Two Languages Are Better Than One: Bilingual Enhancement For Chinese Named Entity Recognition

Jinzhong Ning[1], **Zhihao Yang**[1*], **Zhizheng Wang**[1],
Yuanyuan Sun[1], **Hongfei LIN**[1], **Jian Wang**[1]
School of Computer Science and Technology, Dalian University of Technology
Jinzhong_Ning@mail.dlut.edu.cn, wzz0727@gmail.com
{yangzh,syuan,hflin,wangjian}@dlut.edu.cn

Abstract

Chinese Named Entity Recognition (NER) has continued to attract research attention. However, most existing studies only explore the internal features of the Chinese language but neglect other lingual modal features. Actually, as another modal knowledge of the Chinese language, English contains rich prompts about entities that can potentially be applied to improve the performance of Chinese NER. Therefore, in this study, we explore the bilingual enhancement for Chinese NER and propose a unified bilingual interaction module called the **A**dapted **C**ross-**T**ransformers with Global **S**parse Attention (**ACT-S**) to capture the interaction of bilingual information. We utilize a model built upon several different ACT-Ss to integrate the rich English information into the Chinese representation. Moreover, our model can learn the interaction of information between bilinguals (inter-features) and the dependency information within Chinese (intra-features). Compared with existing Chinese NER methods, our proposed model can better handle entities with complex structures. The English text that enhances the model is automatically generated by machine translation, avoiding high labour costs. Experimental results on four well-known benchmark datasets demonstrate the effectiveness and robustness of our proposed model.

1 Introduction

"One language sets you in a corridor for life. Two languages open every door along the way."

—Frank Smith, Psycholinguist

Named entity recognition (NER) is the task of determining spans and semantic categories of named entities such as organization (*ORG*), person (*PER*) and location (*LOC*) in given free text. As the cornerstone of a wide range of natural language processing tasks, NER plays an essential role in many downstream tasks, such as relation extraction (Zelenko et al., 2003) and question answer (Diefenbach et al., 2018).

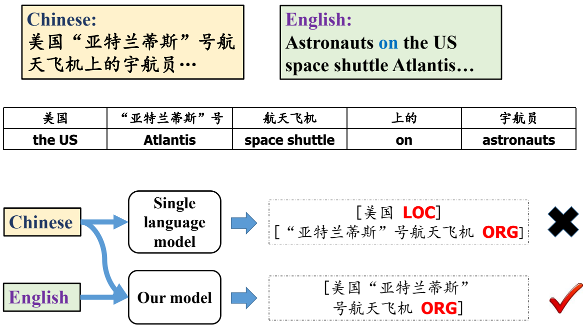


Figure 1: An example of Chinese NER via bilingual enhancement. The entity types of named entities are highlighted.

Compared with English NER, Chinese NER meets a series of challenges caused by the characteristics of Chinese. Aside from the lack of natural word boundary information, Chinese named entities usually vary significantly in length and have complex compositional structures (Dong et al., 2016). At the same time, the annotated data of Chinese NER is relatively scarce, and it is difficult and costly to annotate the data manually (Liu et al., 2022). Hence, without more annotated data, it is a promising approach to improve Chinese NER by leveraging external information resources, which has attracted more and more research attention.

One way to utilize external resources is to perform Chinese NER with bilingual constraints. Previous works (Che et al., 2013; Wang et al., 2013) have demonstrated that the joint use of bilingual information in Chinese and English can significantly improve performance in the Chinese NER task. In addition, incorporating vocabulary knowledge has also become a promising solution(Zhang and Yang, 2018; Li et al., 2020; Mengge et al., 2020). What's more, several studies were proposed to exploit information from other modalities to supplement the representation of Chinese text(Meng et al., 2019;

[*] Corresponding Author

Proceedings of the 29th International Conference on Computational Linguistics, pages 2024–2033
October 12–17, 2022.

Sun et al., 2021b; Wu et al., 2021; Sui et al., 2021).

However, despite the success of the above-mentioned methods in Chinese NER by introducing external information, these methods still have the following limitations. First, the external resources utilized by these methods are mainly obtained manually, which increases the cost significantly. Second, existing methods for boosting NER with bilingual constraints (Che et al., 2013; Wang et al., 2013) rely on bilingual word alignment information and both Chinese and English sentences need to be manually annotated, which limits its usage. Third, as the state-of-the-art approaches based on deep neural networks for Chinese NER, lexical enhancement methods and multimodal methods still fail to effectively handle entities with complex composition structures, which, however, are frequently observed in Chinese NER tasks. The example in Figure 1 illustrates one of such dilemmas in Chinese NER. In this example, due to the complex component structure, the ORG entity “美国‘亚特兰蒂斯’号航天飞机(The US space shuttle Atlantis)" tends to be incorrectly labeled by the NER model as a LOC entity “美国(USA)" and an ORG entity “‘亚特兰蒂斯’号航天飞机(Space shuttle Atlantis)". However, it is encouraging that the clues of the English expression, such as the preposition “on", will potentially alleviate this type of incorrect labeling that often occurs in Chinese NER tasks.

