Joint Chinese Word Segmentation and Part-of-speech Tagging via Multi-channel Attention of Character N-grams

Yuanhe Tian[♥], **Yan Song**[♠♡†], **Fei Xia**[♥]
[♥]University of Washington [♠]The Chinese University of Hong Kong (Shenzhen)
[♡]Shenzhen Research Institute of Big Data
[♥]{yhtian, fxia}@uw.edu [♠]songyan@cuhk.edu.cn

Abstract

Chinese word segmentation (CWS) and part-of-speech (POS) tagging are two fundamental tasks for Chinese language processing. Previous studies have demonstrated that jointly performing them can be an effective one-step solution to both tasks and this joint task can benefit from a good modeling of contextual features such as n-grams. However, their work on modeling such contextual features is limited to concatenating the features or their embeddings directly with the input embeddings without distinguishing whether the contextual features are important for the joint task in the specific context. Therefore, their models for the joint task could be misled by unimportant contextual information. In this paper, we propose a character-based neural model for the joint task enhanced by multi-channel attention of n-grams. In the attention module, n-gram features are categorized into different groups according to several criteria, and n-grams in each group are weighted and distinguished according to their importance for the joint task in the specific context. To categorize n-grams, we try two criteria in this study, i.e., n-gram frequency and length, so that n-grams having different capabilities of carrying contextual information are discriminatively learned by our proposed attention module. Experimental results on five benchmark datasets for CWS and POS tagging demonstrate that our approach outperforms strong baseline models and achieves state-of-the-art performance on all five datasets.[1]

1 Introduction

Chinese word segmentation (CWS) and part-of-speech (POS) tagging are two fundamental tasks in Chinese natural language processing (NLP). Although they can be treated as two separate tasks in a sequential order, it has been demonstrated by previous studies that processing them jointly in a unified sequence labeling framework could be more effective, where CWS and POS tags are predicted in a single step (Ng and Low, 2004; Jiang et al., 2008; Wang et al., 2011; Sun, 2011; Zeng et al., 2013; Zheng et al., 2013; Zhang et al., 2014; Kurita et al., 2017; Shao et al., 2017; Zhang et al., 2018). In doing so, existing studies mainly focused on incorporating contextual information (e.g., n-grams) as features into their joint taggers, which had been widely used as an effective way to improve model performance especially before neural models were widely used. Although neural models are powerful in modeling long text sequences, external features from larger granular texts are still demonstrated to be useful in existing neural models (Zheng et al., 2013; Kurita et al., 2017; Shao et al., 2017; Zhang et al., 2018). In these models, contextual features are leveraged by directly concatenating their embeddings with the character embeddings in the embedding layer, where all contextual features are treated equally without distinguishing their importance to the joint tagging process in the specific context.

However, this concatenation approach to incorporating contextual features into a joint tagger fails to consider that different contextual features could have different contributions to the joint task in a specific context, especially when there are ambiguities in the input sentence. For example, in an example sentence

[†]Corresponding author.

[1]Our code and models are available at https://github.com/cuhksz-nlp/McASP.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 28th International Conference on Computational Linguistics, pages 2073–2084
Barcelona, Spain (Online), December 8-13, 2020

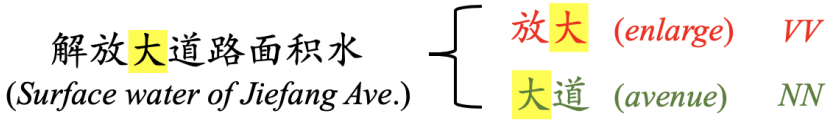


Figure 1: An illustration of the different roles of different n-gram features for the character "大" (highlighted in yellow) in an example sentence. In this case, "放大" (*enlarge*) and "大道" (*avenue*) are two n-gram features associated with "大", where the former one (in red color) suggests incorrect joint CWS and POS labels while the latter one (in green color) suggests the correct labels.

in Figure 1, there are two n-gram features associated with the character "大" (highlighted in yellow), i.e., "放大" (*enlarge*) and "大道" (*avenue*). In this case, both n-grams are common Chinese words; the former (in red color) might suggest incorrect CWS and POS tagging results, while the latter (in green color) suggests correct results.[2] If a model treats both n-gram features equally, it could be misled by the former (i.e., "放大" (*enlarge*) in red color). Therefore, it is important to distinguish the contributions of different n-grams to the joint task in a specific context; extra efforts are needed to effectively and smartly leverage such n-grams. In addition, considering that n-grams with different properties could also contribute differently for the joint task, it can be helpful to categorize the n-gram features into groups according to these properties and then model them separately.

In this paper, we propose a neural character-based joint CWS and POS tagger with multi-channel attention (MCATT) of character n-grams to improve the joint task. Specifically, to tag each character in an input sentence, the proposed MCATT first extracts the n-grams associated with the character from a pre-constructed lexicon and next categorizes such n-grams according to a specific metric (i.e., frequency or length). Then, we feed all n-grams within the same category into each channel, where those n-grams are compared and weighted according to their contribution to the joint label prediction in a specific context. Afterwards, the attentions from different channels are combined to help with the tagging process for each corresponding character. Compared to normal attention, where all associated n-grams are compared and weighted together without categorization, multi-channels provide an alternative approach to discriminatively leverage n-grams with different properties. Therefore, the weights for n-grams with similar properties are computed in their own channel rather than computed globally with all other n-grams. This multi-channel mechanism could be helpful to leverage the infrequent yet important n-grams, because the parameters for those n-grams are updated infrequently during training so that models with normal attentions may fail to distinguish these infrequent important n-grams in a specific context. We experiment our proposed model on five widely used benchmark datasets. Our model with multi-channel attentions outperforms strong baselines and achieves state-of-the-art results on all datasets.

