MCL: Multi-Granularity Contrastive Learning Framework for Chinese NER

Shan Zhao[1], **ChengYu Wang**[1,2*], **Minghao Hu**[3], **Tianwei Yan**[2], **Meng Wang**[1]

[1] School of Computer Science and Information Engineering, HeFei University of Technology, HeFei, China
[2] College of Computer, National University of Defense Technology, Changsha, China
[3] Information Research Center of Military Science, PLA Academy of Military Science, Beijing, China
2022800040@hfut.edu.cn, {chengyu99a, huminghao16, augusyan57, eric.mengwang}@gmail.com

Abstract

Recently, researchers have applied the word-character lattice framework to integrated word information, which has become very popular for Chinese named entity recognition (NER). However, prior approaches fuse word information by different variants of encoders such as Lattice LSTM or Flat-Lattice Transformer, but are still not data-efficient indeed to fully grasp the depth interaction of cross-granularity and important word information from the lexicon. In this paper, we go beyond the typical lattice structure and propose a novel Multi-Granularity Contrastive Learning framework (MCL), that aims to optimize the inter-granularity distribution distance and emphasize the critical matched words in the lexicon. By carefully combining cross-granularity contrastive learning and bi-granularity contrastive learning, the network can explicitly leverage lexicon information on the initial lattice structure, and further provide more dense interactions of across-granularity, thus significantly improving model performance. Experiments on four Chinese NER datasets show that MCL obtains state-of-the-art results while considering model efficiency. The source code of the proposed method is publicly available at https://github.com/zs50910/MCL

Introduction

Named Entity Recognition (NER) mainly involves determining entity boundaries and categories and aims to identify important entities in the text. It plays an important role in many downstream natural language processing (NLP) tasks, such as relation extraction (Bunescu and Mooney 2005) and knowledge base population (Zhang et al. 2017).

Due to the additional word segmentation process of Chinese (Duan and Zheng 2011), Chinese NER is more difficult compared to English NER. One intuitive way to perform Chinese NER is a pipeline task: word segmentation and word sequence labeling (Yang et al. 2017). The major disadvantage of such a framework is error propagation: word segmentation errors negatively impact the identification of named entities (Peng and Dredze 2015). With awareness of the existing word segmentation errors, Zhang et al.(2018) firstly introduces a lattice structure to incorporate word boundary information for character sequences by a variant of LSTM. Soon, the lattice structure becomes a paradigm of following-up works (Liu et al. 2019; Gui et al. 2019a,b; Li et al. 2020; Ma et al. 2020; Zhao et al. 2021).

Most existing lattice-based studies focus on incorporating word information into the general encoder framework. For example, Gui et al. (2019a) design a convolutional neural network (CNN) with a rethinking mechanism to encode matched words at different window sizes. Liu et al.(2019) and Ma et al.(2020) exploit word-character LSTM to encode concatenation of character and word embeddings. Moreover, Gui et al. (2019b) introduce a lexicon-based graph neural network (GNN) that achieves Chinese NER as a node classification task. The whole-sentence semantics and word ambiguities can be effectively tackled. After that,Transformer-based models (Li et al. 2020) propagate lexicon information by relative position encoding. DCSAN (Zhao et al. 2021) integrates separate features, character representation, and lexicon information by a shallow fusion layer (cross-attention) for Chinese NER. Although these approaches have achieved promising results, they solely integrate character representations and lexicon features into a character-based model by different variants of encoders, and fail to fully exploit the multi-granularity features (e.g., character feature and word feature). Thus, their performance relies heavily on the quality of a well-designed encoder, and the further interaction of cross-granularity and important word information from the lexicon are not fully grasped. Therefore, our focus is on how to make the initial lattice framework efficient enough to leverage lexicon information at a multi-granularity level.

Inspired by the work about contrastive learning (He et al. 2020; Misra and Maaten 2020; Li et al. 2021b), we propose a Multi-Granularity Contrastive Learning (MCL) method to help the lattice framework learn efficiently. To achieve this, we first design cross-granularity contrastive learning (CCL) , which minimizes the distance of representations of the character and corresponds to matched word, and maximize that of non-paired character and word. Then, we further construct bi-granularity contrastive learning (BCL) that aims to emphasize word information through pulling positive samples (a subset of original embedding concatenation of character-word pairs) closer and pushing apart negative ones (individual character embeddings), as shown in Figure 2. The key insight of CCL is to close the representation gap between different granularity to encourage character-word

[*] Corresponding Author

interactions as much as possible, while BCL explicitly facilitates our model to be more sensitive to important word information. In this way, our model explicitly leverages lexicon information at data-level on the initial lattice structure, and further provide deeper interactions of across-granularity.

Finally, we conducted extensive experiments on four NER datasets to evaluate the proposed model. Experimental results show that MCL can achieve state-of-the-art performance. In particular, we obtain 78.59%, 95.86%, 95.79%, and 68.17% F1 on OntoNotes, MSRA, Resume, and Weibo datasets respectively.