To address the above issues, in this work, we propose to boost the performance of Chinese NER with the unlabelled English text translated from the corresponding Chinese text. English texts are automatically generated through the publicly available neural machine translation API without any extra human labor all the way through. Besides, considering the Chinese texts and corresponding English translations as two different modalities, we perform bilingual enhancement of Chinese NER with multimodal NER approach. Furthermore, based on the fact that a word in the text will only be strongly correlated with a small fraction of words in the translated text, we propose a bilingual interaction enhancement model based on **A**dapted **C**ross-**T**ransformers with Global **S**parse Attention (short for **ACT-S**). The interaction of bilingual information as well as intra-linguistic interaction are taken into account in the our model.

The primary contributions of this work can be summarized as follows: (1) We improve the performance of Chinese NER by bilingual enhancement, based on the unlabeled translated English text automatically generated using the Neural Machine Translation API. To the best of our knowledge, this is the first end-to-end NER method that effectively exploits bilingual information. (2) We further propose the neural module called Adapted Cross-Transformers with Global Sparse Attention to simultaneously model bilingual interactions and inter-lingual interactions. So far as we know, it is the first attempt to use global sparse attention mechanisms for multimodal information interaction. (3) Experimental results on four Chinese NER datasets show that our proposed model achieves superior performance to other strong baseline models.

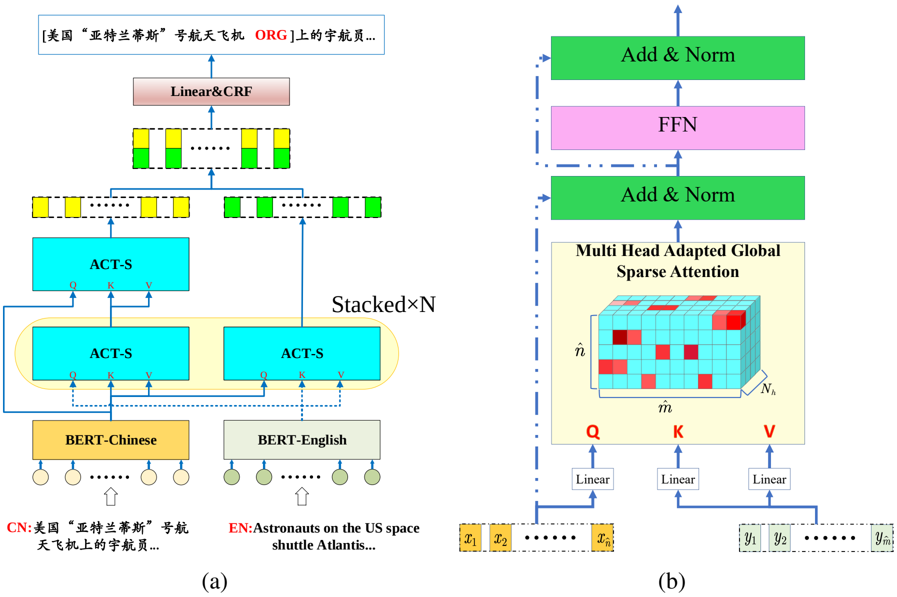
2 Related Work

2.1 Chinese NER with lexicon enhancement

In Chinese NER, a series of recent works focus on introducing lexical boundaries and semantic information by word matching. Zhang *et al.* (2018) proposed to introduce semantic and boundary information of the lexicon through the lattice structure in LSTM. Afterwards, some CNN-based NER lexical enhancement methods, such as LR-CNN (Gui et al., 2019a), were proposed. Graph neural networks have also been applied to Chinese NER word enhancement tasks, a typical one of which is LGN (Gui et al., 2019b). And Transformer-like encoders fusing lexical information are also used for Chinese NER tasks, including PLTE (Mengge et al., 2020) and FLAT (Li et al., 2020). In addition, some work (Ma et al., 2020; Liu et al., 2021) has proposed to fuse lexical information into the word embedding representation instead of integrating word information into the model encoder. However, the lexicon enhanced Chinese NER approach still cannot effectively deal with entities with complex composition structures. And our approach utilizes bilingual clues that can alleviate this problem in Chinese NER.