2 Our Approach

The architecture of our approach is shown in Figure 2. The left side illustrates the backbone model following the sequence labeling paradigm; the right side elaborates the multi-channel attention module used to incorporate contextual n-gram information into the backbone model. Formally, given an input sentence $\mathcal{X} = x_1 x_2 \cdots x_i \cdots x_l$, where l is the input sequence length, our approach predicts its corresponding joint CWS and POS label sequence $\widehat{\mathcal{Y}} = \widehat{y}_1 \widehat{y}_2 \cdots \widehat{y}_i \cdots \widehat{y}_l$ by

$$\widehat{\mathcal{Y}} = f(\mathcal{X}, \mathcal{MA}(\mathcal{S})) \tag{1}$$

where $\mathcal{MA}$ denotes the multi-channel attention module. Let $\mathcal{N}$ denote a lexicon consisting of a list of n-grams collected for the entire corpus; $\mathcal{S} \subset \mathcal{N}$ is the set of n-grams in $\mathcal{X}$ that appears in $\mathcal{N}$. The details of applying the multi-channel attentions of n-grams to such framework are provided below.

[2]The first n-gram feature "放大" (*enlarge*) can suggest that there is word boundary after "大" and the POS label for "大" can be verb (VV), which does not reflect the correct interpretation of this sentence. On the contrary, the second n-gram feature "大道" (*avenue*) suggests "大" could be the initial character of a word, where there is no delimiter after it, and suggest the POS label for "大" can be noun (NN), which reflect the correct interpretation of this sentence.

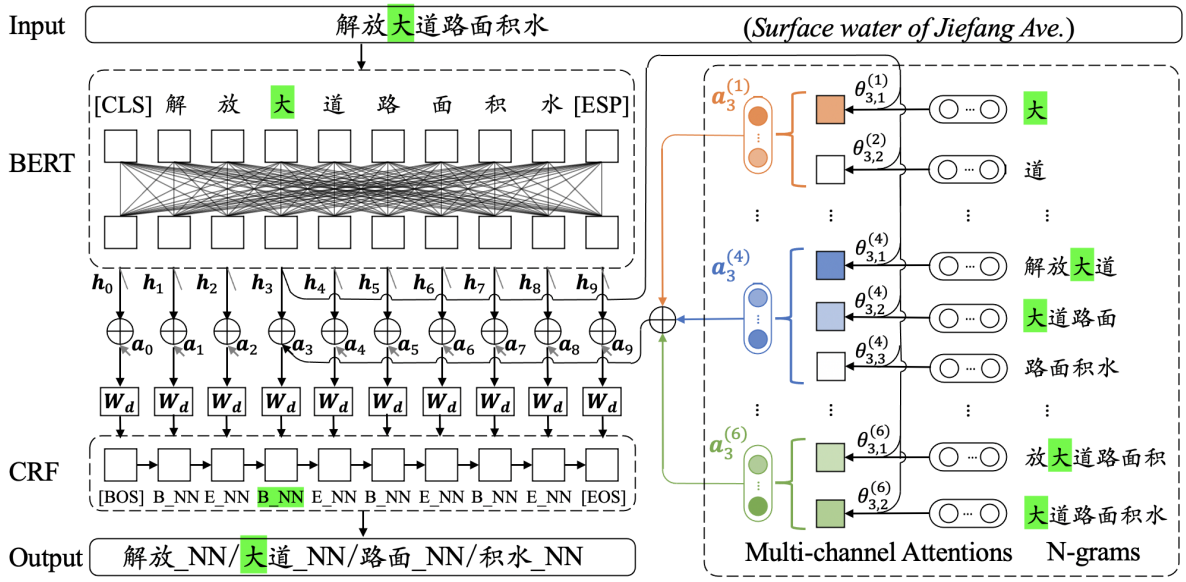


Figure 2: The overall architecture of our character-based model for the joint CWS and POS tagging with an example input and output. On the left is the backbone model following the sequence labeling paradigm; on the right is the multi-channel attention module with n-grams categorized by their length. Different attention channels for n-grams associated with "大" (*big*) are highlighted with distinct colors.

2.1 The Multi-channel Attentions

N-grams have been used as useful contextual features to enhance text representation for CWS and POS tagging in many studies (Song et al., 2009; Song and Xia, 2012; Song et al., 2012; Song and Xia, 2013; Shao et al., 2017; Zhang et al., 2018). However, for joint CWS and POS tagging, previous approaches to leveraging the n-gram features are limited to directly concatenating the n-gram embeddings with the input character embedding, where unimportant n-grams may mislead the model and result in incorrect predictions. Therefore, assigning appropriate weights to different n-grams regarding to their contexts is a potential effective solution (Higashiyama et al., 2019; Tian et al., 2020b) to the joint task and we propose to use multi-channel attention to tackle this mission. In detail, we first categorize n-grams by a specific metric, which in this study is either their frequencies or lengths and then model the grouped n-grams in separate channels of attentions. As a result, the contributions of the salient n-grams are highlighted and the attention weights are not dominated by frequent n-grams or the short ones that tend to appear in more sentences. Our model is thus able to leverage the highlighted n-grams accordingly and avoid being misled by the unimportant ones.

To train the attention module, for each instance $\mathcal{X}$, we collect all n-grams that appear in $\mathcal{N}$ to form a set of n-grams $\mathcal{S}$ to be used in the attention module. The multi-channel attention works as follows: in the first step, all n-grams are categorized into n groups according to their frequencies in a corpus or their lengths. We denote all n-grams as $\mathcal{S} = \{S_1, S_2, \cdots S_k, \cdots S_n\}$ and the n-grams in each group as $\mathcal{S}_k = \{s_1^{(k)}, s_2^{(k)}, \cdots s_j^{(k)}, \cdots s_{m_k}^{(k)}\}$; we use $\mathbf{e}_j^{(k)}$ to represent the vectored embedding of $s_j^{(k)}$. Afterwards, for character x_i, the attention weight of each n-gram $s_j^{(k)}$ in channel k is activated by a weight $a_{i,j}^{(k)}$:

$$a_{i,j}^{(k)} = \frac{\theta_{i,j}^{(k)} \cdot exp(u_{i,j}^{(k)})}{\sum_{j=1}^{m_k} \theta_{i,j}^{(k)} \cdot exp(u_{i,j}^{(k)})} \tag{2}$$

where $u_{i,j}^{(k)} = \mathbf{h}_i^\top \cdot \mathbf{e}_j^{(k)}$ is the inner product of $\mathbf{h}_i$ and $\mathbf{e}_j^{(k)}$ with $\mathbf{h}_i$ referring to the hidden vector from the encoder for x_i. Particularly, $\theta_{i,j}^{(k)}$ is a binary indicator indicating whether x_i is a part of $s_j^{(k)}$, which is formally defined by $\theta_{i,j}^{(k)} = 1$ if $x_i \in s_j^{(k)}$ and $\theta_{i,j}^{(k)} = 0$ otherwise. For example, in the example input illustrated in Figure 2, the extracted n-grams in channel $k = 4$ are $s_1^{(4)}$ = "解放大道" (*Jiefang Ave.*), $s_2^{(4)}$ = "大道路面" (*the surface of the Ave.*), and $s_3^{(4)}$ = "路面积水" (*the surface water of the road*)}. So,