Related Work

The key of the proposed MCL method is to leverage lexical information by a contrastive learning framework. So, we focus on the lexical-based methods and contrastive learning methods in the literature.

Lexicon-based Chinese NER

In Chinese NER, many recent studies use word matching methods to enhance character-based models. There are four main types of neural networks structure for Chinese NER, including CNN-based, RNN-based, Graph-based, and Transformer-based structures. **RNN-based.** Zhang and Yang (2018) first introduced a lattice LSTM to avoid the error propagation of segmentation, in which word information is integrated into a shortcut path between the start and the end of characters of the word. SoftLexicon (LSTM) (Ma et al. 2020) introduced lexical information through label and probability methods at the character representation layer. **CNN-based.** Gui et al. (2019a) proposed the LR-CNN that can model all the characters and potential words that match the sentence in parallel using a rethinking mechanism. **Graph-based.** Gui et al. (2019b) and Sui et al. (2019) converted lattice NER into a node classification task by constructing a graph to incorporate the word information by graph neural networks. **Transformer-based.** Li et al. (2020) converted the lattice structure into a flat structure consisting of spans with Transformer architecture. Zhao et al.(2021) leveraged cross-attention to capture interactions over word-character pairs. MECT (Wu, Song, and Feng 2021) is an extension to FLAT (Li et al. 2020), which use extra multi-metadata embedding to integrate Chinese character features with the radical-level embedding. However, these models tend to over-rely on the quality of variants of the encoder. In contrast, we integrated word information on the initial lattice framework by contrastive learning. We construct the model from a data perspective and our model is adapted to mainstream encoders.

Contrastive Learning

Recently, Chen et al. (2020) proposed influential SimCLR by refining the idea of contrastive learning with the help of modern image augmentation techniques to learn robust sets of features. Soon, contrastive learning becomes a rising domain and has achieved promising results in various tasks(He et al. 2020; Misra and Maaten 2020; Li et al. 2021b,a). In NLP, contrastive learning aims to learn a semantic space such that embeddings of similar text inputs are close to each other while repelling dissimilar ones (Li et al. 2022c). Chen et al. (2021) and Li et al.(2022a) introduce contrastive learning for distantly supervised relation extraction. The former regard the multi-instance learning as the relational triple encoder and constraint positive pairs against negative pairs for each instance. The latter proposes a hierarchical contrastive learning framework to reduce noisy sentences. Then, Hu et al. (2022) propose a unified framework to combine graphs and contrastive learning to better incorporate valuable features for promoting impression generation. Moreover, Das et al. (2022) optimize the inter-token distribution distance for Few-Shot NER by contrastive learning technique. Inspired by these studies, we apply contrastive learning to Chinese NER. This is also the first attempt to incorporate lexical information using a contrast learning framework.

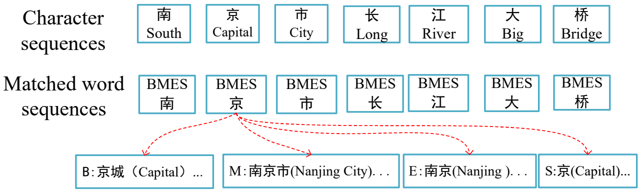


Figure 1: Soft-lexicon strategy used (Ma et al. 2020); “BMES” denotes the the aligned word for the character. B, M and E means all lexicon matched words on a sentence that begin, middle, and end with the character respectively. S is the single-character word.

The Proposed Model

In this section, we introduce the proposed Multi-Granularity Contrastive Learning Framework for Chinese NER (MCL) in detail, as illustrated in Figure 2. Characters and matched words are first represented as distributed representations from pre-trained characters and word embedding. Cross-granularity contrastive learning is then designed to encourage character-word interactions as much as possible by minimizing the representation gap of character-word pairs and maximizing that of irrelevant characters and words. After that, we propose another bi-granularity contrastive learning, which is more sensitive to word boundary information on character sequences. Finally, we apply a conditional random field (CRF) layer to perform the decoding for Chinese NER.

Input Representation

We first introduce input representation including character-level and word-level respectively.

Character Representation Character embeddings are used to map discrete characters into continuous input vectors. Given a Chinese input sentence $s = [c_1, c_2..., c_n]$, where c_i represents the i-th character, we map each character into a real-valued embedding to express its semantic and syntactic meaning. Each character c_i is represented as:

$$x_i = \mathrm{e}^c(c_i), x_i \in \mathbb{R}^d \tag{1}$$

where $\mathrm{e^c}$ denotes a pre-trained character embedding lookup table. The character feature representations can be obtained by:

$$X = [x_1, x_2, x_3, ..., x_n] \in \mathbb{R}^{n*d} \tag{2}$$

Word Representations Regarding the strategy for selecting matched word-character pairs from the lexicon, the soft-lexicon feature strategy proposed by Ma et al.(2020) is widely used for its better adaptability. To unify the word-character representation space, we also use the soft-lexicon feature strategy, as shown in Figure 1. This strategy selects a fixed-dimensional vector that is composed of four word sets marked by the four segmentation labels "BMES", as the aligned word for each character. For example, the word set B (京) consists of all lexicon matched words on the sentence s that begin with the character "京". Similarly, M (京) consists of all lexicon-matched words in the middle of which character "京" occurs, E (京) consists of all lexicon matched words that end with the character "京", and S (京) is the single-character word comprised of character "京". When a word set is empty, we will set a special word "none" to it to indicate this situation.

