Translation as Cross-Domain Knowledge: Attention Augmentation for Unsupervised Cross-Domain Segmenting and Labeling Tasks

Ruixuan Luo, [2*] **Yi Zhang,** [1*] **Sishuo Chen,** [2] **Xu Sun** [1,2]
[1]MOE Key Laboratory of Computational Linguistic, School of EECS, Peking University
[2]Center for Data Science, Peking University
{luoruixuan97,zhangyi16,chensishuo,xusun}@pku.edu.cn

Abstract

The nature of no word delimiter or inflection indicating segment boundaries or word semantics increases the difficulty of Chinese text understanding, and also intensifies the demand for word-level semantic knowledge to accomplish the tagging goal in Chinese segmenting and labeling tasks. However, for unsupervised Chinese cross-domain segmenting and labeling tasks, the model trained on the source domain frequently suffers from the deficient word-level semantic knowledge of the target domain. To address this issue, we propose a novel paradigm based on **attention augmentation** to introduce crucial cross-domain knowledge via a translation system. The proposed paradigm enables the model attention to draw cross-domain knowledge indicated by the implicit word-level cross-lingual alignment between the input and its corresponding translation. Aside from the model requiring cross-lingual input, we also establish an off-the-shelf model which eludes the dependency on cross-lingual translations. Experiments demonstrate that our proposal significantly advances the state-of-the-art results of cross-domain Chinese segmenting and labeling tasks[1].

1 Introduction

As a language that has no word delimiter or inflection indicating segment boundaries or word semantics, Chinese increases the difficulty of segmenting and labeling tasks which need to correctly identify the segment boundaries from a sequence of characters and to thoroughly understand the word meanings. In this paper, we regard the knowledge about segment boundaries and detailed word meanings as the word-level semantic knowledge, and intend to promote the Chinese cross-domain segmenting and labeling tasks where the paucity of such knowledge in the target domain is usually a severe problem and frequently undermines the performance of a cross-domain model.

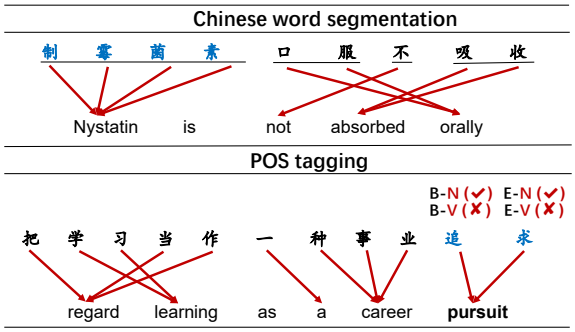


Figure 1: Two typical examples that cross-lingual contexts derived from machine translation help improve segmenting and labeling tasks. The arrows highlight the cross-lingual word that each Chinese character pays most attention to. Regarding the first case, the integrity of the out-of-domain word 制霉菌素 is indicated by its translated English word. In the second case, the translation of the ambiguous word 追求 is *pursuit* instead of *pursue*, which indicates the correct tag *noun*.

Existing work achieves cross-domain segmenting and labeling either by crafting domain-specific knowledge (Zhang et al., 2018; Liu et al., 2014), which is inflexible and cumbersome in adapting to different domains, or by learning on unannotated data of target domain (Ye et al., 2019; Jia et al., 2019) via language modeling, which cannot comprehend the detailed word-level semantics apart from grasping the general meaning of a sentence.

To address these issues, the key is to search for cross-domain knowledge that is easily accessible yet contains crucial word-level semantic knowledge. Motivated by the previous studies that segmenting and labeling tasks can benefit machine translation (Chang et al., 2008; Wang et al., 2014; Niehues and Cho, 2017; Zaremoodi and Haffari, 2019), in turn, we speculate that the machine-translated sentences would conceivably reveal some fundamental segmenting knowledge

*Equal contribution.
[1]Our code is available at https://github.com/lancopku/Attention-Augmentation

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1896–1906
November 7–11, 2021.

and help infer detailed word-level semantics.

The cross-domain knowledge inferred from translations can be illustrated by the two examples in Figure 1. For Chinese word segmentation, the word in blue from the target domain is originally excessively segmented. However, a translated version of the input sentence provides a strong indication of the integrity of this word. The part-of-speech tag of the word 追求 is originally ambiguous since it can be both a noun (as in the given context) and a modal verb. Nonetheless, it corresponds to distinct English words depending on the meaning it presents as a noun or as a verb, thus its translated counterpart actually hints at the correct label.

Motivated by the above observations, we propose a novel paradigm based on **attention augmentation** to introduce word-level cross-domain knowledge via cross-lingual translation. The proposed paradigm complements the input sentence with its cross-lingual translation and enables the model attention to draw word-level knowledge implicitly embodied in the alignment of the input sentence pair. It then incorporates cross-lingual masked language modeling to further strengthen the word-level alignment, evolving into the **masked attention augmentation**. The enhanced alignment, in turn, helps boost segmenting and labeling tasks. To make our proposal more practical, we use this model to tag the raw text in target domains to reap abundant synthetic data, which aims to elicit the originally implicit cross-domain knowledge implied by the word-level alignment, and then use the synthetic data to train an off-the-shelf model relying on no translation inputs. Experimental results on a series of cross-domain segmenting and labeling datasets demonstrate that our model substantially advances the state-of-the-art results.