	CTB5			CTB6			CTB7			CTB9			UD		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Word	494K	7K	8K	641K	60K	82K	718K	237K	245K	1,696K	136K	242K	99K	13K	12K
Sent	18K	350	348	23K	2K	3K	31K	10K	10K	106K	10K	16K	4K	500	500
OOV	-	8.1	3.5	-	5.4	5.6	-	5.5	5.2	-	2.9	3.1	-	12.1	12.4

Table 1: The statistics of all datasets in terms of the number of words and sentences. The out of vocabulary (OOV) rate in the development and test sets are computed based on the words appearing in the training set.

for $x_3 =$ "大" (*big*), $\theta^{(4)}_{3,1} = \theta^{(4)}_{3,2} = 1$ while $\theta^{(4)}_{3,3} = 0$ because "大" is a component of $s^{(4)}_1$ and $s^{(4)}_2$ but is not a part of $s^{(4)}_3$[3]. As a result, for each entire channel, its resulted weight is computed by

$$\mathbf{a}^{(k)}_i = \sum_{j=1}^{m_k} a^{(k)}_{i,j} \mathbf{e}^{(k)}_j \tag{3}$$

Finally, the overall attention of different n-grams for x_i is the concatenation of attentions from all channels:

$$\mathbf{a}_i = \bigoplus_n \delta_k \mathbf{a}^{(k)}_i \tag{4}$$

with a trainable positive parameter δ_k to balance the contribution of each channel.

2.2 Joint Tagging with the Attentions

To leverage the n-grams through the proposed attention module, we first obtain the hidden vector $\mathbf{h}_i$ of each x_i in the input sequence from the encoder (e.g., BERT (Devlin et al., 2019)) of the backbone model. Next, we feed the resulting $\mathbf{h}_i$ to the attention module and obtain its output $\mathbf{a}_i$, which contains the weighted contextual information carried by the n-gram features. Then, we incorporate such weighted information into the backbone model by concatenating $\mathbf{a}_i$ with $\mathbf{h}_i$ and align the resulting vector to the output dimension by a trainable matrix $\mathbf{W}_d$, which is represented by

$$\mathbf{u}_i = \mathbf{W}_d \cdot (\mathbf{h}_i \oplus \mathbf{a}_i) \tag{5}$$

Afterwards, we pass $\mathbf{u}_i$ to a conditional random field (CRF) decoder to estimate the joint label $\widehat{y}_i$ for x_i.

3 Experiment Settings

3.1 Datasets

In our experiments, five Chinese benchmark datasets are used, including CTB5, CTB6, CTB7, and CTB9 from the Penn Chinese TreeBank (Xue et al., 2005) and the Chinese GSD Treebank of Universal Dependencies (UD) (Nivre et al., 2016).[4] All CTB datasets are in simplified Chinese while UD is in traditional Chinese. Following Shao et al. (2017), we translate the UD dataset into simplified Chinese before experiments are conducted.[5] To obtain the training, development, and test data for each datasets, we follow previous studies (Jiang et al., 2008; Wang et al., 2011; Zhang et al., 2014; Shao et al., 2017) to split CTB5, CTB6, CTB7, and CTB9, and use the official splits for UD. Since UD contains two types of POS tags, namely, universal and language-specific tags, we follow the notation in Shao et al. (2017) and mark the former one as UD1 and the latter one as UD2. For the POS tag set, CTB has 33 tags; UD1 and UD2 have 15 and 42 tags, respectively. The statistics of all five datasets in terms of the number of words and sentences in the training, development, and test sets, respectively, are reported in Table 1. We also report out-of-vocabulary (OOV) rates in the development and test sets.

[3]This is highlighted in Figure 2.

[4]We obtain the official CTB5 (LDC2005T01), CTB6 (LDC2007T36), CTB7 (LDC2010T07), and CTB9 (LDC2016T13) from Linguist Data Consortium (LDC, `https://catalog.ldc.upenn.edu`) and the UD Chinese data (version 2.4) from `https://universaldependencies.org/`.

[5]We use the translation scripts from `https://github.com/skydark/nstools/tree/master/zhtools`.

	CTB5			CTB6			CTB7			CTB9			UD		
	AV	DLG	PMI	AV	DLG	PMI	AV	DLG	PMI	AV	DLG	PMI	AV	DLG	PMI
$\mathcal{N}$	34.5K	13.5K	10.6K	29.7K	16.6K	13.6K	27.9K	18.9K	16.6K	38.1K	27.5K	28.9K	19.6K	5.3K	1.1K

Table 2: The size of lexicon $\mathcal{N}$ constructed by AV, DLG, and PMI.

3.2 Lexicon Construction

To enhance the joint CWS and POS tagging through the multi-channel attentions, we need to construct the lexicon $\mathcal{N}$ which is simply a list of n-grams.[6] In this study, we do not want our approach to rely on existing n-gram resources. Therefore, we use three unsupervised methods to obtain n-grams from each datasets, namely, accessor variety (AV) (Feng et al., 2004), description length gain (DLG) (Kit and Wilks, 1999), and pointwise mutual information (PMI) (Sun et al., 1998).