Generally, the aligned word w_i for each corresponding character c_i is represented as:

$$y_i^{bmes} = [v(b(c_i)); v(m(c_i)); v(e(c_i)); v(s(c_i))] \tag{3}$$

$$y_i^{bmes} \in \mathbb{R}^{4d} \tag{4}$$

where v denotes the function that maps a single word set to a dense vector. The function works as:

$$v(p) = \frac{1}{Z} \sum_{w \in p} (z(w) + b')e^w \tag{5}$$

where $z(w)$ denote the frequency of w_c occurring in the statistic data set; w_c is the character sequence constituting w; e^w represents a pre-trained word embedding lookup table; b denotes the value that there are 10% of training words occurring less than b times within the statistic data set. Z can be computed by:

$$Z = \sum_{w \in (B \sqcup M \sqcup E \sqcup S)} z(w) + b' \tag{6}$$

To facilitate calculation, we utilize a linear projection to transform dimensions, and finally, word feature representations can be obtained as:

$$Y^{bmes} = \mathrm{Linear}[y_1^{bmes}, y_2^{bmes}, ..., y_n^{bmes}] \in \mathbb{R}^{n*d} \tag{7}$$

Encoding and Decoding

We employ BiLSTM as our encoder, which is superior in building contextualized representations for various NLP tasks. In order to be consistent with the latter contrastive learning processes, the aforementioned character features X are first encoded by an MLP layer, which contains a nonlinear layer.

$$X^{mlp} = \mathrm{MLP}([\mathrm{x_1, x_2, x_3, ..., x_n}]) \in \mathbb{R}^{\mathrm{n*d}} \tag{8}$$

Then, taking the character features X^{mlp} and word features Y^{bmes} as inputs, a BiLSTM can be used to output hidden representations $H \in \mathbb{R}^{n*4d}$ as:

$$\mathrm{H} = \mathrm{BiLSTM}[\mathrm{X^{mlp}}; \mathrm{Y^{bmes}}] \tag{9}$$

A standard CRF layer is used to predict NER taggings, which takes H as inputs, and outputs a sequence of predicted tagging probabilities $T = [t_1, ..., t_n]$. Let $T^{'}$ denotes an arbitrary label distribution sequence, the probability of the label sequence T can be calculated using a softmax function:

$$Pr(T|H) = \frac{\prod_{i=1}^{n} \varphi_n(t_{n-1}, t_n, H)}{\sum_{t' \in T'} \prod_{i=1}^{n} \varphi_n(t'_{n-1}, t'_n, H)} \tag{10}$$

where $\varphi_n(t_n, t_{n-1}, H) = \exp(W_n H + b_n)$ is the scoring function and W_n and b_n are the weight vector and bias. During training, we optimize model parameters by minimizing the following conditional likelihood:

$$\mathcal{L}_{task} = -\mathrm{log} Pr(T|H) \tag{11}$$

Multi-Granularity Contrastive Learning

Taking the aforementioned character granularity and word granularity as inputs, we perform multi-granularity contrastive learning by carefully combining cross-granularity contrastive learning (CCL) and bi-granularity contrastive learning(BCL).

CCL Only relying on a network encoder to fuse word boundary information into character sequences still lacks the capability to fully grasp the interaction of cross-granularity from simple embedding concatenation. Since the character and word embedding is relatively static in training or testing process. Recently, contrastive learning has shown strong power in learning and distinguishing significant knowledge by concentrating positive samples and contrasting with negative samples, and results in promising performance improvement in many tasks. Unlike traditional contrastive learners (Gao, Yao, and Chen 2021; Yan et al. 2021) that optimize similarity objective between sentence-level representations, we proposed a token-level cross-granularity contrastive learning (CCL) because NER is a task that is sensitive to tokens. CCL minimizes the distribution distance of character and matched word and maximizes that of non-paired character and word, which aims to capture deeper interactions of across-granularity.