2.2 Multimodal NER

In recent years, with the development of multimodal information processing technology, the multimodal NER has emerged. In the field of English information extraction, existing multimodal NER works(Yu et al., 2020; Sun et al., 2021a; Zhang et al., 2021) have focused on using image clues to improve NER on Twitter. As for Chinese NER, introducing information from other modalities has


[美国“亚特兰蒂斯”号航天飞机 ORG]上的宇航员...
Linear&CRF
ACT-S
Q
K
V
Stacked×N
ACT-S
ACT-S
BERT-Chinese
BERT-English
CN:美国“亚特兰蒂斯”号航天飞机上的宇航员...
EN:Astronauts on the US space shuttle Atlantis...
Add & Norm
FFN
Add & Norm
Multi Head Adapted Global Sparse Attention
$\hat{n}$
N_h
$\hat{m}$
Q
K
V
Linear
Linear
Linear
x_1
x_2
$x_{\hat{n}}$
y_1
y_2
$y_{\hat{m}}$

(a)

(b)

Figure 2: (a) The overall architecture of our model. (b) The implementation details of ACT-S.

also become a promising solution. On the one hand, the multimodal information of Chinese characters is used to mining the semantics in the structure of Chinese characters(Sun et al., 2021b; Wu et al., 2021). On the other hand, multimodal information of the whole Chinese sentence, such as the audio content of the text (Sui et al., 2021), is also used to improve the word representation for Chinese NER. Compared with existing methods, the method we propose makes use of relative distance information while focusing on strongly correlated units during modal interactions.

2.3 Sequence labelling with bilingual clues or translation

Some previous works (Che et al., 2013; Wang et al., 2013) have demonstrated that constraints in bilingual parallel annotated corpora can be used to improve NER performance in two languages. But different from them, our proposed approach can take advantage of the hints in the unannotated English texts which are automatically translated by Neural machine translation tools. In addition, in cross-lingual sequence labelling tasks, translation methods(Mayhew et al., 2017; Fei et al., 2020; Zhen et al., 2021) are used to migrate annotation information from rich languages to low-resource languages. Unlike these previous works, the only resources used by our model in the additional language are texts with no annotation information. Furthermore, the model automatically learns all bilingual lexical alignment information with no assistance from any bilingual alignment tool.

3 Methodology

3.1 Overall Architecture

Task Formulation: Given a Chinese text $S_c = (cw_1, ..., cw_i, ..., cw_n)$ and its corresponding English translated text $E_c = (ew_1, ..., ew_l, ..., ew_m)$, where cw_i represents the i-th Chinese character and ew_l represents the l-th character of English translated text, the goal of the task is to utilize the information in the bilingual text to determine the spans and types of all named entities in the Chinese text. In this work, we formulate the task as a sequence labeling task. And the BMES (Beginning, Middle, End, Singleton)(Xue, 2003) tagging scheme is adopted.

The architecture of our proposed model is shown in Figure 2. In our model, we absorb the inspiration from the unified multimodal Transformer encoder widely used in vision-language tasks (Tsai et al., 2019; Yu et al., 2020). And similar to MECT (Wu et al., 2021), we introduce distance-aware and direction-aware components in Transformer attention. However, unlike previous works, we propose to use global sparse attention in the Cross-Transformer to reduce the noise in the information fusion process of the two modalities. Each part of the model is introduced in detail in the following sections.

3.2 Word Representations

Since many previous works have proven the effectiveness of the pre-trained language model in Named Entity Recognition task (Li et al., 2020; Wu et al., 2021), we use BERT (Devlin et al., 2019) as our contextualized representation encoder for both Chinese text and English translated text. To fit the BERT encoding procedure, we add two special symbols [CLS] and [SEP] at the beginning and end of the input sentence respectively, and we discard the representation vectors of [CLS] and [SEP] at the end of the BERT encoding computation. If a word is tokenized into several subwords by the Byte Pair Encoding (BPE) algorithm used in BERT, we found empirically that the model performs better when the average pooling method is used to merge the representations of the subwords that belong to the same word into a single representation vector. Thus, we can obtain the word representation generated using BERT:

$$(c_1, c_2, ..., c_n) = BERT_{-Ch}\left(cw_1, cw_2, ..., cw_n\right) \tag{1}$$

$$(e_1, e_2, ..., e_m) = BERT_{-En}\left(ew_1, ew_2, ..., ew_m\right) \tag{2}$$

3.3 Adapted Cross-Transformer with Global Sparse Attention

This section presents our first proposed Adapted Cross-Transformer with Global Sparse Attention (ACT-S) for bilingual information interaction in detail. As illustrated in Figure 2(a), several parameter independent ACT-Ss are used in our model for both inter- and intra-language interactions between bilinguals. The implementation details of ACT-S are shown in Figure 2(b).

Motivation: In multimodal tasks, the widely used cross-modal Transformers (Yu et al., 2020; Sui et al., 2021; Wu et al., 2021) typically use Softmax to normalize the cross-modal attention distribution of each head. As a result, in multi-head cross-modal attention, each unit is represented by all the units in the other modality by multiple different weighted averages. However, it is essential that a word is only associated with a small number of words in the other language in bilingual enhanced Chinese NER. Based on the above observations, we propose for the first time to incorporate the global sparse attention mechanism in a cross-modal Transformer to learn bilingual interactive word representations, which exclude the interference of irrelevant words in the other language. To our knowledge, it is the first time that global sparse attention is used for a multimodal task.