Accessor Variety Given a character n-gram s, let left access number $L_{av}(s)$ be the number of distinct characters that precede s in the corpus. The right access number $R_{av}(s)$ is defined similarly. The AV score of s is the minimal number of the left and right access numbers:

$$AV(s) = min(L_{av}(s), R_{av}(s)) \tag{6}$$

In general, n-grams with higher AV scores are more likely to be words in Chinese. Since AV is sensitive to the size of dataset, in our experiments, we use different thresholds for the five datasets: 2 for CTB5, 3 for CTB6, 4 for CTB7, 5 for CTB9, and 1 for UD. For each dataset, we collect all n-grams whose AV scores are higher than the corresponding threshold to build the lexicon $\mathcal{N}$.

Description Length Gain DLG measures the wordnesshood of an n-gram s according to the change of the description length of a dataset $\mathcal{D}$ with and without treating s as a segment; formally, $DLG(s)$ is calculated by

$$DLG(s) = DL(\mathcal{D}) - DL(\mathcal{D}[r \rightarrow s] \oplus s) \tag{7}$$

where $\mathcal{D}[r \rightarrow s] \oplus s$ is the revised dataset of the original $\mathcal{D}$ with all the occurrences of s in $\mathcal{D}$ replaced by a single symbol r, and with the original n-gram s appended to the end. Besides, the description length of a corpus $\mathcal{D}$ is calculated by

$$DL(\mathcal{D}) = -\sum_{x \in \mathcal{V}} c(x) log \frac{c(x)}{|\mathcal{D}|} \tag{8}$$

where $\mathcal{V}$ is a character vocabulary containing all character types appearing in $\mathcal{D}$ and $c(x)$ denotes the count of character x in $\mathcal{D}$. In our experiments, the threshold for DLG is set to 0; that is, for each dataset $\mathcal{D}$, its lexicon $\mathcal{N}$ contains all n-grams whose DLG scores are higher than that threshold.

Pointwise Mutual Information PMI measures the co-occurrence of two adjacent characters x', x'' by

$$PMI(x', x'') = log \frac{p(x'x'')}{p(x')p(x'')} \tag{9}$$

where p is the probability distribution of a given n-gram (i.e., x', x'' and $x'x''$) in a dataset. For each dataset, we check all the character bi-grams in the corpus; a delimiter is inserted between the two characters if their PMI score is below a threshold. The n-grams in the resulted segmented corpus form the lexicon $\mathcal{N}$. In our experiments, we set the threshold for all datasets to 0.

To construct $\mathcal{N}$, we perform the aforementioned unsupervised methods on the raw text of the training set and the development set combined for each dataset.[7] Next, we filter out n-grams whose frequency is no more than a threshold.[8] Finally, for all datasets we keep the n-grams whose lengths are within the range of $[1, 10]$. Table 2 shows the sizes of the lexicons for the five datasets.

[6] Technically, this lexicon can be constructed through a series of existing resources or automatic methods.

[7] Note that we only use the raw text in the development set, without using its gold labels.

[8] The thresholds for CTB5-CTB9 and UD are 2, 3, 4, 5, and 1, respectively

Model	Settings		CTB5		CTB6		CTB7		CTB9		UD1		UD2	
	Cat.	$\mathcal{N}$	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint
BERT	N/A	N/A	98.09	96.85	97.42	94.79	96.94	93.47	97.37	94.22	97.86	94.90	97.90	94.86
Norm Att.	N/A	AV	98.06	96.99	97.47	94.83	97.02	93.64	97.42	94.35	98.07	95.16	98.10	95.06
		DLG	98.19	97.06	97.49	94.84	97.04	93.66	97.50	94.37	98.15	95.09	98.14	95.04
		PMI	98.18	97.03	97.53	94.86	96.92	93.54	97.46	94.39	98.16	95.13	98.08	95.01
Our Model	Freq.	AV	98.30	97.11	97.54	**94.96**	97.03	**93.81**	97.59	94.57	98.33	95.38	98.24	95.24
		DLG	98.27	97.06	97.53	94.94	97.00	93.75	97.62	94.64	98.30	95.42	98.21	95.23
		PMI	98.39	97.12	**97.56**	94.92	**97.05**	93.80	**97.65**	**94.66**	98.37	95.36	98.24	95.25
	Len.	AV	98.40	97.16	97.54	94.92	97.01	93.71	97.59	94.57	98.27	95.41	98.17	**95.26**
		DLG	**98.45**	**97.19**	97.55	94.93	97.02	93.77	97.62	94.63	**98.38**	**95.46**	98.15	95.20
		PMI	98.39	97.15	97.50	94.95	97.03	93.75	97.60	94.56	98.35	95.44	**98.25**	95.17

(a) Results from BERT

Model	Settings		CTB5		CTB6		CTB7		CTB9		UD1		UD2	
	Cat.	$\mathcal{N}$	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint
ZEN	N/A	N/A	98.20	97.05	97.51	94.90	97.90	93.58	97.50	94.60	98.21	95.15	98.20	94.98
Norm Att.	N/A	AV	98.35	97.17	97.56	94.95	97.97	93.77	97.57	94.70	98.30	95.38	98.31	95.19
		DLG	98.30	97.09	97.57	94.98	97.97	93.76	97.55	94.68	98.23	95.30	98.30	95.18
		PMI	98.37	97.19	97.60	95.00	97.95	93.72	97.58	94.69	98.28	95.36	98.29	95.14
Our Model	Freq.	AV	98.42	97.24	97.64	95.02	97.05	93.91	97.60	94.79	98.36	95.57	98.38	95.32
		DLG	**98.52**	97.33	97.63	95.01	97.01	93.85	97.60	94.77	98.34	95.54	98.41	95.40
		PMI	98.39	97.26	**97.67**	**95.08**	97.03	93.90	**97.62**	**94.80**	**98.37**	95.60	98.40	95.39
	Len.	AV	98.43	97.27	97.63	95.06	97.03	93.87	97.58	94.75	98.31	**95.63**	98.41	95.39
		DLG	98.49	97.33	97.61	95.05	97.06	93.89	97.57	94.72	98.30	95.50	98.40	95.37
		PMI	98.48	**97.35**	97.64	95.07	**97.08**	**93.93**	97.59	94.79	98.35	95.57	**98.42**	**95.41**

(b) Results from ZEN

Table 3: F scores for segmentation and joint tagging of MCAPOST under different settings on the development set of five datasets, where the results of models with BERT encoder and ZEN encoder are reported in (a) and (b), respectively. "Freq." and "Len." refer to the n-gram categorization strategies based on n-gram frequency and n-gram length; "AV", "DLG", and "PMI" stands for different ways to construct the lexicon $\mathcal{N}$; "N/A" is the abbreviation for *not applicable*.