Since the character and word embedding are static, the aforementioned character and word features X and Y^{bmes} are first encoded by a MLP layer. Then, given a updating character representation x_i^{mlp} as original one, we take a matched word representation $y_{i,mlp}^{bmes}$ as the positive example. To construct the negative examples, we take the remaining word representations in the sentence that do not repeat with positive examples. The number of negative cases in this way will is just interrelated to the sentence length removing duplicate characters. Next, the objective of contrastive learning is to minimize the following loss, which can be formulated as:

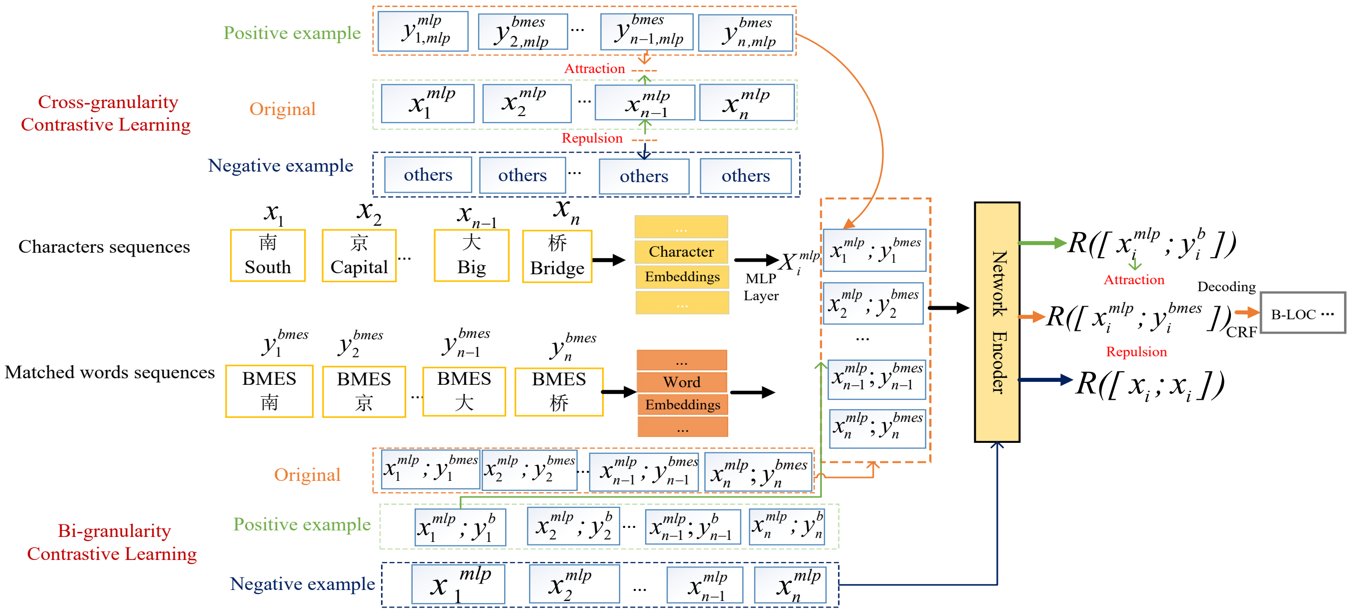


Figure 2: Overall flowchart of MCL. Characters and matched words are first represented as distributed representations from character and word embedding. Cross-granularity contrastive learning is then designed to encourage character-word interactions as much as possible by minimizing the representation gap of character-word pairs and maximizing that of irrelevant characters and words. After that, we propose another bi-granularity contrastive learning, which is more sensitive to word boundary information on character sequences. Finally, we apply a conditional random CRF to perform the decoding for Chinese NER. Others denote the remaining character and word representations in the sentence that do not repeat with the original and positive examples.

$$\mathcal{L}_{bcl} = -log\frac{e^{sim(x_i^{mlp}, y_{i,mlp}^{bmes})/\tau_1}}{\sum_{i\neq j}^{T} e^{sim(x_i^{mlp}, y_{j,mlp}^{bmes})/\tau_1}} \tag{12}$$

where, $sim()$ calculates the similarity of different characters. T is the number of remaining characters after removing duplicate characters in a sentence. τ_1 is the temperature.

BCL Inspired by incorporating extra knowledge in radiology findings summarization task (Hu et al. 2022), we design another token-level contrastive learning, namely bi-granularity contrastive learning. We expect our model to be more sensitive to word information on character sequences. For this purpose, the word-character pairs $[x_i^{mlp}; y_i^{bmes}]$ are original examples and the subset $[x_i^{mlp}; y_i^b]$ are considered as positive examples and the corresponding character representation x_i^{mlp} in the sentence as the negative examples. y_i^b denotes the representations of matched words on s that begin with the character c_i. Since we aim to enhance word information sensitivity in character sequences instead of expanding differences between various characters in one mini-batch. we do not consider the other characters in the same mini-batch as the negative examples. Moreover, we use the linear layer so that the dimensions are equal. We can calculate the training objective of the contrastive module:

$$h_i^{ori} = f_\theta([x_i^{mlp}; y_i^{bmes}]) \tag{13}$$

$$h_i^{pos} = f_\theta([x_i^{mlp}; y_i^b]) \tag{14}$$

$$h_i^{neg} = f_\theta(x_i^{mlp}) \tag{15}$$

$$\mathcal{L}_{ccl} = -log\frac{e^{sim(h_i^{ori}, h_i^{pos})/\tau_2}}{e^{sim(h_i^{ori}, h_i^{pos})/\tau_2} + e^{sim(h_i^{ori}, h_i^{neg})/\tau_2}} \tag{16}$$

where τ_2 is the temperature, and $f_\theta()$ is the BiLSTM parameterized by θ.