For inputs $X = (x_1, ..., x_u, ..., x_{\hat{n}})$ and $Y = (y_1, ..., y_v, ..., y_{\hat{m}})$, we treat $X \in \mathbb{R}^{\hat{n} \times d}$ as queries, and $Y \in \mathbb{R}^{\hat{m} \times d}$ as keys and values. The input Q, K, and V are obtained by linear transformation of X and Y:

$$Q^{(h)}, K^{(h)}, V^{(h)} = XW_q^{(h)}, YW_k^{(h)}, YW_v^{(h)} \tag{3}$$

where $h \in \{1, 2, ..., N_h\}$ is the index of the h-th attention head and N_h is the number of attention heads. $\left\{W_q^{(h)}, W_k^{(h)}, W_v^{(h)}\right\} \in \mathbb{R}^{d \times d_k}$ are learnable parameters and $d_k = \frac{d}{N_h}$.

To provide the attention mechanism in ACT-S with the ability of both distance perception and direction perception, we adopt the component of sensing relative distances similar to that in MECT(Wu et al., 2021) in the attention matrix computation process:

$$\begin{aligned} \widetilde{A}_{u,v}^{(h)} &= Q_u^{(h)} \left(K_v^{(h)}\right)^T + Q_u^{(h)} R_{u-v}^T \\ &+ K_v^{(h)} R_{u-v}^T + \alpha \left(K_v^{(h)}\right)^T + \beta R_{u-v}^T \end{aligned} \tag{4}$$

$$R_{u-v} = \left[..., \sin\left(\frac{u-v}{10^{6p/d_k)}}\right), \cos\left(\frac{u-v}{10^{6p/d_k}}\right), ...\right] \tag{5}$$

where u is the index of the word in the target language and v is the token in the other language, Q_u, K_v is the query vector and key vector of word x_u, y_v respectively , $R_{u-v} \in \mathbb{R}^{d_k}$ is the relative position encoding, p in Eq. 5 is in the range $\left[0, \frac{d_k}{2}\right]$, $\alpha \in \mathbb{R}^{d_k}$ and $\beta \in \mathbb{R}^{d_k}$ are learnable parameters.

Different from MECT, we add a key bias term $K_v^{(h)} R_{u-v}^T$ to attention matrix to represent the bias of v-th token in the key sequence on certain relative distance and we empirically found that models perform better with it. And in this work, we consider that words in the target language are only relevant to a small number of words in another language. To learn a better representation of the target language words guided by the relevant words in another language. For the first time, we introduce sparse prior information to the global attention distribution via the top-k mask operation $Tkm(\cdot)$, which is formulated as follows:

$$Tkm(\widetilde{A}_{u,v}^{(h)}, k) = \begin{cases} \widetilde{A}_{u,v}^{(h)} & \text{if} \quad \widetilde{A}_{u,v}^{(h)} \in top(\widetilde{A}_{u,:}^{(h)}, k) \\ C & \text{if} \quad \widetilde{A}_{u,v}^{(h)} \notin top(\widetilde{A}_{u,:}^{(h)}, k) \end{cases} \tag{6}$$

Where k is a hyperparameter, masking constant $C \ll 0$, $A_{u,:} = Q_u^{(h)} \left(K^{(h)}\right)^T \in \mathbb{R}^{\hat{m}}$ contains the attention values between $Q_u^{(h)}$ and all keys in $K^{(h)}$,

$top(\widetilde{A}^{(h)}_{u,:}, k)$ is a set containing the largest k values in $\widetilde{A}^{(h)}_{u,:}$. The Adapted Global Sparse Attention matrix for the h-th attention head is then calculated as:

$$A^{(h)}_{u,v} = Tkm(\widetilde{A}^{(h)}_{u,v}, k) \tag{7}$$

And the attention score is calculated as follows:

$$Atten^{(h)} = soft\max\left(A^{(h)}\right)V^{(h)} \tag{8}$$

$$Atten = \left[Atten^{(1)}; ...; Atten^{(N_h)}\right]W^{o} \tag{9}$$

where $W^{o} \in \mathbb{R}^{d\times d}$ is a learnable parameter. Then, we stack the following sub-layers on top to obtain the text representation based on bilingual interaction:

$$\hat{O} = LN\left(X + Atten\right) \tag{10}$$

$$O = LN\left(\hat{O} + FFN\left(\hat{O}\right)\right) \tag{11}$$

where FFN is the feed-forward network, LN is the layer normalization. Both of these operations are consistent with those in vanilla Transformer (Vaswani et al., 2017).

3.4 Chinese representation with bilingual interaction

Given a Chinese text sequence $C = (c_1, c_2, ..., c_n)$ and an English text sequence $E = (e_1, e_2, ..., e_m)$, obtaining the Chinese representation of fused bilingual information requires two steps: inter-modal fusion and intra-modal fusion.

