NE	NM	Overall
BERT	65.77	62.05	63.80
LatticeLSTM	53.04	62.25	58.79
TENER	55.34	64.98	60.21
SoftLexicon	70.94	67.02	70.50
LEBERT	-	-	70.75
W^{2}NER	-	-	72.32
BoundarySmoothing	-	-	72.66
FLAT	66.96	70.78	68.68
Ours	**73.06**	**72.50**	**72.79**

Table 3: Results for Chinese NER.

	Span F			
	Ontonotes	MSRA	Weibo	Resume
FLAT	84.33	96.32	74.00	96.34
NerCo	**84.94**	**96.57**	**76.61**	**96.94**
	Type Acc			
	Ontonotes	MSRA	Weibo	Resume
FLAT	98.24	97.91	92.81	99.75
NerCo	**98.44**	**99.72**	**95.02**	**99.87**

Table 4: Span F and Type Acc of different models.



(a) Weibo (b) Ontonotes

Figure 3: F1 score against the sentence length.

splits of Weibo and Ontonotes. In general, our method shows better performance and robustness on both datasets. Although FLAT performs similarly to our method when the sentence length is less than 80, its performance deteriorates as the sentence length increases and lags significantly behind our method. Notably, FLAT was even surpassed by BERT when the sentence length exceeds 100. However, our method can always perform well and, most importantly, remain stable when the sentence length varies. For Ontonotes, our method can always maintain a margin over FLAT in all sentence length sections, demonstrating more adequate and rational utilization of labeled data.

5.2 Ablation Study

To investigate the contribution of several key components and our proposed two-stage method, we conduct several ablation experiments on the four datasets. First, without contrastive learning, we only finetune the encoder and observe a large performance drop in our implementation, which proves that optimizing the contrastive loss function could help improve the inner-entity consistency and avoid *Entity Representation Segmentation in Label-semantics*. As the ablation of our proposed two-stage learning framework, we remove the first stage of contrastive learning and optimize the multi-task loss(i.e. the sum of InfoNCE loss and finetuning loss). The performance deteriorates even worse than only finetuning the model. This may be due to the fact that joint training for multiple objectives simultaneously may make the learning mission unclear, letting the training process uncontrollable. When removing filtering operations on `Non-Entity` tokens and performing contrastive learning on all contrastive pairs, the performance also decreases slightly, indicating that these meaningless contrasting may add noises to parameter optimization.

Models	OntoNotes	MSRA	Weibo	Resume
NerCo	**83.62**	**96.29**	**72.79**	**96.82**
w/o Contrastive learning	82.85	95.86	68.68	96.10
w/o Two-stage learning	80.83	95.24	68.32	95.36
w/o Filtering Non-Entity tokens	83.41	96.13	71.93	95.74

Table 5: An ablation study of the proposed model. (1)**w/o Contrastive learning**: Only fine-tuning is implemented, and *Semantic Class* guided contrastive learning is not used. (2)**w/o Two-stage learning**:Multi-task learning is conducted, where we optimize the sum of fine-tuning loss and InfoNCE loss in a single stage. (3)**w/o Filtering Non-Entity tokens**: Filtering operation on Non-Entity tokens as queries is removed.

5.3 How NerCo Brings Improvement

To explore why NerCo works better than the baseline, we visualize the representations of entity tokens encoded by both models and also demonstrate the learning process of our method. We use t-SNE to project these representations into two dimensions, taking the test split of MSRA as a representative example.

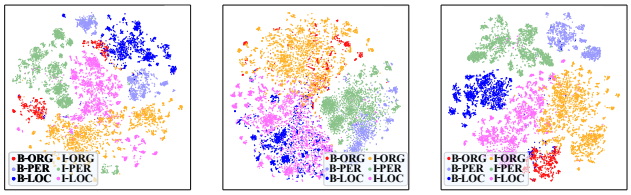
As shown in Figure 4, we select 6 classes(respectively `B-ORG`,`I-ORG`, `B-PER`, `I-PER`, `B-LOC` and `I-LOC`). Figure 4(a) presents the clustering of FLAT's output embeddings, and the right two figures illustrate our proposed two-stage learning strategy. Figure 4(b) shows the token representations after contrastive learning of pulling tokens of the same *Semantic Class* together, regardless their differences in chunk tags(i.e. `B-PER` and `I-PER` are considered the same *Semantic Class*). We can see tokens of the same *Semantic Class* are clustered together, and these three clusters are scattered separately in different directions, as desired by our proposed contrastive learning. And Figure 4(c) is the visualization of the final representations after the second stage finetuning. We can see that after NerCo's two-stage training, the clusters show much clearer margins and token representations within each cluster are significantly more tightly packed together, compared with the situation of that in FLAT. Since NerCo can generate such better token representations, it is much easier for it to make correct predictions near the decision boundaries, which naturally leads to performance gains in downstream tasks.