During the training of MCL, the model can be optimized by jointly minimizing the contrastive training loss and NER loss:

$$\mathcal{L} = \mathcal{L}_{task} + \mathcal{L}_{bcl} + \mathcal{L}_{ccl} \tag{17}$$

Experimental Setup

To evaluate the performance of our method, we conduct experiments on four datasets, including OntoNotes 4.0 (Weischedel et al. 2011), Weibo (Peng and Dredze 2015), MSRA (Levow 2006), and Chinese Resume dataset (Zhang and Yang 2018). The corpus of MSRA and OntoNotes comes from news, the corpus of Weibo comes from social media, and the corpus of Resume comes from the resume data in Sina Finance. Since the lexicon-based model, SoftLexicon (LSTM) (Ma et al. 2020) adopts BiLSTM-CRF as backbone network, we therefore set it as the baseline model. We adopt standard Precision (Prec), Recall (Rec), and F1 score to evaluate the model.

[2] https://github.com/ljynlp/W2NER

[3] https://github.com/zs50910/DCSAN-for-Chinese-NER

Models	OntoNotes	MSRA	Resume	Weibo
Lattice LSTM (2018)	73.88	93.18	94.46	58.79
LR-CNN (2019a)	74.45	93.71	95.11	59.92
LGN (2019b)	74.85	93.64	95.41	60.15
CGN (2019)	74.79	93.47	94.12	63.09
FLAT (2020)	76.45	94.12	95.45	60.32
MECT (2021)	76.92	94.32	95.89	63.30
DCSAN (2021)	76.23	94.86	95.02	65.26
W^2NER* (2022b)	75.66	94.55	94.26	64.32
Baseline	75.64	93.66	95.53	61.42
MCL	**78.59**	**95.12**	**95.96**	**68.17**
LEBERT (2021)	82.08	95.70	96.08	70.75
BERT	80.14	94.95	95.53	68.20
BERT+MCL	**82.96**	**96.11**	**96.46**	**73.08**

Table 1: We compare our MCL with recent state-of-the-art models on four Chinese benchmarks. * denotes the model is no-lexical. The results of W^2NER and DCSAN were obtained by running the public codes[2] [3] with the pre-trained character embedding used in our experiment. The rest results of the models are taken from the respective original paper.

Implementation Details

We regularize our network using dropout with a rate tuned on the development set (the dropout rate is 0.5 for embeddings and encoder). We utilize 1 layer encoder in our network and set the dimensionality of hidden size was set to 100 for Weibo and 300 for the rest three datasets. The pre-trained character embedding is the same as (Zhang and Yang 2018). Following (Zhao et al. 2021), we use the word embedding dictionary (Song et al. 2018) as default lexicon. The learning rate was set to 0.007 for all datasets with Adamax. The temperatures are 0.3 for CCL and 0.05 for BCL.

Overall Results

As shown in Table 1, our model outperforms other models on four Chinese NER datasets. It can be seen that our model achieves state-of-the-art performance by obtaining 78.59, 95.12, 95.96, and 68.17 F1 respectively. Compared with the best results among Lattice LSTM, LR-CNN, LGN, CGN, FLAT, MECT and DCSAN, our approach gets absolute F1 improvements of 1.67%, 0.26%, 0.07% and 2.91% on datasets respectively. When compared to the baseline model, we find stronger performance improvement with respect to OntoNotes (+3.15%), Resume (+0.43%), MSRA(+1.46%) and Weibo (+6.75%). The above results indicate that MCL is able to better leverage word information and learn character-word pair representations compared to other lexicon-based models. To investigate the comprehensiveness of the advantages of our model. We also compare our model with the state-of-the-art no-lexical model W^2NER. We can observe that our model also outperforms W^2NER by 2.26 in the average F1 score. Besides the single-model evaluation on the four datasets, we also evaluated MCL when combined with the Pre-trained Model, BERT. The results of BERT are taken from the FLAT paper. We can find that MCL further improves the performance of BERT significantly. Moreover, we compare the proposed model with a lexicon-based Pre-trained Model, LEBERT, and show the results in Table 1. It can be seen that our model outperforms the LEBERT on all datasets. Especially, our proposed model has a significant improvement (+2.33%) on Weibo dataset.