Inter-modal interaction: In order to learn a better representation of Chinese with the aid of information in the English text, we first employ ACT-Ss to get the English-Aware Chinese Representation. Similarly, to align each English word with its closely related Chinese characters, i.e., assigning high/low attention weights to its related/unrelated Chinese characters, we also use parameter independent ACT-S to gain the Chinese-Aware English Representation:

$$K^{(t)} = ACT\text{-}S\_cn(K^{(t-1)}, J^{(t-1)}) \tag{12}$$

$$J^{(t)} = ACT\text{-}S\_en(J^{(t-1)}, K^{(t-1)}) \tag{13}$$

where $t \in \{1, \cdots, N\}$, N denotes the number of interactions between bilingual modalities, $K^{(t)} = \left(k^{(t)}_1, k^{(t)}_2, ..., k^{(t)}_n\right)$, $J^{(t)} = \left(j^{(t)}_1, j^{(t)}_2, ..., j^{(t)}_m\right)$, $K^{(0)} = C$ and $J^{(0)} = E$.

Intra-modal interaction: In order to learn the dependencies between Chinese words, we use another ACT-S to obtain the Intra-Chinese Interaction Representation:

$$A = ACT\text{-}S\_in(C, J^{(N)}) \tag{14}$$

Then, we concatenate $A = (a_1, a_2, ..., a_n)$ and $J^{(N)}$ and input them into a fully connected layer to obtain the final hidden representations $H = (h_1, h_2, ..., h_n)$:

$$H = \left(a_1 \oplus j^{(N)}_1, ..., a_n \oplus j^{(N)}_n\right)W^{F} + b \tag{15}$$

where $\oplus$ denotes the concatenation operation, W^{F} and b are learnable parameters.

At last, we pass the final hidden representations H to a Conditional Random Field (CRF) (LAFFERTY, 2001) module.

4 Experiments

4.1 Experiment Settings

4.1.1 Datasets

We used four publicly available Chinese NER datasets, including Weibo NER (Peng and Dredze, 2015), Resume NER (Zhang and Yang, 2018), MSRA (Levow, 2006) and Ontonotes 4.0 (Weischedel et al., 2010). The corpus of MSRA and Ontonotes 4.0 comes from news, the corpus of Weibo comes from social media, and the corpus of Resume comes from the resume data in Sina Finance. The splitting methods and other pre-processing methods of datasets follow those in (Zhang and Yang, 2018; Li et al., 2020; Wu et al., 2021).

4.1.2 Text Translation

Neural machine translation (NMT) methods have achieved state-of-the-art performance for the translation of a wide range of language pairs(Vaswani et al., 2017). Therefore, automatic text translation with NMT is applicable. In this work, we employ Baidu Translation API[1] and Tencent Translation API[2] to automatically translate Chinese text into English, respectively, and both machine translation systems have achieved the highest BLUE score on the WMT Chinese-English task(Sun et al., 2019; Wang et al., 2021b) in recent years.

[1] https://fanyi-api.baidu.com/

[2] https://cloud.tencent.com/product/tmt

Models	Resources	Weibo			Resume			MSRA			Ontonotes 4.0		
		P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BERT-Tagger$^{\blacklozenge\dagger}$	-	-	-	68.20	-	-	95.53	-	-	94.95	-	-	80.14
LSTM-CRF[BERT]$^{\blacklozenge *}$	-	68.21	68.38	68.29	95.11	96.01	95.56	95.33	95.04	95.18	79.92	80.56	80.24
TENER[BERT]$^{\blacklozenge *}$	-	67.19	69.49	68.32	94.98	96.20	95.59	95.39	95.44	95.41	78.69	82.21	80.41
FLAT[BERT]$^{\blacklozenge\dagger}$	L	-	-	68.55	-	-	95.86	-	-	96.09	-	-	81.82
DyLex$^{\blacklozenge\dagger}$	L	-	-	71.12	-	-	95.99	-	-	96.49	-	-	81.48
MECT[BERT]$^{\blacklozenge\dagger}$	L+RC	-	-	70.43	-	-	95.98	-	-	96.24	-	-	82.57
Ours(N=2)$^{\blacklozenge *}$	T-T	**72.60**	**74.16**	**73.37**	96.29	**96.99**	**96.87**	**96.62**	96.82	96.72	83.95	83.77	83.86
Ours(N=2)$^{\blacklozenge *}$	T-B	72.57	73.95	73.25	**96.30**	96.91	96.60	96.59	**96.89**	**96.74**	**83.98**	**83.85**	**83.91**
BERT-Tagger$^{\blacktriangle\dagger}$	-	67.12	66.88	67.33	96.12	95.45	95.78	94.43	93.86	94.14	78.01	80.35	79.16
LSTM-CRF[BERT]$^{\blacktriangle *}$	-	67.63	67.18	67.40	95.84	95.61	95.72	94.46	93.89	94.17	78.92	79.56	79.24
TENER[BERT]$^{\blacktriangle *}$	-	66.69	68.21	67.44	95.05	96.63	95.83	94.45	94.19	94.32	78.99	79.70	79.34
LEBERT$^{\blacktriangle\dagger}$	L	-	-	70.75	-	-	96.08	-	-	95.70	-	-	82.08
PLTE[BERT]$^{\blacktriangle\dagger}$	L	72.00	66.67	69.23	96.16	96.75	96.45	94.91	94.15	94.53	79.62	81.82	80.60
SLex-LSTM[BERT]$^{\blacktriangle\dagger}$	L	70.94	67.02	70.50	96.08	96.13	96.11	95.75	95.10	95.42	83.41	82.21	82.81
Ours(N=2)$^{\blacktriangle *}$	T-T	**73.21**	**71.90**	**72.55**	96.21	**96.89**	96.55	96.35	95.91	96.13	83.24	**83.82**	83.53
Ours(N=2)$^{\blacktriangle *}$	T-B	73.15	71.84	72.48	**96.32**	96.87	**96.59**	**96.37**	**95.92**	**96.14**	**83.41**	83.71	**83.56**