Our contributions are highlighted as follows:

- We propose a novel paradigm of attention augmentation that addresses the deficiency of word-level semantic knowledge for Chinese cross-domain segmenting and labeling tasks via augmented translation.
- We leverage this paradigm to derive plentiful synthetic data and then train a new tagging model with the synthetic data, relieving the model from the dependency on translation and further improving the practicability.
- The proposed approach significantly advances the state-of-the-art results of cross-domain Chinese segmenting and labeling tasks without any human-annotated data.

2 Related Work

Some previous work improves domain adaption in a single task framework. Domain-specific lexicon (Zhang et al., 2018; Liu et al., 2014) is usually adopted for cross-domain tasks. However, high-quality dictionaries for target domains are not always available. Recent work (Ye et al., 2019; Ding et al., 2020) uses raw text in the target domain to train word embeddings or construct word collections (Liu et al., 2014; Zhang et al., 2018). To relieve the burdensome work in crafting domain-specific knowledge, some researches attempt to align different domains, including mapping entity label space (Daumé III, 2007), sharing hidden feature representations (Yang et al., 2017), aligning feature distributions (Ganin et al., 2016) and using adaptation layers (Lin and Lu, 2018). However, it is difficult to align different domains in an unsupervised way.

Increasing work turns to multi-task learning since related NLP tasks can boost each other in a joint-learning framework. Language modeling (LM) is a common auxiliary objective that has been shown to be beneficial for sequence tagging (Rei, 2017). A natural idea is to learn contextualized embeddings by masked language modeling on the text from target domain (Han and Eisenstein, 2019). Jia et al. (2019) consider unsupervised domain adaptation for Named Entity Recognition (NER) via cross-domain language modeling tasks. Zhao et al. (2018) propose to incorporate unlabeled data for Chinese Word Segmentation (CWS) by combining segmentation model with language models. In addition to language modeling, some work learns to jointly optimize syntactic parsing and semantic parsing objectives (Niehues and Cho, 2017; Zaremoodi and Haffari, 2019). Liu and Zhang (2012) propose to use character clustering and self-training to jointly train CWS and POS tagging task. Tian et al. (2020) use a two-way attention mechanism to incorporate both context feature and their corresponding auto-analyzed syntactic knowledge for each input character and trains CWS and POS tagging tasks jointly. Some extensions turn to a multilingual setting and they leverage NMT systems to help cross-lingual NER for low-resource languages (Jain et al., 2019) or jointly model bilingual

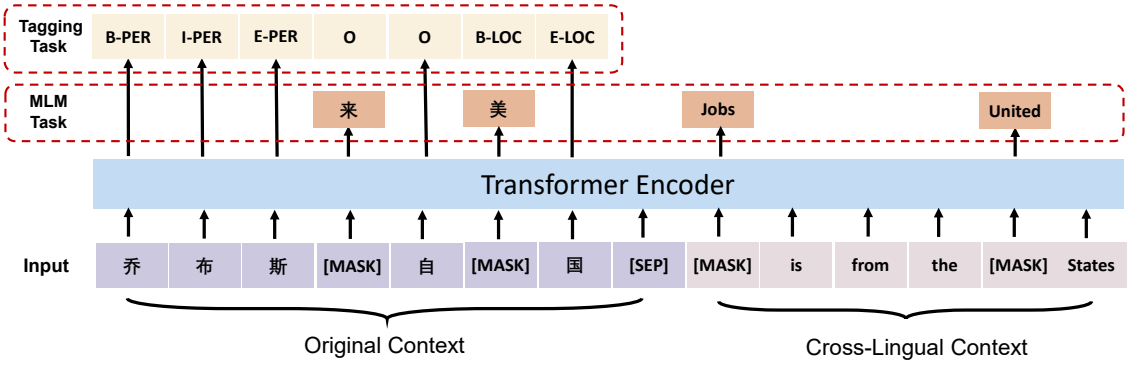


Figure 2: An illustration of the proposed approach $m\mathcal{A}^2$. The tagging task is only applied to the original context while the masked language modeling is performed on both contexts.

POS tagging (Snyder et al., 2008). However, existing methods are still inefficient in learning word-level semantic knowledge which is essential for segmenting and labeling tasks.

3 Approach

In this section, we introduce our approach in detail. We suppose that an annotated source domain dataset $\mathcal{S}$=$\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|\mathcal{S}|}$ consisting of input-label pairs is available, and our goal is to tag the unlabeled inputs $\mathcal{T}$=$\{\boldsymbol{u}_i\}_{i=1}^{|\mathcal{T}|}$ from the target domain. Each instance $\boldsymbol{x}_i$ in $\mathcal{S}$ represents an input sequence and its ground-truth label $\boldsymbol{y}_i$ is also a sequence. We implement our approach based on BERT. In addition, we consider the situation that the target domain shares the same label set as the source domain in this work.