3.3 Implementations

Since text representation plays an important role in model performance (Conneau et al., 2017; Song et al., 2017; Song et al., 2018), in our experiment, we try two well-known Chinese text encoders as the backbone model: Chinese version of pre-trained BERT[9] (Devlin et al., 2019) and ZEN[10] (Diao et al., 2019). For both BERT and ZEN, we follow their default settings in our experiments (i.e., for both BERT and ZEN, we use 12 layers of multi-head attentions on character encoding with the dimension of hidden vectors set to be 768; for ZEN, we use 6 layers of n-gram representations). For the models with the multi-channel attention module, we use two criteria to categorize the n-grams that are used in different channels. The first is by frequency, where n-grams whose counts in the dataset are within the same range $[c_k, c_{k+1})$ are categorized into one group and are compared and weighted within the same channel in the attention module. In our experiments, we set $c_k = 2^k$, for $k \in [1, 10]$ and $c_{11} = +\infty$. The second criterion is by n-gram length, where n-grams with the same n value are in the same group and fed into the same channel in the attention module.

For other settings, we randomly initialize the n-gram embeddings used in the attention module, with their dimension matching the hidden vector size of the BERT/ZEN encoder, i.e. 768; we set the dropout rate to 0.2, the training batch size to 16, and learning rate to 1e-5. We fine-tune all parameters in BERT and ZEN and use the negative log-likelihood loss function to optimize all models. For evaluation, we follow previous studies (Zheng et al., 2013; Kurita et al., 2017; Shao et al., 2017; Zhang et al., 2018) to use the F scores of the segmentation and joint label, where the latter one is the main focus of this

[9]We use the Chinese base model from `https://s3.amazonaws.com/models.huggingface.co/`.

[10]We obtain the pre-trained ZEN model from `https://github.com/sinovation/ZEN`.

Models	CTB5		CTB6		CTB7		CTB9		UD1		UD2	
	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint
Jiang et al. (2008)	97.85	93.41	-	-	-	-	-	-	-	-	-	-
Kruengkrai et al. (2009)	97.87	93.67	-	-	-	-	-	-	-	-	-	-
Zhang and Clark (2010)	97.78	93.67	-	-	-	-	-	-	-	-	-	-
Sun (2011)	98.17	94.02	-	-	-	-	-	-	-	-	-	-
Wang et al. (2011)	98.11	94.18	95.79	91.12	95.65	90.46	-	-	-	-	-	-
Qian and Liu (2012)	97.85	93.53	-	-	-	-	-	-	-	-	-	-
Shen et al. (2014)	98.03	93.80	-	-	-	-	-	-	-	-	-	-
Kurita et al. (2017)	98.41	94.84	-	-	96.23	91.25	-	-	-	-	-	-
Shao et al. (2017)	98.02	94.38	-	-	-	-	96.67	92.34	95.16	89.75	95.09	89.42
Zhang et al. (2018)	98.50	94.95	96.36	92.51	96.25	91.87	-	-	-	-	-	-
BERT	98.28	96.03	97.36	94.65	96.78	93.38	97.33	94.40	97.74	94.82	97.70	94.76
Norm Att. (PMI)	98.71	96.45	97.31	94.68	97.08	93.74	97.54	94.55	98.09	95.34	98.01	94.96
Our Model (Freq. + PMI)	98.64	96.59	97.28	94.71	**97.18**	94.01	**97.72**	94.79	98.23	95.35	98.26	95.37
Our Model (Len. + PMI)	98.73	96.60	97.30	94.74	97.13	93.98	97.69	94.78	**98.29**	95.50	**98.27**	95.38
ZEN	98.61	96.60	97.35	94.70	97.09	93.80	97.64	94.64	98.14	95.15	98.02	95.05
Norm Att. (PMI)	98.68	96.73	97.30	94.75	97.10	93.95	97.58	94.79	98.19	95.33	98.13	95.26
Our Model (Freq. + PMI)	98.74	96.80	97.32	94.81	97.13	94.06	97.67	**94.87**	98.25	95.55	98.25	95.38
Our Model (Len. + PMI)	**98.79**	**96.82**	**97.38**	**94.82**	97.16	**94.09**	97.66	94.82	98.28	**95.59**	98.23	**95.41**

Table 4: F scores of segmentation and joint tagging on the test set of five datasets from previous studies, and our models and baselines (using BERT/ZEN encoder) with and without the multi-channel attentions.

paper[11]. We train all models on the training set, preserve the one achieving the highest joint F-score on the development set, and finally evaluate it on the test set.