5.4 Case Study

Table 6 presents two tagging examples predicted by our model NerCo and the baseline FLAT. In the first example from MSRA, the word '联合'(Uniting) in '欧洲联合'(European uniting) is very similar to the noun '联盟'(Union) of another entity '欧洲联盟'(European Union), leading to an extremely blurry and confusing entity boundary for correctly detecting the entity '欧洲'(European). It turned out that FLAT mistakenly predicted '欧洲联合'(European uniting) as an `ORG` entity, presumably confusing the verb '联合'(Uniting) with the noun '联盟'(Union) and failing to recognize the right `LOC` entity '欧洲'(European). However, our model correctly detected the entity's boundary and classified it into the right type. The possible reason is that our method's first stage contrastive learning makes inner tokens' representations more consistent

Case 1 from MSRA Dataset								
Sentence	[中国]LOC支持[欧洲]LOC联合和[欧盟]ORG一体化进程 [China]LOC supports [European]LOC uniting and the process of [EU]ORG integration							
Characters	中 和	国 欧	支 盟	持 一	欧 体	洲 化	联 进	合 程
Gold	[B-LOC O	I-LOC] LOC [B-ORG	O I-ORG]ORG	O O	[B-LOC O	I-LOC] LOC O	O O	O O
FLAT	[B-LOC O	I-LOC] LOC [B-ORG	O I-ORG]ORG	O O	[B-ORG O	I-ORG O	I-ORG O	I-ORG]ORG O
Our Method	[B-LOC O	I-LOC] LOC [B-ORG	O I-ORG]ORG	O O	[B-LOC O	I-LOC] LOC O	O O	O O
Case 2 from Ontonotes Dataset								
Sentence	[成都通用医疗设备（西南）有限公司]ORG [Chengdu General Medical Equipment (Southwest) Co., Ltd]ORG							
Characters	成 （	都 西	通 南	用 ）	医 有	疗 限	设 公	备 司
Gold	[B-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG]ORG
FLAT	[B-ORG I-ORG]ORG	I-ORG [B-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG]ORG
Our Method	[B-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG	I-ORG I-ORG]ORG

Table 6: Examples of tagging results.



(a) FLAT (b) NerCo(stage 1) (c) NerCo(stage 2)

Figure 4: A t-SNE Visualization on representations of entity tokens on MSRA. (a) shows the representations output by FLAT. (b) shows the intermediate token representations after *Semantic Class* guided contrastive learning. (c) shows the final token representations of our method. Our proposed NerCo generates clearer margins between clusters. To give a more specific example, tokens of `I-ORG`, dispersed at the lower part of the canvas (a), are mixed up with `I-LOC` tokens and `I-PER` tokens. Also, tokens clusters of `I-LOC`, `B-ORG`, `B-PER` and `I-PER` are split into several parts, without distinct boundaries outwards. In contrast, token representations of NerCo in (c) are packed closely within their clusters.

and similar, while keeping them away from the representations of outer-entity tokens. As a result, it forms a clear boundary for the sequence labeling model to detect in the second stage. In the second example from Ontonotes, the whole sequence of '成都通用医疗设备（西南）有限公司'(Chengdu General Medical Equipment (Southwest) Co., Ltd) is a complete company name, apparently forming an `ORG` entity. However, FLAT disappointingly segmented it into two parts and wrongly regarded the sequence as two `ORG` entities. We speculate that FLAT may be disturbed by the word '西南'(Southwest), because these words usually mean the beginning of a new entity. At the same time, the unusual length also limits the recognition of FLAT. In contrast, NerCo could learn cohesive representations corresponding to entities after contrastive learning stage, thus easily tackling the long entity and naturally making the accurate prediction.

6 Conclusion

In this work, we propose a novel framework NerCo: Named Entity recognition with Contrastive Learning to address the representation deficiency which we term *Entity Representation Segmentation in Label-semantics* in sequence labeling NER. We naturally introduce contrastive learning to harness the representation learning process such that the token representations should be similar within the same *Semantic Class*, and discriminative of different. We simply construct in-batch contrastive pairs based on entity's label-semantics and utilize InfoNCE loss for the first stage. Finetuning on the learned label-semantics based model using sequence labeling is followed for final entity prediction. Experiments demonstrate that our proposed two-stage method benefits model's generalization performance in Chinese named entity recognition. Our results claim that the simple and long-adopted sequence labeling methods are powerful instead of out-of-time after addressing the above problem. Future work will concentrate on adapting the idea of NerCo to more complex settings, such as nested and non-continuous NER, and integrating more advanced and powerful contrastive learning techniques to further enhance our method.