3.1 Attention Augmentation

In light of the fact that segmenting and labeling tasks can benefit the machine translation task, we speculate that the translation process implicitly incorporates the segmenting process and comprehends the detailed word semantics. We then intend to exploit the word-level semantic knowledge embodied in the translation process, including the segmenting knowledge and the detailed word meanings. We enable the prevailing attention-based model to attend to both the input sentence and its translation counterpart to infer labels, resulting in an augmented attention pattern, and this paradigm is abbreviated to **A**ttention **A**ugmentation ($\mathcal{A}^2$) .

In the paradigm of attention augmentation, the original input sentence and its cross-lingual translation compose a translation pair. This pair is fed into a self-attention based segmenting or labeling model for further processing, which in our case is BERT. Owing to the self-attention mechanism, the model can not only attend to the original input but also its translated version to predict labels for the original input. For one thing, the implicit cross-lingual word-level alignment embodied in the translation pair indicates the knowledge about segment boundaries. For another, the original context and the cross-lingual context complement each other and help understand the detailed word meanings in a reciprocal manner if the monolingual context is insufficient to infer the accurate word semantics. Especially, these two kinds of word-level semantic knowledge are usually not covered by the original context in the source domain.

Concretely, given an input sequence $\boldsymbol{x}_i$ from source domain, we first fetch its translated version $\boldsymbol{t}_i$ with an available cross-lingual translation model. Then the original input $\boldsymbol{x}_i$ and its translation $\boldsymbol{t}_i$ are packed together into a single sequence $\tilde{\boldsymbol{x}}_i$ and are separated by a special `[SEP]` token. With $\tilde{\boldsymbol{x}}_i$ as input, the pre-trained model encodes $\tilde{\boldsymbol{x}}_i$ by a number of blocks consisting of self-attention mechanism and outputs a predicted label sequence $\hat{\boldsymbol{y}}_i$. The model is then updated with the cross-entropy loss $\mathcal{L}_{cls}$ against the ground-truth labels $\boldsymbol{y}_i$ of the input sequence $\boldsymbol{x}_i$:

$$\boldsymbol{h}_i = \text{BERT}(\tilde{\boldsymbol{x}}_i) \quad (1)$$

$$\hat{\boldsymbol{y}}_i = \text{Softmax}(\boldsymbol{W}_{cls}\boldsymbol{h}_i + \boldsymbol{b}_{cls}) \quad (2)$$

$$\mathcal{L}_{cls} = \text{CrossEntropy}(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) \quad (3)$$

where $\boldsymbol{h}_i$ denotes the encoded representations of the input $\tilde{\boldsymbol{x}}_i$, $\boldsymbol{W}_{cls}$ and $\boldsymbol{b}_{cls}$ are learnable parameters. Since the ground-truth labels of the translated sentence $\boldsymbol{t}_i$ is unavailable, the outputs of the positions corresponding to the translated sentence will be ignored during training.

3.2 Masked Attention Augmentation

In order to better align the representations of the source language and the translated language, we incorporate cross-lingual masked language modeling (Conneau and Lample, 2019) into our work and develop the **m**asked **A**ttention **A**ugmentation approach ($m\mathcal{A}^2$). We randomly mask some tokens in the concatenated sequence $\tilde{\boldsymbol{x}}_i$ and encourage the model to reconstruct the masked tokens. Since the tokens in both languages can be masked, the model can attend to the tokens in the other language to predict the masked tokens for the current language, which enhances the alignment between the source language and the translated language. We simultaneously optimize the cross-lingual language modeling and the preceding tagging objective. An overview of $m\mathcal{A}^2$ is illustrated in Figure 2.

To be precise, given a translation pair, 15% of the tokens in both the original input sentence and the translated sentence will be chosen at random for prediction. Each chosen token is replaced with a `[MASK]` token for 80% of the time, a random token for 10% of the time, and unchanged for the rest 10% of the time, resulting in the corrupted input sequence $\boldsymbol{x}_i^c$. $\boldsymbol{x}_i^c$ is then fed into a tagging model for encoding. The hidden states of the last layer will be used to reconstruct the masked tokens apart from predicting labels. To this end, we add an additional transformation layer to predict the masked tokens in $\boldsymbol{x}_i^c$. The learning process of the masked language modeling objective $\mathcal{L}_{mlm}$ is formulated as follows:

$$\boldsymbol{x}_i^c = \text{Mask}(\tilde{\boldsymbol{x}}_i) \tag{4}$$

$$\boldsymbol{h}_i = \text{BERT}(\boldsymbol{x}_i^c) \tag{5}$$

$$\hat{\boldsymbol{z}}_i = \text{Softmax}(\boldsymbol{W}_{mlm}\boldsymbol{h}_i + \boldsymbol{b}_{mlm}) \tag{6}$$

$$\mathcal{L}_{mlm} = \text{CrossEntropy}(\hat{\boldsymbol{z}}_i, \tilde{\boldsymbol{x}}_i) \tag{7}$$

where $\boldsymbol{W}_{mlm}$ and $\boldsymbol{b}_{mlm}$ are learnable parameters in the additional transformation layer, $\hat{\boldsymbol{z}}_i$ is the predicted sequence, Mask$(\cdot)$ denotes the function that masks some percentage of the input tokens at random. The overall loss $\mathcal{L}$ of $m\mathcal{A}^2$ comprises the classification loss for the tagging purpose and the masked language modeling loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{mlm} \tag{8}$$

Note that the process of cross-lingual language modeling has no requirements for segmenting or labeling tags of the input and thus the unannotated raw text $\mathcal{T}$ of the target domain can be engaged just for cross-lingual language modeling, which means the paradigm of masked attention augmentation is able to exploit both the annotated data in the source domain and the unannotated data in the target domain.