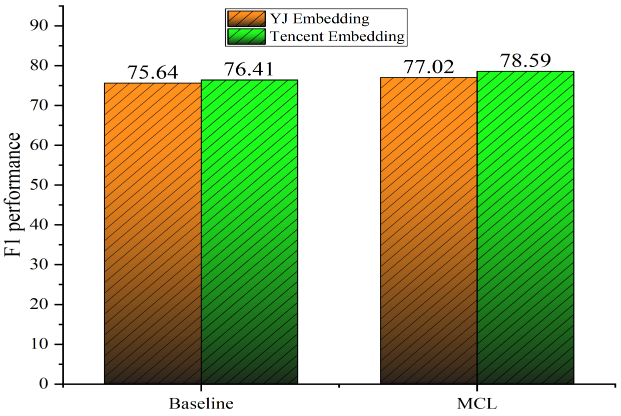


Figure 3: Comparison of our MCL and baseline against different lexicons on OntoNotes dataset, where "YJ embedding" denotes the lexicon used in (2018) and "Tencent embedding" denotes the lexicon used in (2018; 2021).

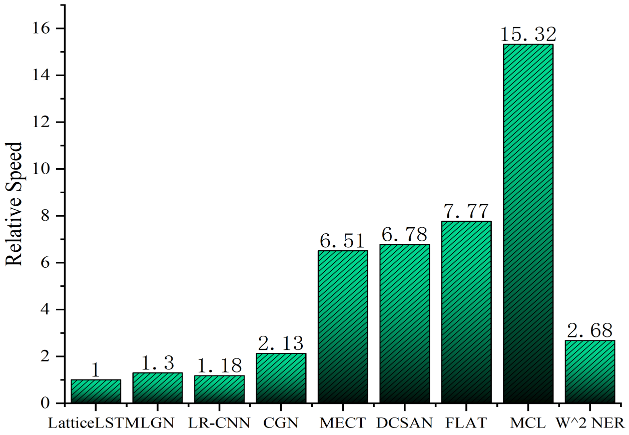


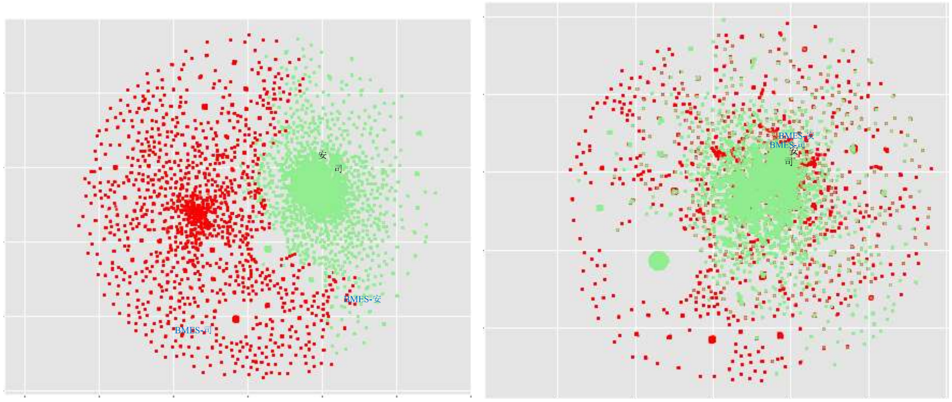
Figure 4: Relative inference speed of different models, compared with Lattice LSTM. Due to variable-sized set of matched words, LR-CNN are non-batch parallel.

F1 against Different Lexicons

To explore the effectiveness of different lexicons, we analyze the performance of our MCL and the baseline against two lexicons, which are used in (2018) and (2018; 2021), on the OntoNotes dataset. The results are shown in Figure 3. It can be seen that models with "Tencent Embedding" outperform models with "YJ Embedding". We think the reason is that "Tencent Embedding" is a larger lexicon that contains over 8000k Chinese characters and words. Moreover,

	Case1: Baseline ✓ MCL ✓
Sentence	北海市的崛起是近年来广西壮族自治区对外开放取得卓著成就的重要标志之一。
	The rise of Beihai City is one of the important indicators that the Guangxi Zhuang Autonomous Region has achieved outstanding success in opening up to the outside world in recent years
Gold Labels	B-GPE M-GPE E-GPE O O O O O O O O B-GPE M-GPE M-GPE M-GPE M-GPE M-GPE E-GPE O O O O O O O O O O O O O O O O O O
Baseline	B-GPE M-GPE E-GPE O O O O O O O O B-GPE M-GPE M-GPE M-GPE M-GPE M-GPE E-GPE O O O O O O O O O O O O O O O O O O
MCL	B-GPE M-GPE E-GPE O O O O O O O O B-GPE M-GPE M-GPE M-GPE M-GPE M-GPE E-GPE O O O O O O O O O O O O O O O O O O
	Case2: Baseline × MCL ✓
Sentence	由中国自主设计建设 、达到当今世界先进技术水平的安阳彩色显像管玻壳有限公司今天建成投产。
	The Anyang Color Image Tube Glass Shell Limited Company, which was designed and built by China itself and reached the level of advanced technology in the world today, was put into production today.
Gold Labels	OB-GPE E-GPE O O O O O O O O O O O O O O O O O O O O B-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG E-ORG O O O O O O O O
Baseline	O B-GPE E-GPE O O O O O O O O O O O O O O O O O O O O B-GPE E-GPE O O O O O O O O O O O O O O O O O O O O O O O
MCL	O B-GPE E-GPE O O O O O O O O O O O O O O O O O O O O B-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG E-ORG O O O O O O O O O O O
	Case2: Baseline × MCL ×
Sentence	黑山头口岸联检部门将原来要二至三天办完的出入境手续改为一天办完
	The joint inspection department at the head of Montenegro has changed the entry and exit procedures that originally took two to three days to be completed in one day
Gold Labels	B-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG M-ORG E-ORG O O O O O O O O O O O O O O O O O O O O O O O
Baseline	O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O
LECL	O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O