Table 1: Main results. Bold marks the highest score. † marks results quoted directly from the original papers. ♦ marks results produced with BERT-wwm. ▲ marks results produced with BERT-base-Chinese. $*$ marks the results implemented in the fastNLP[3] framework. 'L' denotes using the lexicon resources. 'RC' denotes the radical information of Chinese. 'T-T' denotes using the bilingual information from the Tencent Translation API and 'T-B' denotes using the bilingual information from the Baidu Translation API.

4.1.3 Baseline Methods

To demonstrate the effectiveness of our proposed model, we compare it with several strong baseline models for Chinese NER: (1) **BERT-Tagger** (Devlin et al., 2019)(2) **LSTM-CRF[BERT]** (Huang et al., 2015) (3) **TENER[BERT]** (Yan et al., 2019). Besides, we also compare our method with the lexicon enhancement methods, which are the state-of-the-art methods for Chinese NER: (1) **LEBERT** (Liu et al., 2021) (2) **FLAT[BERT]** (Li et al., 2020) (3) **PLTE[BERT]** (Mengge et al., 2020)(4) **SoftLexicon-LSTM[BERT]** (Ma et al., 2020)(In this paper, we call Soft-Lexicon SLex for short.) (5)**DyLex** (Wang et al., 2021a) (6) **MECT[BERT]** (Wu et al., 2021).

4.1.4 Implement Details

The English word representations E are initialized with the cased BERT-base-English model pretrained by Devlin et al. (2019). In addition, to make the comparison with the results of the baseline models convincing, we use BERT-base-Chinese(Devlin et al., 2019) and BERT-wwm(Cui et al., 2021) to initialize the Chinese word representations C separately to get different experimental results for comparison and fine-tuned during training. All the neural models are implemented with PyTorch and fastNLP[3]. More implementation details are described in Appendix A.

4.2 Main Results

We compared our proposed method with the state of the art methods. The experimental results are reported in Table 1, which is divided into two blocks. The methods in the first block use the Chinese word representation from BERT-wwm. And the methods in the second block use the Chinese word representation from BERT-base-Chinese. Our model achieves a significant and consistent performance boost over current SOTA models on four Chinese NER datasets. From the results, we can observe that:

(1) In comparison with the methods without external resources (BERT-Tagger, LSTM-CRF[BERT] and TENER[BERT]), our model achieves a significant performance boost. Because our model makes use of the rich prompt information in the English text that help to determine the boundaries and types of entities. It demonstrates the significant effect of introducing bilingual information compared to just using the internal features of Chinese.

(2) Compared with the lexicon enhancement method using pre-trained BERT-wwm Chinese representation, our model has superior performance, i.e., +2.25 , +0.88, +0.25 ,+1.34 on Weibo, Resume, MSRA, Ontonotes4.0, respectively. When compared with baselines with BERT-base-Chinese, the performance of our model is still be competitive, i.e., +1.8, +0.14, +0.44, +0.75 on Weibo, Resume, MSRA, Ontonotes4.0, respectively. This verifies our claim that, compared with the exter-

[3] https://github.com/fastnlp/fastNLP

Models	Weibo	Resume	MSRA	OntoNotes
Ours	73.37	96.87	96.72	83.86
-GS	71.68	96.09	96.02	82.72
-RA	72.23	96.18	96.14	82.81
-APW	73.28	96.80	96.65	83.79
-KB	73.25	96.75	96.64	83.81

Table 2: An ablation study of the proposed model. F1 scores were evaluated on the test sets. 'GS' denotes the Global Sparse operation. 'RA' denotes the relative distance-aware attention. 'APW' denotes the average pooling operation used to obtain the word-level representations. 'KB' denotes the key bias term in the attention matrix.

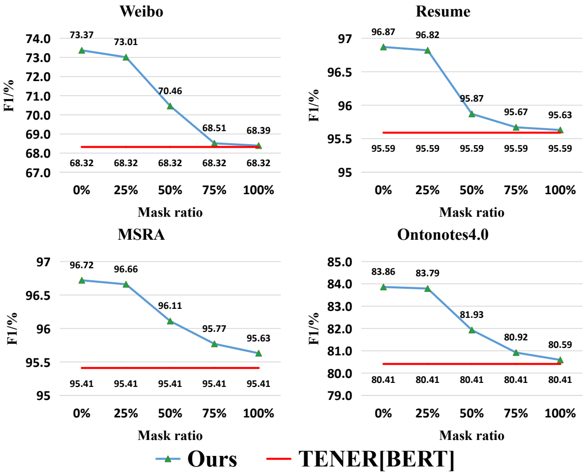


Figure 3: Impact of the machine translation quality. F1 scores were evaluated on the test sets.