3.3 Attention Augmentation Rebooted

The proposed $\mathcal{A}^2$ and $m\mathcal{A}^2$ take translation pair as input for training. As a result, they require the translation process during the inference stage. To make the models more practical, we then intend to enable the models to be deployed in the scenario where the translation of input is not available.

To achieve this, we take advantage of the well-trained model with masked attention augmentation to annotate the unlabeled inputs $\mathcal{T}$ of the target domain, establishing synthetic training data for the target domain. This annotation process is expected to elicit the word-level cross-domain knowledge which is originally implicit in the translation pairs. We then use the synthetic training data that imbibes the word-level semantic knowledge to train a new tagging model. This new model, termed as $m\mathcal{A}^2$-reboot for short, is then unchained from the translation input and can be easily deployed for regular inference. In addition, it can be initialized with the fine-tuned language model on the source domain to be endowed with the basic knowledge of tagging, as we do in practice.

4 Experiments

We conduct experiments on three Chinese cross-domain segmenting and labeling tasks, namely part-of-speech tagging (POS tagging), Chinese Word Segmentation (CWS) and Named Entity Recognition (NER). As most studies do, we conduct cross-domain experiments on datasets from different domains.

4.1 Tasks

POS tagging Following Tian et al. (2020), we conduct experiments on eight genres of CTB9 (Xue et al., 2005): broadcast conversation (BC), broadcast news (BN), conversational speech(CS), discussion forums (DF), magazine articles (MZ), newswire (NW), SMS / chat messages (SC), and weblog (WB). Each genre is regarded as the target domain data for evaluation and the combined data of the other genres serves as the source domain data.

Method	Target Data								Avg.
	BC	BN	CS	DF	MZ	NW	SC	WB	
BiLSTM-CRF	88.98	92.36	84.82	88.70	86.60	91.44	86.91	87.41	88.40
XLM	92.02	92.21	86.12	89.47	88.36	91.98	85.91	88.02	89.26
BERT-base-Chinese	92.08	93.73	89.18	92.02	91.97	94.44	91.13	89.59	91.77
BERT-base-multilingual	90.13	94.81	87.20	90.94	91.17	94.18	89.15	89.33	90.86
TwASP (Tian et al., 2020)	92.34	94.02	89.46	92.44	92.17	94.64	91.77	89.85	92.09
$\mathcal{A}^2$	92.45	95.82	92.42	93.08	92.51	95.02	92.53	90.91	93.09 (+1.00)
$m\mathcal{A}^2$	**92.63**	**95.99**	**92.55**	**93.60**	**92.70**	**95.05**	**93.36**	**92.11**	**93.50 (+1.41)**
$m\mathcal{A}^2$-reboot	92.13	95.83	92.17	93.26	92.35	94.59	92.88	91.09	93.04 (+0.95)

Table 1: Comparison with the state-of-the-art results for unsupervised cross-domain POS tagging.

Method	Target Data		Avg.
	DM	PT	
Partial CRF (Ding et al., 2020)	82.8	85.0	83.9
CWS-DICT (Zhang et al., 2018)	81.2	85.9	83.6
WEB-CWS (Ye et al., 2019)	82.2	85.1	83.7
DAAT (Ding et al., 2020)	85.0	89.6	87.3
XLM	85.7	90.1	87.9
BERT-base-Chinese	85.4	91.8	88.6
BERT-base-multilingual	85.3	91.6	88.5
$\mathcal{A}^2$	87.8	92.4	90.1 (+1.5)
$m\mathcal{A}^2$	**89.7**	93.3	91.5 (+2.9)
$m\mathcal{A}^2$-reboot	89.6	**93.5**	**91.6 (+3.0)**

Table 2: Comparison with the state-of-the-art results for unsupervised cross-domain CWS.

Method	Target Data		Avg.
	MSRA	PFR	
BiLSTM-CRF	75.22	76.65	75.94
XLM	78.29	77.84	78.07
BERT-base-Chinese	80.20	80.58	80.39
BERT-base-multilingual	79.87	80.48	80.18
$\mathcal{A}^2$	80.67	80.96	80.82 (+0.43)
$m\mathcal{A}^2$	**81.19**	**82.50**	**81.85 (+1.46)**
$m\mathcal{A}^2$-reboot	80.65	81.97	81.31 (+0.92)

Table 3: Comparison with the state-of-the-art results for unsupervised cross-domain NER.