4 Results and Analyses

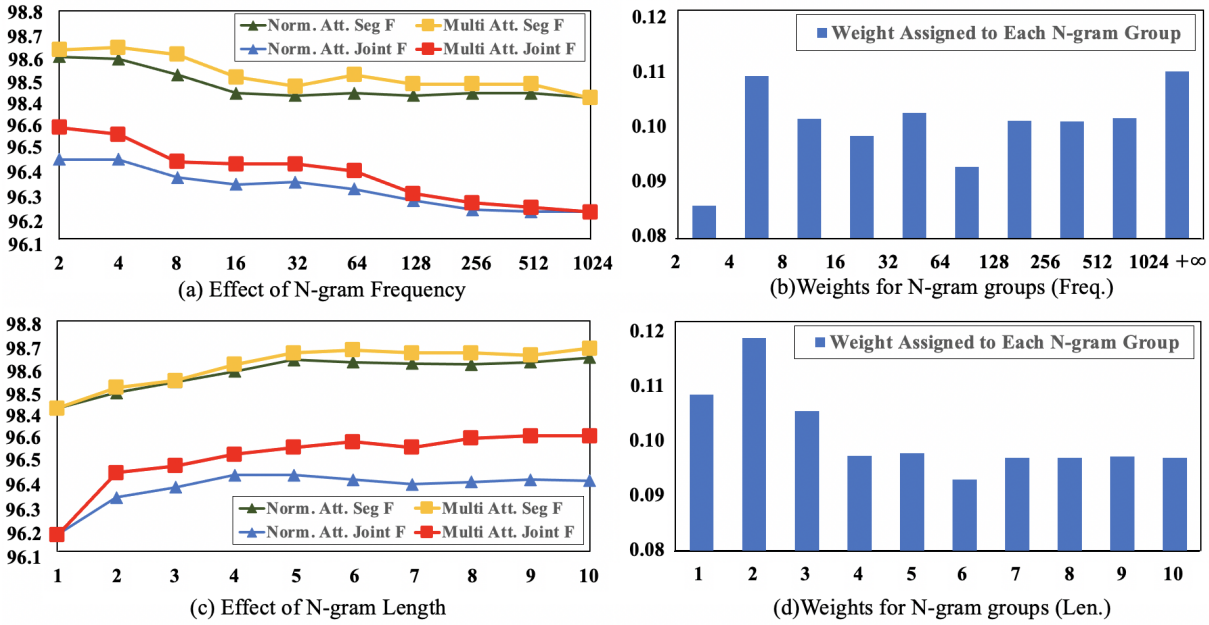
4.1 Overall Performance

In experiments, we test our model with the multi-channel attention module under different settings, where two different encoders (i.e., BERT and ZEN), two strategies to categorize n-grams (one is based on n-gram frequency (denoted by "Freq.") and the other is based on n-gram length (denoted by "Len.")), and three ways to construct the lexicon $\mathcal{N}$ are used. In addition, we also run baseline models without using the attention module as well as the ones using normal attentions (single-channel, denoted by "Norm Att.") to model all n-grams. The results (the F scores of the segmentation and joint labels) of our models and baseline models on the development sets are reported in Table 3.

There are several observations. First, the multi-channel attention works well with different (i.e., BERT and ZEN) encoders; it helps both segmentation and the joint task consistently on all datasets when compared with the BERT/ZEN baseline without using it. Second, the proposed multi-channel attention module can be applied with different n-gram categorization strategies (i.e., by n-gram frequency and length). Especially, even though the performance of BERT/ZEN model with normal attentions is rather good on the joint task, the multi-channel attention is still able to further boost its performance. This observation shows that grouping and modeling n-grams in different channels could better leverage the n-gram features compared with modeling them together in normal attentions. Third, our model shows its robustness with respect to different ways of constructing $\mathcal{N}$, where similar results are observed over construction methods of AV, DLG, and PMI on all datasets. For example, on CTB7, the absolute differences of the F score between our models under different settings are no more than 0.06%.

Moreover, we also compare our experimental results with representative studies in the past decade on the test set of five benchmark datasets. The F-scores of their studies and the ones from our models and baselines are reported in Table 4, where the lexicon $\mathcal{N}$ used for our models and baselines is constructed based on PMI. From the results, our model with BERT and multi-channel attentions outperforms all baselines and previous studies on all datasets with respect to the F score of the joint labels. In addition, when equipped with ZEN, our model can further outperform BERT-based models on the F-scores of the joint CWS and POS tagging task. Compared with previous studies, where extra knowledge or

[11]The evaluations are performed by `https://github.com/chakki-works/seqeval`.



(a) Effect of N-gram Frequency

(b)Weights for N-gram groups (Freq.)

(c) Effect of N-gram Length

(d)Weights for N-gram groups (Len.)

Figure 3: The effect of different n-gram groups categorized by frequency and length on CTB5, where the n-gram lexicon $\mathcal{N}$ is constructed by PMI. The segmentation and joint tagging F scores of the models (using BERT encoder) with normal and multi-channel attentions are illustrated in (a) and (c), where $\mathcal{N}$ is constructed by including n-grams whose frequency is in range $[2^i, +\infty)$ (1≤i≤10) and whose length is in range $[1, n]$ (1≤n≤10), respectively. The weights (i.e., δ_k in Eq. 4) assigned to n-gram groups categorized by frequency and length in the multi-channel attention module are shown in (b) and (d), respectively.

resources, such as well-defined dictionaries (Wang et al., 2011; Zhang et al., 2014), syntactic features form manual crafted resources (Zhang and Clark, 2010), information of Chinese radicals (Shao et al., 2017), or large auto-processed data (Zhang et al., 2018) are used, our approach only leverages the resource from the datasets, which reduces the cost to train a joint CWS and POS tagger. Overall, the above results demonstrate that weighting n-grams separately is an appropriate approach to improve joint CWS and POS tagging without requiring extra knowledge.

4.2 Effect of N-gram Categorization Methods

We analyze the effect of different categorization methods, i.e., n-gram frequency and n-gram length, to the joint task. For frequency-based methods, we tried different frequency thresholds from 2^1 to 2^{10}, where n-grams whose frequency in the dataset is less than 2^i $(1 \leq i \leq 10)$ are ignored in the attention module; for length-based methods, we try the number from 1 to 10, where n-grams with their length form 1 to n (≤ 10) are considered. We run experiments with our models and normal attentions using BERT encoder under these settings on CTB5 with the lexicon $\mathcal{N}$ constructed by PMI, where the curves (F-scores) of the models with the two categorization methods are reported in Figure 3(a) and 1(c), respectively.

For frequency-based categorization method, the performance of models with normal and multi-channel attentions drops when the frequency threshold increases (see Figure 3(a)). Yet, our model shows a smaller drop over the normal attention model, indicating that multi-channel could be a solution to enhance the attention learned on the same data. We also find that the curves tend to stabilize when the frequency reaches 128. This could be explained by that although frequent n-grams may provide useful information to the task, this information can also be learned by the backbone, leading to a relatively small improvement. In addition, we compare the weights assigned to each n-gram group in Figure 3(b), from which we find n-grams with the frequency of $[4, 8)$ receives a relatively high weight. One possible reason could be that these n-grams provide important cues for the joint task, which is hard for the backbone model to learn because these n-grams do not appear frequently. Therefore, by categorizing n-grams based on their frequency, the model can highlight important n-grams that are infrequent, and thus the model is not dominated by the frequently appearing n-grams.