Table 2: Examples of OntoNotes dataset.



(a) MCL of before training

(b) MCL of after training

Figure 5: Two UMAP visualisation of embeddings before and after training. The green points is character embedding, the red points is matched words embedding. This figure illustrates that the character and word representations are drawn closer after applying contrastive learning.

when utilizing a larger lexicon, we can observe that the performance gap becomes more obvious. Particularly, using the “Tencent Embedding” lexicon, MCL with contrastive learning is able to improve by 1.31 F1, and then without it was only improved by 0.87. These results demonstrate that our model is more effective in terms of using lexicons.

Computational Efficiency Study

To explore the efficiency of our model, we conducted experiments of inference time on the OntoNotes dataset, as shown in Figure 4. Since baseline and MCL contain the same structure in inference time, we compare the rest models on efficiency. It can be seen that MCL significantly runs faster than compared models. Compared with lexical-based models, MCL runs 15.32, 12.98, 11.78, 7.19,2.35,2.25 and 1.97 times faster than lattice LSTM, LR-CNN, LGN, CGN, DCSAN, MECT, and FLAT respectively. Especially, MCL has 5.71 times the inference-speed compared to the no-lexical model W^2NER. We think the reason is that W^2NER is a complex model containing multiple neural networks such as BiLSTM, multi-granularity 2D convolution, and co-predictor layer. In contrast, MCL just contains 1 layer of BiLSTM and CRF in inference time.

Model	P	R	F1
MCL	77.64	78.48	78.59
- CCL	77.24	74.96	76.08
- BCL	77.03	76.47	76.75
- Both Contrastive Learning	77.28	74.07	75.64

Table 3: Ablations on OntoNotes test set.

Ablation Study

We conduct an ablation study on the OntoNotes test set to investigate the influence of different modules in our proposed model in Table 3. Modules are tested in three ways: (1) We remove CCL and only use BCL to learn word information. In this case, transfer learning is difficult. We find that the F1 score significantly decreases by 2.51, indicating that interaction of cross-granularity is critical for the task. (2) To test the effectiveness of BCL, we just adopt CCL to fuse lexical information. We can find that the performance drops to 76.75 (-1.84%) F1. This indicates that being more sensitive to key word information is necessary for improving the performance of the task. (3) Finally, removing both contrastive learning modules cause model to degenerate into the baseline model, and leads to further worse results on NER (-2.95%), which suggests that leveraging lexicon information at data-level over lattice structure play a vital role in the Chinese NER task.

Qualitative Analysis

To further demonstrate how our approach with contrastive learning explicitly leverages lexicon information, we perform qualitative analysis on three cases, and the results are shown in Table 2. In the first case, there is an overlapped entity “广西壮族自治区 (Guangxi Zhuang Autonomous Region) ”, which is easy to incorrectly recognizes “广西 (Guangxi) ” as an entity. Due to the “广西壮族自治区 (Guangxi Zhuang Autonomous Region) ” as a common entities in the lexicon, the baseline and MCL both correctly detects the GPE-entity. In the second case, there is an organization entity “安阳彩色显像管玻壳有限公司 (The Anyang Color Image Tube Glass Shell Limited Company) ”. It is difficult for baseline to detect the uncommon organization entity since it lacks cross-granularity information, which wrongly recognizes “安阳 (Anyang)” as a GPE-entity. However, MCL is more sensitive to critical word information and can learn better interaction of cross-granularity. For example, the matched word “安阳彩色显像管玻壳有限公司 (The Anyang Color Image Tube Glass Shell Limited Company)” is the crucial word information for character “司 (Company)” and the deep interaction between the character and the matched word is helpful for entity recognition. Finally, for case 3, we can find baseline and MCL both wrongly recognizes “黑山头口岸联检部门(The joint inspection department at the head of Montenegro)” as non-entity. The reason is that “黑山头口岸联检部门(The joint inspection department at the head of Montenegro) ” is a fairly uncommon entity and cannot be matched in the lexicon. These results indicate that lexical information is essential to our task. Furthermore, deep interactions of cross-granularity and being more sensitive to critical word information are also indispensable.