Chinese Word Segmentation Following Ye et al. (2019) and Ding et al. (2020), we regard the News dataset (Qiu and Shi, 2015) as the source domain data, and the dermatology (DM) and patent (PT) datasets (Qiu and Shi, 2015) derived from dermatology domain and patent domain as the target domain data.

Named Entity Recognition We employ MSRA (Levow, 2006) dataset and the People's Daily (PFR) dataset. Each dataset is used for evaluation with the model trained on the other dataset.

Since unannotated data of the target domain can be engaged just for cross-lingual language modeling, for these tasks, we exclude the labels of the available training set from the target domain data and treat it as the unannotated data of the target domain for cross-lingual language modeling.

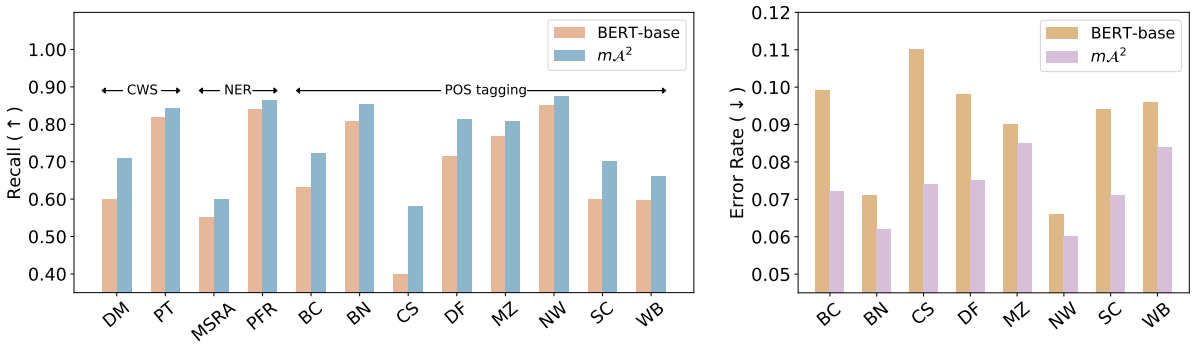
4.2 Experimental Settings

The implementation of our approach is based on BERT-base-multilingual since it needs to process the cross-lingual context. In our work, the cross-lingual context is provided by translating Chinese into English and we employ the commercial translation engine, i.e., Google Translation, to obtain high-quality translations. The learning rate is set to 2×10^{-5} for CWS and NER tasks, and is 5×10^{-5} for POS tagging. We adopt the Adam optimizer (Kingma and Ba, 2015) and train all the tasks for 20 epochs. Other hyper-parameters follow the default settings in BERT-base-multilingual. For $m\mathcal{A}^2$, half of the batches are from the source domain and used for both sequence tagging and masked language modeling while the other half batches are from the target domain and only used for the masked language modeling task.

We report F1-score for these three tasks. Regarding Chinese POS-tagging task, it jointly considers word segmentation and POS tagging for evaluation as the way in NER. We compare our proposal with some general baselines, i.e., BERT-base-Chinese, BERT-base-multilingual and XLM (Conneau and Lample, 2019) which incorporates multilingual pre-training. Besides, we include some state-of-the-art methods on target domain data for comparison.

4.3 Overall Performance

The performance of POS tagging, CWS and NER are presented in Table 1, 2 and 3 respectively. As shown, our proposal substantially advances the state-of-the-art results on all three tasks. For POS



(a) Evidence to suggest our model brings knowledge of segment boundaries. The boost in recall score of out-of-domain words verifies that our approach helps identify the integrity of out-of-domain words.

(b) The error rate reduction of ambiguous words (with multiple possible labels) verifies that our model helps understand the detailed word meanings.

Figure 3: Quantitative results for analyzing the two advantages of our approach.

tagging, our approach outperforms previous state-of-the-art model on a total of 8 datasets. For CWS, our proposal surpasses the previous best-performing BERT-base-Chinese, achieving 89.7 (+4.3 improvement) on the DM dataset and 93.5 (+1.7 improvement) on the PT dataset. Consistent improvement is also observed in NER task. We notice that our best result outperforms BERT-base-Chinese by 1.46 improvement on average.

BERT-base-Chinese, BERT-base-multilingual, XLM and our approach are general approaches that can be applied to both segmenting and labeling tasks, and the last three are multilingual approaches. We notice that these models outperform the conventional state-of-the-art models on most datasets. We conjecture that the large-scale pre-training process involved in these approaches helps model learn better contextualized word embeddings of some specific domains from heterogeneous raw text, resulting in a performance boost for cross-domain segmenting and labeling tasks. However, what sets our approach apart is that the multilingual context in our approach provides advantageous alignments to help learn segment boundaries and detailed word meanings. As a result, our approach brings further significant improvement compared to other models.