Acknowledgements

This research was partially supported by the National Key Research and Development Project of China No. 2021ZD0110700, the Key Research and Development Project in Shaanxi Province No. 2022GXLH-01-03, the National Science Foundation of China under Grant Nos. 62037001, 62002282, 62250009 and 61721002, the Major Technological Innovation Project of Hangzhou No. 2022AIZD0113, the "Pioneer" and "Leading Goose" R&D Program of Zhejiang No. 2022C01107, the China Postdoctoral Science Foundation No. 2020M683492, the MOE Innovation Research Team No. IRT_17R86, and Project of XJTU-SERVYOU Joint Tax-AI Lab.

References

[Che *et al.*, 2013] Wanxiang Che, Mengqiu Wang, Christopher D. Manning, and Ting Liu. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 52–62, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Chiu and Nichols, 2016] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370, 2016.

[Das *et al.*, 2022] Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. Container: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, 2022.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Diao *et al.*, 2019] Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. Zen: Pre-training chinese text encoder enhanced by n-gram representations. *arXiv preprint arXiv:1911.00720*, 2019.

[Fang *et al.*, 2020] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

[Gao *et al.*, 2021] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Gillick *et al.*, 2019] Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. Learning dense representations for entity retrieval. *arXiv preprint arXiv:1909.10506*, 2019.

[Giorgi *et al.*, 2021] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online, August 2021. Association for Computational Linguistics.

[Gu *et al.*, 2022] Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Yi Zheng, Baoxing Huai, and Nicholas Jing Yuan. Delving deep into regularity: A simple but effective method for Chinese named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1863–1873, Seattle, United States, July 2022. Association for Computational Linguistics.

[Gui *et al.*, 2019a] Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, pages 4982–4988, 2019.

[Gui *et al.*, 2019b] Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuanjing Huang. A lexicon-based graph neural network for Chinese NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1040–1050, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Hadsell *et al.*, 2006] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[He and Wang, 2008] Jingzhou He and Houfeng Wang. Chinese named entity recognition and word segmentation based on character. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*, 2008.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[Henderson *et al.*, 2017] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

[Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[Jaiswal *et al.*, 2020] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[Levow, 2006] Gina-Anne Levow. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, 2006.

[Li *et al.*, 2014] Haibo Li, Masato Hagiwara, Qi Li, and Heng Ji. Comparison of the impact of word segmentation on name tagging for chinese and japanese. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2532–2536, 2014.

[Li *et al.*, 2020] Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*, 2020.

[Li *et al.*, 2022a] Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973, 2022.

[Li *et al.*, 2022b] Linyang Li, Yong Dai, Duyu Tang, Zhangyin Feng, Cong Zhou, Xipeng Qiu, Zenglin Xu, and Shuming Shi. Markbert: Marking word boundaries improves chinese bert. *arXiv preprint arXiv:2203.06378*, 2022.

[Liu *et al.*, 2010] Zhangxun Liu, Conghui Zhu, and Tiejun Zhao. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? In *International Conference on Intelligent Computing*, pages 634–640. Springer, 2010.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2021a] Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. Lexicon enhanced chinese sequence labeling using bert adapter. *arXiv preprint arXiv:2105.07148*, 2021.

[Liu *et al.*, 2021b] Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. Lexicon enhanced Chinese sequence labeling using BERT adapter. 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 5847–5858, Online, August 2021. Association for Computational Linguistics.

[Ma and Hovy, 2016] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[Ma *et al.*, 2020] Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. Simplify the usage of lexicon in Chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960, Online, July 2020. Association for Computational Linguistics.

[Peng and Dredze, 2016] Nanyun Peng and Mark Dredze. Improving named entity recognition for chinese social media with word segmentation representation learning. *arXiv preprint arXiv:1603.00786*, 2016.

[Ramshaw and Marcus, 1995] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*, 1995.

[Sang and De Meulder, 2003] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[Si *et al.*, 2022] Shuzheng Si, Shuang Zeng, Jiaxing Lin, and Baobao Chang. SCL-RAI: Span-based contrastive learning with retrieval augmented inference for unlabeled entity problem in NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2313–2318, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[Sui *et al.*, 2019] Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3830–3840, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Sun *et al.*, 2019] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Weischedel *et al.*, 2011] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 2011.

[Wu *et al.*, 2020] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*, 2020.

[Yan *et al.*, 2019] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*, 2019.

[Zhang and Yang, 2018] Yue Zhang and Jie Yang. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*, 2018.

[Zhang *et al.*, 2023] Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. Optimizing bi-encoder for named entity recognition via contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[Zhu and Li, 2022] Enwei Zhu and Jinpeng Li. Boundary smoothing for named entity recognition. *arXiv preprint arXiv:2204.12031*, 2022.