Freq.	Ranked N-gram Examples According to Their Total Weights ($\times 10^{-3}$) in CTB5
$[2^1, 2^2)$	“家用纺织品进出口公司” (home textiles import & export corp.) 1.6; ...; “有关各方” (interested parties) 1.2; ...
$[2^2, 2^3)$	“中国红十字会” (Red Cross Society of China) 1.6; ...; “进出口货物” (import and export goods) 0.8; ...
$[2^3, 2^4)$	“社会主义市场经济” (Socialist Market Economy) 2.1; ...; “进一步改善” (further improvement) 1.2; ...
$[2^4, 2^5)$	“高新技术产业” (high-tech industry) 2.0; ...; “公务员” (civil servant) 1.3; ...
$[2^5, 2^6)$	“比去年同期增长” (increase from the same period last year) 2.8; ...; “中国政府” (Chinese government) 2.5; ...
$[2^6, 2^7)$	“香港特别行政区” (Hong Kong Special Administrative Region) 6.3; ...; “人口” (population) 3.7; ...
$[2^7, 2^8)$	“外商投资” (foreign investment) 5.3; ...; “包括” (include) 4.6; “希望” (hope) 4.4; “七十” (seventy) 4.4; ...
$[2^8, 2^9)$	“合作” (cooperation) 8.9; “人民” (people) 8.7; “第一” (first) 8.7; “自己” (myself) 8.6; ...
$[2^9, 2^{10})$	“百分之” (percentage) 18.3; “国家” (nation) 14.5; ...; “好” (good) 6.8; ...; “明” (clear) 6.6; ...
$[2^{10}, +\infty)$	“的” (of) 63.0; “一” (one) 23.7; “国” (country) 17.8; “在” (at) 16.2; “年” (year) 15.2; “是” (year) 14.2; ...

Table 5: Ranked n-gram examples w.r.t. their received weights in their frequency groups in our model trained on CTB5 with PMI to construct $\mathcal{N}$. The weight of each n-gram is attached to its English translation.

Len.	Ranked N-gram Examples According to Their Total Weights ($\times 10^{-2}$) in CTB5
1	“的” (of) 4.4; “国” (nation) 1.7; “一” (one) 1.7; “中” (middle) 1.2; “在” (at) 1.1; “年” (year) 1.1; ...
2	“中国” (China) 1.2; “台湾” (Taiwan) 0.8; “经济” (economy) 0.7; “企业” (company) 0.6; “投资” (investigate) 0.6; ...
3	“百分之” (percentage) 2.2; “新华社” (Xinhua News Agency) 1.6; “十二月” (December) 0.6; ...;
4	“外商投资” (foreign investment) 1.3; “百分之十” (ten percentage) 1.3; “利用外资” (use foreign capital) 1.0; ...
5	“一九九七年” (the year of nineteen ninety seven) 2.9; “百分之二十” (twenty percent) 2.4; ...
6	“外商投资企业” (foreign-invested enterprise) 6.8; “国内生产总值” (GDP) 3.0; ...
7	“香港特别行政区” (Hong Kong Special Administrative Region) 19.7; ...
8	“社会主义市场经济” (Socialist Market Economy) 8.7; “人均国内生产总值” (GDP per capita) 4.3; ...
9	“香港特别行政区政府” (Government of the Hong Kong Special Administrative Region) 17.7; ...
10	“中国石油天然气总公司” (China National Petroleum Corporation) 22.6; ...

Table 6: Ranked n-gram examples w.r.t. their received weights in their length groups in our model trained on CTB5 with PMI to construct $\mathcal{N}$. The weight of each n-gram is attached to its English translation.

For length-based categorization method, the performance of both models with normal and multi-channel attentions tend to improve with higher n-gram length threshold. In this case, our model shows a bigger improvement over the normal attention model and the curves tend to stabilize when the X-axis reaches 6 (see Figure 3(c)). A possible reason could be that the number of new n-grams being leveraged with the raise of n-gram length threshold is decreasing so that their influence to the overall performance could be hard to observe. Similarly, we also compare the weights assigned to n-grams grouped by their length and illustrate them in Figure 3(d). In the histogram, we find that bi-grams receive the highest weight over all n-gram groups, which could be attributed to the fact that most words in Chinese contains two characters. On the contrary, n-grams with more characters tend to have fewer influence to the task because it is uncommon to see very long words in Chinese.

4.3 N-gram Analyses

To explore the way of our model leveraging n-grams in different properties, we apply our model (with BERT encoder) trained on CTB5 with PMI lexicon construction method to the whole data of CTB5[12] and for each n-gram we sum its assigned attention. The n-gram examples categorized by their frequency and length are shown in Table 5 and Table 6, respectively, where n-grams and their assigned attentions are presented in the decreasing order. For the n-gram examples in both tables, we find that almost all n-grams with high weights are normal Chinese words or phrases. Due to the lexicon is constructed by an unsupervised method, where many n-grams in it are not well-formed words, this observation indicates

[12]This means we run the model on the training, development and test set of CTB5.

our model in both settings can distinguish important n-grams and assign them high weights. In addition, for each frequency group in Table 5, there are cases where long n-grams receive a higher weight than the short n-grams, e.g., for n-grams in group $[2^5, 2^6)$, the weight assigned to the seven-gram "比去年同期增长" (*increase from the same period last year*) is 2.8×10^{-3} while the weight assigned to the four-gram "中国政府" (*Chinese government*), which is more frequent than the seven-gram in CTB5, is 2.5×10^{-3}.[13] Similarly, in the n-gram examples in Table 6, top ranked long n-grams tend to have higher weights compared with the top ranked short ones. This observation shows that our model with multi-channel attentions can appropriately model the important long n-grams, i.e., assigning high weights to the long n-grams that are common words or phrases, even though they may be infrequent in the dataset.