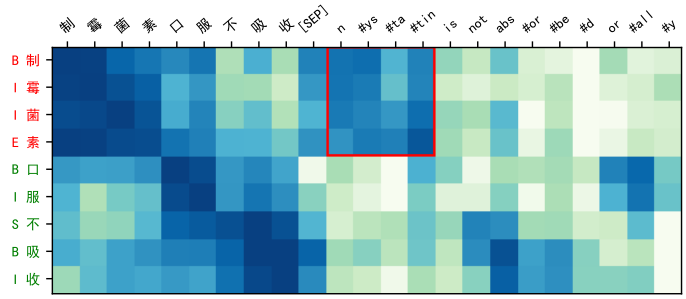
In addition to the general methods, we also compare the proposal with some state-of-the-art methods for different tasks. For POS tagging, TwASP provides additional auto-analyzed syntactic features and alleviates the deficiency of word-level semantic knowledge to some extent. However, it still suffers from the problems that monolingual context is inherently hard to tackle. For CWS, DAAT struggles to design intricate strategies to build annotated datasets of target domain. Despite the inefficient process, a major downside is that such an annotation process can only identify limited out-of-domain terms with high reliability. We also implement some competitive baseline methods on NER task. Compared to these models, our cross-lingual context provides hints at the word-level semantic knowledge required by the target task, resulting in better performance.

4.4 Analysis

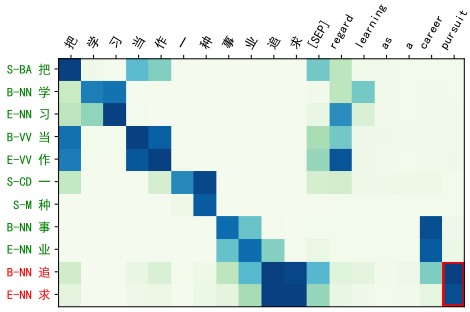
4.4.1 Quantitative Analysis

Analysis reveals the superiority of our proposal is two-fold: (1) the word-level cross-lingual alignment discloses the information of segment boundaries; (2) the cross-lingual context helps the model comprehend the detailed word meanings.

For the first advantage, the word-level alignment embodied in the translation pair indicates the integrity of a segment, especially for the out-of-domain words which are not covered by the source domain. Since all three tasks for Chinese necessitate the basic segmentation process, we calculate the recall score of the out-of-domain words in all datasets to analyze this benefit. We compare the strong baseline BERT-base-multilingual and the proposed $m\mathcal{A}$ which is based on BERT-base-multilingual to directly show the effectiveness. Results are reported in Figure 3. Taking the DM dataset from cross-domain CWS as an example, BERT-base-multilingual achieves a recall score of 59.5% for out-of-domain words while our proposal advances the score to 70.1%. Consistent improvement is observed in other datasets, which suggests the cross-lingual word alignment helps identify the integrity of segment, especially for the out-of-domain words in cross-domain tasks.



(a) A case where the cross-lingual word-level alignment indicates the knowledge about segment boundaries of the out-of-domain word in red.



(b) A case where the cross-lingual translation helps comprehend the detailed word meaning. The English word implies the correct tag of the word in red, enabling disambiguation.

Figure 4: Qualitative results for analyzing the two advantages of our approach $m\mathcal{A}^2$.

Regarding the second advantage, the detailed word meanings can be comprehended sufficiently through combining the cues revealed by multilingual contexts. To be specific, the cross-lingual alignment helps disambiguation since the ambiguity of a word or segment can also be reduced if it is expressed in two languages. For the POS tagging task, the accurate understandings of the detailed word meanings are crucial, especially for Chinese where no inflection could indicate the POS tags and ambiguity is common. We regard the words with multiple possible labels as ambiguous words and report the error rate of these words in Figure 3(b). As shown, an error rate reduction is obtained when our approach is applied. Taking the BC dataset as an instance, there are initially 9,470 ambiguous words that are incorrectly labeled by the baseline model. However, after introducing the cross-lingual context, this number is reduced by 27.98%, which confirms the effectiveness of our approach in comprehending detailed word meanings.

4.4.2 Qualitative Analysis

In addition to the preceding quantitative analysis, we instantiate the two aforementioned advantages and give an analysis from a qualitative perspective. Figure 4 shows the augmented attention heatmaps of two cases regarding the two advantages. Note that the words in red are incorrectly tagged by BERT-base-multilingual but are correctly recognized by our proposed $m\mathcal{A}^2$.

Figure 4(a) demonstrates a case where the cross-lingual alignment suggests the segmenting knowledge. Concretely, the out-of-domain word in red attends to its corresponding English translation which strongly indicates its integrity as a whole word, and further helps the model to correctly infer the segmentation labels.

Concerning the second advantage, Figure 4(b) shows a case where cross-lingual alignment helps the model comprehend the detailed word meanings via disambiguation. BERT-base-multilingual fails to predict the word 追求 as a noun since it also occurs frequently as a verb in Chinese natural text. However, the word 追求 corresponds to different English words according to its meaning as a noun or as a verb, and the attention heatmap reveals a strong alignment to the English word *pursue* which represents what 追求 means as a noun. Given this clue, our approach correctly tags the word as a noun, confirming that cross-lingual context promotes the understanding of detailed word semantics and enables disambiguation.