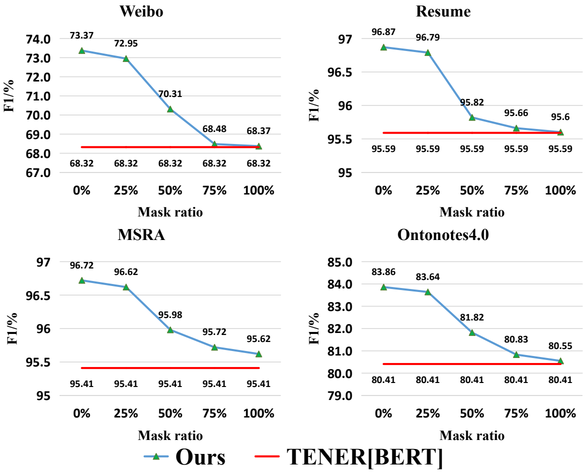


Figure 4: Impact of missing translated texts. F1 scores were evaluated on the test sets.

nal resources used by most Chinese NER models, incorporating the rich information in English is a promising way to improve the performance of Chinese NER tasks.

(3) All translated English texts used in this work are automatically generated from the publicly available machine translation API. Compared with the baseline models with external resources, our proposed bilingual enhanced approach improves performance and requires no artificially generated external knowledge. From another perspective, we believe that our proposed model transfers the knowledge in the neural machine translation model to enhance the Chinese NER model.

(4) Our proposed model could achieve state-of-the-art performance in both cases with English texts automatically translated by two different machine translation systems. This suggests that the performance improvement of our model is not dependent on a specific machine translation system.

(5) Even when the size of training set is small, such as Weibo NER, the performance improvement of our model over other baselines is still significant. This demonstrates that our proposed model is not data-hungry and has promising potential in low-resource NER scenarios.

5 Analysis and Discussion

5.1 Ablation Study

To study the contribution of the main components in our model, we conducted an ablation study on all four datasets. The results are reported in Table 2. And we can observe the following facts:

(1) To demonstrate the advantage of global sparse operation, we remove it from the model. The results show that, without global sparse operation, the performance of the model degrades severely, which indicates that there exist severe interference of irrelevant words during the process of bilingual interactions. The global sparse operation in ACT-S substantially improves the performance of our model, demonstrating its effectiveness while achieving higher interpretability for our method.

(2) The component of sensing relative distance in ACT-S has a significant positive impact on the performance of the proposed model, and the model without it shows a certain degree of performance degradation. This illustrates the validity of the relative distance-aware component in our model.

(3) We empirically found that the model performs better at word-level bilingual interactions than at subword-level.

(4) There are positive effects of introducing a key bias term into the cross-attention matrix.

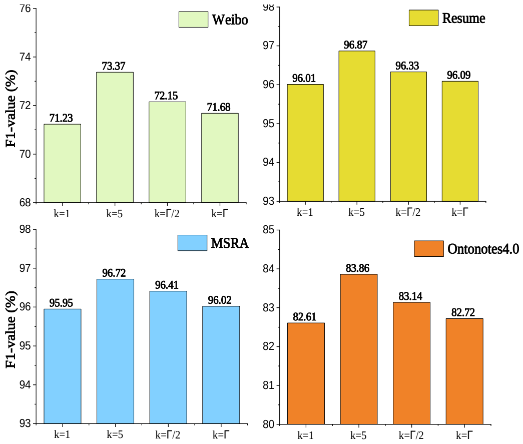


Figure 5: Influence of attention sparsity k. Γ is the length of the text sequence.

Method	NER results
Gold labels	[美国“亚特兰蒂斯号”航天飞机　ORG]上的宇航员... Astronauts on [the US space shuttle Atlantics　ORG]...
BERT-Tagger	[美国　LOC]“亚特兰蒂斯号”航天飞机上的宇航员...
MECT	[美国　LOC][“亚特兰蒂斯号”航天飞机　ORG]上的宇航员...
Translation-1	Astronauts on the US space shuttle Atlantics...
Ours-1	[美国“亚特兰蒂斯号”航天飞机　ORG]上的宇航员...
Translation-2	Astronauts on space shuttle Atlantics of US...
Ours-2	[美国“亚特兰蒂斯号”航天飞机　ORG]上的宇航员...

Table 3: Example-1 of the NER results.