5 Related Work

There are basically two approaches to CWS and POS tagging: to perform the two tasks it in a pipeline framework; or to treat them as a joint task where the two tasks are conducted simultaneously, which is known as joint CWS and POS tagging. Ng and Low (2004) provided a comprehensive study to compare the two approaches and found that the joint approach outperform the pipeline one. Therefore, in the past two decades, the majority of studies on CWS and POS tagging applied the joint approach to these tasks (Ng and Low, 2004; Jiang et al., 2008; Jiang et al., 2009; Wang et al., 2011; Sun, 2011; Zeng et al., 2013), where n-grams are widely used as features carrying contextual information to improve model performance. Recently, neural methods, especially the recurrent neural networks (e.g., bi-LSTM) have demonstrated their effectiveness to encode contextual information, and thus significantly improve the model performance in joint CWS and POS tagging. Even though, improvements can still be obtained when n-grams are incorporated into the neural taggers (Zheng et al., 2013; Kurita et al., 2017; Shao et al., 2017; Zhang et al., 2018; Tian et al., 2020a). For example, Kurita et al. (2017) used a stacked bi-LSTM model to incorporate n-grams and achieved state-of-the-art results on CTB5 and CTB7. In addition to n-grams, approaches leveraging external resources are also used to improve joint CWS and POS tagging: Shao et al. (2017) leveraged n-grams and radical information of Chinese characters to enhance model performance; Zhang et al. (2018) pre-trained their character embeddings on large data where both segmentation and POS lebels are auto-tagged by an existing model. Compared to these studies, our model provide a way to leverage n-grams through a multi-channel attention mechanism, where n-grams are categorized by their frequencies or lengths and the n-grams in the same category are compared and weighted. Therefore, the n-grams that contribute more to the joint task in a specific context are highlighted and the model will not be dominated by the frequent or short n-grams (short n-grams also tend to be frequent) because their parameters are intensively updated during training.

6 Conclusion

In this paper, we propose a neural character-based tagger for joint CWS and POS tagging, where a multi-channel attention mechanism is used to leverage context information carried by n-grams. In detail, for multi-channel attention, we categorize n-grams according to a specific metric, such as their frequencies or lengths, and model the n-grams separately in each attention channel. In doing so, n-grams with different properties (frequencies or lengths) can be weighted and distinguished separately under a specific context, so that they can be reasonably treated for the joint CWS and POS tagging task. Experimental results on five Chinese benchmark datasets shows that our approach with the multi-channel attentions can work well with n-grams extracted by different methods and provides consistent improvements over strong baseline taggers (i.e., BERT and ZEN) without using it. Particularly, our model with ZEN achieves the state-of-the-art performance for joint CWS and POS on all datasets.

Acknowledgements

This work is supported by The Chinese University of Hong Kong (Shenzhen) under University Development Fund UDF01001809.

[13] In CTB5, the frequency of "比去年同期增长" (*increase from the same period last year*) is 44 and that of "中国政府" (*Chinese government*) is 55.

References

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, pages 4171–4186.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2019. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. *ArXiv*, abs/1911.00720.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor Variety Criteria for Chinese Word Extraction. *Computational Linguistics*, 30(1):75–93.

Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, and Isaac Okada. 2019. Incorporating Word Attention into Character-Based Word Segmentation. 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)*, pages 2699–2709, Minneapolis, Minnesota, June.

Wenbin Jiang, Liang Huang, Qun Liu, and Yajuan Lü. 2008. A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of ACL-08: HLT*, pages 897–904, Columbus, Ohio, June.

Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 522–530, Suntec, Singapore, August.

Chunyu Kit and Yorick Wilks. 1999. Unsupervised Learning of Word Boundary with Description Length Gain. In *EACL 1999: CoNLL-99 Computational Natural Language Learning*.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 513–521, Suntec, Singapore, August.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2017. Neural Joint Model for Transition-based Chinese Syntactic Analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1214, Vancouver, Canada, July.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 277–284, Barcelona, Spain, July.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666.

Xian Qian and Yang Liu. 2012. Joint Chinese Word Segmentation, POS Tagging and Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 501–511, Jeju Island, Korea, July.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 173–183, Taipei, Taiwan, November.

Mo Shen, Hongxiao Liu, Daisuke Kawahara, and Sadao Kurohashi. 2014. Chinese Morphological Analysis with Character-level POS Tagging. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258, Baltimore, Maryland, June.

Yan Song and Fei Xia. 2012. Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *LREC*, pages 3853–3860.

Yan Song and Fei Xia. 2013. A Common Case of Jekyll and Hyde: The Synergistic Effect of Using Divided Source Training Data for Feature Augmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 623–631, Nagoya, Japan, October.

Yan Song, Chunyu Kit, and Xiao Chen. 2009. Transliteration of Name Entity via Improved Statistical Translation on Character Sequences. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, Suntec, Singapore, August.

Yan Song, Prescott Klassen, Fei Xia, and Chunyu Kit. 2012. Entropy-based Training Data Selection for Domain Adaptation. In *Proceedings of COLING 2012: Posters*, pages 1191–1200, Mumbai, India, December.

Yan Song, Chia-Jung Lee, and Fei Xia. 2017. Learning Word Representations with Regularization from Prior Knowledge. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 143–152, Vancouver, Canada, August.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 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*, pages 175–180, New Orleans, Louisiana, June.

Maosong Sun, Dayang Shen, and Benjamin K. Tsou. 1998. Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1265–1271, Montreal, Quebec, Canada, August.

Weiwei Sun. 2011. A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1385–1394, Portland, Oregon, USA, June.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. 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, July.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020b. Improving Chinese Word Segmentation with Wordhood Memory Networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317, Chiang Mai, Thailand, November.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural language engineering*, 11(2):207–238.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013. Graph-based Semi-Supervised Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 770–779, Sofia, Bulgaria, August.

Yue Zhang and Stephen Clark. 2010. A Fast Decoder for Joint Word Segmentation and POS-Tagging Using a Single Discriminative Model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 843–852, Cambridge, MA, October.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Type-Supervised Domain Adaptation for Joint Segmentation and POS-Tagging. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 588–597, Gothenburg, Sweden, April.

Meishan Zhang, Nan Yu, and Guohong Fu. 2018. A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(9):1528–1538.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep Learning for Chinese Word Segmentation and POS Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 647–657, Seattle, Washington, USA, October.