4.5 Effect of Key Components

4.5.1 Effect of Cross-lingual Language Modeling

As the cross-lingual language modeling is used to enhance the alignment embodied in the input translation pairs, we conduct ablation study to verify its effectiveness. The models with and without cross-language modeling are exactly $m\mathcal{A}^2$ and $\mathcal{A}^2$. The comparison results on POS tagging, CWS and NER tasks are shown at the bottom of Table 1, Table 2 and Table 3, respectively. We observe a significant performance boost with cross-language modeling involved on all the datasets, which verifies the effectiveness of the cross-lingual language modeling in our approach. During training, the model can either attend to the original context or the cross-lingual context to predict the masked tokens, which drives the model to grasp the alignment between

two contexts. More explicit and accurate alignment strengthens the foregoing two advantages of our approach and results in better performance.

4.5.2 Effect of Translation System

The machine translation system contributes cross-domain knowledge and thus plays an important role in our approach. Here we explore the effect of the translation system in our approach. As regards the effect of the languages to translate into, please refer to the Appendix. Generally, translations of input can be obtained by standard machine translation packages. Besides the Google Translation system we used, an alternative way is to train a translation model from scratch with existing datasets. Here we use a customized model, i.e., DynamicConv (Wu et al., 2019) ZH-EN translation model trained on WMT17 (Bojar et al., 2017), to conduct comparative experiments on CWS.

The results of using different translation models are shown in Table 4. We observe that our customized translation model is slightly inferior to the commercial translation engine. This is expected because a better translation system provides more accurate indication of word-level cross-domain knowledge. Nevertheless, both translation models outperform the baseline model by a large margin and the gap between themselves is negligible. We speculate that the existing general training set for a machine translation model has covered most of the knowledge that the commercial translation engine can provide. In other words, it suggests that our approach can be easily deployed with existing datasets and customized models.

Method	Target Data	
	DM	PT
BERT-base-multilingual	85.3	91.6
$m\mathcal{A}^2$ (DynamicConv)	89.5 (+4.2)	92.8 (+1.2)
$m\mathcal{A}^2$ (Google)	89.7 (+4.4)	93.3 (+1.7)
$m\mathcal{A}^2$-reboot (DynamicConv)	89.5 (+4.2)	93.3 (+1.7)
$m\mathcal{A}^2$-reboot (Google)	89.6 (+4.3)	93.5 (+1.9)

Table 4: Effect of different translation systems on CWS. The negligible performance gaps indicate that our approach does not rely heavily on the translation system and similar performance can be achieved with customized models.

4.6 Discussion of Future Work

4.6.1 Extending to in-domain Tasks

Previous analysis points out that the proposal supplements information of both segment boundaries and detailed word meanings. Such information is also required in in-domain tasks if the source domain data itself is insufficient to tackle the segmenting and labeling problems. In the future, we intend to extend the approach to in-domain tasks. Here we conduct experiments on in-domain NER task and take it as an appraisal of the extension. We employ the same datasets used in the cross-domain NER setting but train and test the model with consistent domain.

Experiments show some encouraging results. To be specific, our approach outperforms the best-performing model and achieves 94.39 (+1.30 improvement) and 96.29 (+0.56 improvement) F1-score on MSRA and PFR datasets respectively. Detailed results are reported in Appendix. The explanation of the improvement accords with our earlier analysis. We also notice that the average improvement gap narrows on in-domain datasets, as is expected, because the deficiency of word-level semantic knowledge is not as severe as in cross-domain datasets and our proposal is mainly oriented to cross-domain tasks. Nevertheless, the proposal performs well in both settings, which indicates that our proposal is also a promising approach for in-domain segmenting and labeling tasks.

4.6.2 Extending to other Languages

Although Chinese text processing shows a strong need for the knowledge of segment boundaries and detailed word meanings, such knowledge is also required in tagging tasks of other languages. Another potential application of our approach is the tagging tasks of other languages. Here we choose English, the commonly-used language, as the task language, and conduct experiments on English NER task for a trial. Following Jia et al. (2019), we take CoNLL-2003 dataset as our source domain data and CBS SciTech News (CBS News) representing science and technology domain as the target domain data.

Experiments show that $m\mathcal{A}^2$ and $m\mathcal{A}^2$-reboot achieves 75.31 and 76.04 F1-score respectively while the previous state-of-the-art result achieved by Jia et al. (2019) is 73.59. More experimental results are reported in Appendix. Since the cross-lingual context can also provide indication of entity boundaries in English and help disambiguation of word meanings to some extent, we observe improvement on English NER task too. This result suggests that extending our approach to the tasks of other languages is also a potential direction.

5 Conclusions

We propose a novel paradigm of attention augmentation that supplements cross-domain word-level knowledge via machine translation for Chinese cross-domain segmenting and labeling tasks. We also construct a model unchained from the dependency on translation. The proposed approach substantially advances the state-of-the-art results in Chinese cross-domain segmenting and labeling tasks without any human-annotated data, demonstrating the effectiveness of our proposal.

Acknowledgements

We thank all the anonymous reviewers for their constructive comments and Xuancheng Ren for the helpful discussion in preparing the manuscript. Xu Sun is the corresponding author of this paper.