Moreover, in order to visualize the representations of across-granularity, we retrieve the character representation x_i and word representation y_i^{bmes} before and after training, resulting in a mini-batch sample in the high-dimensional space. To facilitate visualization, we apply uniform manifold approximation and projection (UMAP) dimension reduction to reduce the representations to 2-dim in Figure 5. It is first observed that MCL draws the representations across two granularities much closer after training. For example, the character “司 (Company)” and the corresponding matched words “BMES-司 ” are farther apart in the embedded space, but after training, the distance between them is significantly closer. This again justifies the effectiveness of MCL in merging lexical information with the knowledge transfer across different granularity representations.

Conclusion

In this work, we discuss the long-standing lattice framework and argue the lattice is not data-efficient indeed, as it aims to just incorporate word information by different variants of encoders. Thus, we propose a Multi-Granularity Contrastive Learning (MCL) method containing bi-granularity and cross-granularity contrastive learning to boost the word-character lattice performances. The results show that MCL achieves new state-of-the-art performance with highly competitive efficiency.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.72188101, No.62006243).

References

Bunescu, R. C.; and Mooney, R. J. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 724–731.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Shi, H.; Tang, S.; Chen, Z.; Wu, F.; and Zhuang, Y. 2021. CIL: Contrastive Instance Learning Framework for Distantly Supervised Relation Extraction. 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)*, 6191–6200.

Das, S. S. S.; Katiyar, A.; Passonneau, R. J.; and Zhang, R. 2022. 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)*, 6338–6353.

Duan, H.; and Zheng, Y. 2011. A study on features of the crfs-based chinese named entity recognition. *International Journal of Advanced Intelligence*, 3(2): 287–294.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Gui, T.; Ma, R.; Zhang, Q.; Zhao, L.; Jiang, Y.-G.; and Huang, X. 2019a. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 4982–4988. AAAI Press.

Gui, T.; Zou, Y.; Zhang, Q.; Peng, M.; Fu, J.; Wei, Z.; and Huang, X.-J. 2019b. 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)*, 1039–1049.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hu, J.; Li, Z.; Chen, Z.; Li, Z.; Wan, X.; and Chang, T.-H. 2022. Graph Enhanced Contrastive Learning for Radiology Findings Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4677–4688.

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

Li, D.; Zhang, T.; Hu, N.; Wang, C.; and He, X. 2022a. HiCLRE: A Hierarchical Contrastive Learning Framework for Distantly Supervised Relation Extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2567–2578.

Li, J.; Fei, H.; Liu, J.; Wu, S.; Zhang, M.; Teng, C.; Ji, D.; and Li, F. 2022b. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10965–10973.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2021b. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. 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)*, 2592–2607.

Li, X.; Yan, H.; Qiu, X.; and Huang, X. 2020. FLAT: Chinese NER Using Flat-Lattice Transformer. *In Proceedings of ACL,2020.*

Li, Y.; Liu, F.; Collier, N.; Korhonen, A.; and Vulić, I. 2022c. Improving Word Translation via Two-Stage Contrastive Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4353–4374.

Liu, W.; Fu, X.; Zhang, Y.; and Xiao, W. 2021. 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)*, 5847–5858.

Liu, W.; Xu, T.; Xu, Q.; Song, J.; and Zu, Y. 2019. An Encoding Strategy Based Word-Character LSTM for Chinese NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2379–2389.

Ma, R.; Peng, M.; Zhang, Q.; Wei, Z.; and Huang, X.-J. 2020. Simplify the Usage of Lexicon in Chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5951–5960.

Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6707–6717.

Peng, N.; and Dredze, M. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 548–554.

Song, Y.; Shi, S.; Li, J.; and Zhang, H. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 175–180.

Sui, D.; Chen, Y.; Liu, K.; Zhao, J.; and Liu, S. 2019. 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)*, 3821–3831.

Weischedel, R.; Pradhan, S.; Ramshaw, L.; Palmer, M.; Xue, N.; Marcus, M.; Taylor, A.; Greenberg, C.; Hovy, E.; Belvin, R.; et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Wu, S.; Song, X.; and Feng, Z. 2021. MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition. 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)*, 1529–1539.

Yan, Y.; Li, R.; Wang, S.; Zhang, F.; Wu, W.; and Xu, W. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Yang, J.; Teng, Z.; Zhang, M.; and Zhang, Y. 2017. Combining Discrete and Neural Features for Sequence Labeling. *arXiv preprint arXiv:1708.07279*.

Zhang, Y.; and Yang, J. 2018. Chinese NER using lattice LSTM. *arXiv preprint arXiv:1805.02023*.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Zhao, S.; Hu, M.; Cai, Z.; Chen, H.; and Liu, F. 2021. Dynamic modeling cross-and self-lattice attention network for Chinese ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14515–14523.