5.2 Impact of machine translation quality & missing translated texts

To illustrate the robustness of our model, we set up separate experiments to investigate the effect of machine translation quality and missing translated text on our model. We randomly replace words in the automatically translated English text with the mask token [MASK] in a fixed proportion to simulate a reduction in the machine translation quality. Similarly, we randomly replace the entire automatically translated English text with a fixed-length(average length of samples in the training set) sequence of the mask token [MASK] at a certain rate for all samples. As shown in Figure 3 and Figure 4, our model can achieve excellent performance even in cases where the translated text has little noise or is missing. And when the translated text is almost full of noise or even 100% missing, our model still outperforms TENER [BERT], a strong baseline that does not use any external resources, which demonstrates the robustness of our model. In addition, it demonstrates that the performance improvement of our model originates not only from bilingual resources but is also related to the model itself.

5.3 Impact of attention sparsity k.

We also conducted experiments to verify the impact of the attention sparsity control factor k in ACT-Ss. The results reported on the four datasets are shown in Figure 5. From the results, we can see that the inappropriate sparsity of global attention significantly degrades the performance of the model. If the global cross-attention matrix is too sparse, such as in the case of $k = 1$, the ACT-S module cannot learn sufficiently about the dependencies between bilingual texts. At the other extreme, if the attention matrix takes too many interactions between bilingual words into account, the model will not achieve the best performance either. This indicates the sparse nature of lexical dependencies between bilingual texts. Furthermore, the experimental results suggest that it is applicable and interpretable to introduce sparse attention rather than full attention in bilingual interactions.

5.4 Case Study

Table 3 illustrates one typical example where our proposed bilingual enhancement model successfully tackles the dilemma of the complex structure of entity composition in the Chinese NER task. Most of the existing Chinese NER methods utilize only the internal features of the Chinese language, which makes it difficult to tag entities with complex composition structures correctly. When rich cues in English are leveraged, this problem can be alleviated. In addition, it can be seen from these two cases that the clues in the different English translations are all beneficial for the correct labeling of complex entities.

6 Conclusion

In this paper, we propose an Adapted Cross-Transformers with Global Sparse Attention (ACT-S) module to explore bilingual interaction information to improve the performance of the Chinese NER task. Several parameter independent ACT-Ss are employed in our work to capture the rich information in both English and Chinese. We evaluate the proposed model on four Chinese NER datasets and the experimental results illustrate that our method achieves significant and consistent improvement compared to other baselines. In the future, we will explore how to improve Chinese NER with features from languages other than English and extend our model to other sequence labelling tasks.

Acknowledgements

This research is supported by the Fundamental Research Funds for the Central Universities (No.DUT22ZD205). Finally, we would like to thank the anonymous reviewers for their valuable comments.

References

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *NAACL*, pages 52–62.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, 55(3):529–569.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based LSTM-CRF with radical-level features for chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, volume 10102 of *Lecture Notes in Computer Science*, pages 239–250. Springer.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *IJCAI*.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019b. A lexicon-based graph neural network for chinese ner. In *EMNLP*, pages 1040–1050.

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

JD LAFFERTY. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML-2001*.

Gina-Anne Levow. 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*, pages 108–117.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. In *ACL*.

Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473:37–53.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using BERT adapter. In *ACL*, pages 5847–5858.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuan-Jing Huang. 2020. Simplify the usage of lexicon in chinese ner. In *ACL*, pages 5951–5960.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. *Advances in Neural Information Processing Systems*, 32:2746–2757.

Xue Mengge, Bowen Yu, Tingwen Liu, Yue Zhang, Erli Meng, and Bin Wang. 2020. Porous lattice transformer encoder for chinese ner. In *COLING*, pages 3831–3841.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *EMNLP*, pages 548–554.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale chinese multimodal ner dataset with speech clues. In *ACL*, pages 2807–2818.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021a. Rpbert: A text-image relation propagation-based bert model for multimodal ner. In *AAAI*, pages 13860–13868.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 374–381.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021b. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *ACL*, pages 2065–2075.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Baojun Wang, Zhao Zhang, Kun Xu, Guang-Yuan Hao, Yuyang Zhang, Lifeng Shang, Linlin Li, Xiao Chen, Xin Jiang, and Qun Liu. 2021a. Dylex: Incorporating dynamic lexicons into bert for sequence labeling. In *emnlp*, pages 2679–2693.

Longyue Wang, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021b. Tencent translation system for the wmt21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 216–224.

Mengqiu Wang, Wanxiang Che, and Christopher Manning. 2013. Effective bilingual constraints for semi-supervised learning of named entity recognizers. In *AAAI*, volume 27.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, et al. 2010. Ontonotes release 4.0.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. 2021. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. In *ACL*, pages 1529–1539.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In *Int. J. Comput. Linguist.Chin. Lang. Process*, pages 29–48.

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

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *ACL*, pages 3342–3352.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *AAAI*, pages 14347–14355.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *ACL*, pages 1554–1564.

Ranran Zhen, Rui Wang, Guohong Fu, Chengguo Lv, and Meishan Zhang. 2021. Chinese opinion role labeling with corpus translation: A pivot study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10139–10149.