References

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 169–214. Association for Computational Linguistics.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain chinese word segmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6662–6671. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4237–4247. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. *CoRR*, abs/1909.05356.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22-23, 2006*, pages 108–117. Association for Computational Linguistics.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2012–2022. Association for Computational Linguistics.

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and pos-tagging. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Posters, 8-15 December 2012, Mumbai, India*, pages 745–754. Indian Institute of Technology Bombay.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*

2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 864–874. ACL.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020. Coach: A coarse-to-fine approach for cross-domain slot filling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 19–25, Online. Association for Computational Linguistics.

Jan Niehues and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. *CoRR*, abs/1708.00993.

HW Likun Qiu and Linlin Shi. 2015. Construction of multi-domain chinese dependency treebanks and analysis of influencing factors on dependency parsing. *Journal of Chinese Information Processing*, 29(5):69.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. *CoRR*, abs/1704.07156.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for POS tagging. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1041–1050. ACL.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online. Association for Computational Linguistics.

Xiaolin Wang, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2014. Refining word segmentation using a manually aligned corpus for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1654–1664, Doha, Qatar. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11:207–238.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yuxiao Ye, Weikang Li, Yue Zhang, Likun Qiu, and Jian Sun. 2019. Improving cross-domain chinese word segmentation with word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2726–2735. Association for Computational Linguistics.

Poorya Zaremoodi and Gholamreza Haffari. 2019. Adaptively scheduled multitask learning: The case of low-resource neural machine translation. In *NGT@EMNLP-IJCNLP*.

Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for chinese word segmentation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5682–5689. AAAI Press.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4602–4608. ijcai.org.

A Effect of the Translation Language

Method	Target Data		Avg.
	DM	PT	
BERT-base-Chinese	85.4	91.8	88.6
BERT-base-multilingual	85.3	91.6	88.5
$m\mathcal{A}^2$ (English)	**89.7**	93.3	91.5 (+2.9)
$m\mathcal{A}^2$-reboot (English)	89.6	**93.5**	**91.6 (+3.0)**
$m\mathcal{A}^2$ (French)	89.4	92.3	90.9 (+2.3)
$m\mathcal{A}^2$-reboot (French)	89.4	92.5	91.0 (+2.4)
$m\mathcal{A}^2$ (Japanese)	88.9	92.1	90.5 (+1.9)
$m\mathcal{A}^2$-reboot (Japanese)	89.0	91.9	90.5 (+1.9)

Table 5: Performance with different translation languages for unsupervised cross-domain Chinese word segmentation (CWS).

Taking Chinese word segmentation as an instance, we employ three languages, i.e. English, French and Japanese, as target languages to explore

Method	Target Data		Avg.
	MSRA	PFR	
BiLSTM-CRF	88.28	92.93	90.61
XLM	87.57	93.00	90.29
BERT-base-Chinese	93.09	95.73	94.41
BERT-base-multilingual	92.09	94.81	93.45
$\mathcal{A}^2$	93.41	95.74	94.58 (+0.17)
$m\mathcal{A}^2$	**94.39**	**96.29**	**95.34** (**+0.93**)

Table 6: Comparison result on in-domain results for Chinese NER task.

Method	Target Data
	CBS News
BiLSTM-CRF	61.77
Coach+TR (Liu et al., 2020)	64.54
DANN (Ganin et al., 2016)	69.22
cross-domain LM (Jia et al., 2019)	73.59
BERT-base-multilingual	73.79
$\mathcal{A}^2$	74.32 (+0.53)
$m\mathcal{A}^2$	75.31 (+1.52)
$m\mathcal{A}^2$-reboot	**76.04** (**+2.25**)

Table 7: Comparison with the state-of-the-art results for English cross-domain NER.

the effect of the language to translate into. The results of different target languages are shown in Table 5. First, results with all three languages outperform the baseline model by a large margin, which verifies that introducing cross-lingual context is beneficial for our tagging task. Second, we observe a small performance difference across the three languages. English and French achieve comparable results and Japanese performs slightly inferior to the other languages. We speculate that Japanese is also a kind of language requiring segmentation process and thus the segment boundary information provided by Chinese-Japanese alignment is not as explicit as that of the other two languages. Nevertheless, the proposed approach proves effective with all three languages in general.

B Evaluation on in-domain Chinese NER

Table 6 shows the comparison result on in-domain datasets for Chinese NER task. As we can see from Table 6, our approach outperforms the competitive baseline models on both datasets with the in-domain setting. The results suggest that our proposal is also a promising approach for in-domain tasks.

C Evaluation on Cross-domain English NER

As extending the proposal to more languages is also a potential application direction, we conduct experiments on English cross-domain NER for a trial. Chinese translation of the English input serves as the cross-lingual context. Table 7 shows the comparison result of our approach and the previous state-of-the-art methods on these datasets. As shown, our approach obtains significant improvement compared to the previous methods. English NER task also necessitates the knowledge of segment boundaries as well as detailed word meanings. We assume that the cross-lingual context also helps supplement such knowledge for English to some extent. Therefore, the proposal can benefit this task. This result also reveals that extending the proposal to tasks of more languages is a potential direction.