That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory

Xuemei Tang[1,2] **and Qi Su**[2,3,4*]
[1]Department of Information Management, Peking University, Bejing, China
[2]Digital Humanities Center of Peking University, Bejing, China
[3]School of Foreign Languages, Peking University, Bejing, China
[4]MOE Key Lab of Computational Linguistics, School of EECS, Peking University, Bejing, China
tagxuemei@stu.pku.edu.cn
sukia@pku.edu.cn

Abstract

The evolution of language follows the rule of gradual change. Grammar, vocabulary, and lexical semantic shifts take place over time, resulting in a diachronic linguistic gap. As such, a considerable amount of texts are written in languages of different eras, which creates obstacles for natural language processing tasks, such as word segmentation and machine translation. Although the Chinese language has a long history, previous Chinese natural language processing research has primarily focused on tasks within a specific era. Therefore, we propose a cross-era learning framework for Chinese word segmentation (CWS), CROSSWISE, which uses the Switch-memory (SM) module to incorporate era-specific linguistic knowledge. Experiments on four corpora from different eras show that the performance of each corpus significantly improves. Further analyses also demonstrate that the SM can effectively integrate the knowledge of the eras into the neural network.

1 Introduction

As a human-learnable communication system, language does not remain static but instead evolves over time. The rate of change between different aspects of language, such as grammar, vocabulary, and word meaning, vary due to language contact and many other factors, which has led to the diachronic linguistic gap. An example of this can be seen in, "That slepen al the nyght with open ye (That sleep all the night with open eye)," which is a sentence from The Canterbury Tales, written in Middle English by Geoffrey Chaucer at the end of the 14th century. It is difficult for people without an understanding of Middle English to make sense of this sentence. Furthermore, some discourses contain both modern English and Old English due to citation or rhetorical need. For example, Shakespeare's fourteen lines of poetry are often quoted in contemporary novels. This kind of era-hybrid text creates barriers to natural language processing tasks, such as word segmentation and machine translation.

Sample from MSR										
Golds	(wait)		(who)	(come)	(slove)			(ne)	?	
	等待		谁	来	解决			呢	?	
PKUSeg	等待		谁	来	解决			呢	?	
JiaYan	等	待	谁	来	解	决		呢	?	
Sample from AWIKI										
Golds	(Qi)	(Cui Shu)		(lead army)		(attack)		(Lv)	。	
	齐	崔杼		帅师		伐		莒	。	
PKUSeg	齐崔	杼帅			师伐莒				。	
JiaYan	齐	崔杼		帅师		伐		莒	。	

Table 1: Illustration of the different segmentation results for a modern Chinese sentence and an ancient Chinese sentence with different segmentation toolkits.

The Chinese language has the honor of being listed as one of the world's oldest languages and, as such, has seen several changes over its long history. It has undergone various incarnations, which are recognized as Archaic (Ancient) Chinese, Middle Ancient Chinese, Near Ancient Chinese, and Modern Chinese. Notably, most Chinese NLP tasks skew towards Modern Chinese. Previous research has primarily focused on addressing the CWS problem in Modern Chinese and has achieved promising results, such as Chinese Word Segmentation (CWS) (Zheng et al., 2013; Chen et al., 2015; Zhang et al., 2016; Xu and Sun, 2016; Shao et al., 2017; Yang et al., 2017; Zhang et al., 2018; Tian et al., 2020b,a). Although CWS for ancient Chinese has been recognized in recent years, the processing of language-hybrid texts is still an open question. As shown in

* Corresponding author

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7830 - 7840

Table 1, PKUSeg (Luo et al., 2019a) is a Chinese segmenter that is trained with a modern Chinese corpus; while it can segment modern Chinese sentences correctly, its accuracy drops sharply when applied to ancient Chinese. Conversely, the ancient Chinese segmenter JiaYan[1] performs well on ancient Chinese text but fails to perform well on Modern Chinese texts. Therefore, it is necessary to develop appropriate models to undertake cross-era NLP tasks.

To address this need, we propose CROSSWISE (CROsS-ear Segmentation WIth Switch-mEmory), which is a learning framework that deals with cross-era Chinese word segmentation (CECWS) tasks. The framework integrates era-specific knowledge with the Switch-memory mechanism to improve CWS for era-hybrid texts. More specifically, we utilized the abilities of both CWS and sentence classification tasks to predict segmentation results and era labels. We also incorporated the Switch-memory module to include knowledge of different eras, which consists of key-value memory networks (Miller et al., 2016) and a switcher. In order to store era-specific knowledge by several memory cells, key-value memory networks are used. The sentence discriminator is considered to be a switcher that governs the quantity of information in each memory cell that is integrated into the model. For each memory cell, we map candidate words from the dictionary and word boundary information to keys and values..

The main contributions of this paper are summarized as follows:

- Cross-era learning is introduced for CWS, we share all the parameters with a multi-task architecture. The shared encoder is used to capture information that several datasets from different eras have in common. This single model can produce different words segmentation granularity according to different eras.

- The Switch-memory mechanism is used to integrate era-specific knowledge into the neural network, which can help improve the performance of out of vocabulary (OOV) words. This study proposes two switcher modes (*hard-switcher* and *soft-switcher*) to control the quantity of information that each cell will feed into the model.

- Experimental results from four CWS datasets with different eras confirm that the performance of each corpus improves significantly. Further analyses also demonstrate that this model is flexible for cross-era Chinese word segmentation.

2 Related Work

Chinese word segmentation is generally considered to be a sequence labeling task, namely, to assign a label to each character in a given sentence. In recent years, many deep learning methods have been successfully applied to CWS (Zheng et al., 2013; Chen et al., 2015; Zhang et al., 2016; Xu and Sun, 2016; Shao et al., 2017; Yang et al., 2017; Kurita et al., 2017; Liu et al., 2018; Zhang et al., 2018; Ye et al., 2019a; Higashiyama et al., 2019; Huang et al., 2020b; Tian et al., 2020b,a,c; Liu et al., 2021). Among these studies, some indicate that context features and external knowledge can improve CWS accuracy (Kurita et al., 2017; Yang et al., 2017; Zhang et al., 2018; Liu et al., 2018; Tian et al., 2020b,a,c). Studies from Liu et al. (2018) and Zhang et al. (2018) leveraged the dictionary to improve the task; n-gram is also an effective context feature for CWS (Kurita et al., 2017; Tian et al., 2020b; Shao et al., 2017). The use of syntactic knowledge generated by existing NLP toolkits to improve CWS and part-of-speech (POS) has been established by Tian et al. (2020b). Furthermore, Tian et al. (2020c) incorporated wordwood information for neural segmenters and achieved a state-of-the-art performance at that time.

It is common practice to jointly train CWS and other related tasks based on a multi-task framework. Chen et al. (2017) took each segmentation criterion as a single task and proposed an adversarial multi-task learning framework for multi-criteria CWS by extracting shared knowledge from multiple segmentation datasets. Yang et al. (2017) investigated the effectiveness of several external sources for CWS by a globally optimized beam-search model. They considered each type of external resource to be an auxiliary classification, and then leveraged multi-task learning to pre-train the shared parameters used for the context modeling of Chinese characters. Liu et al. (2018) jointly trained the CWS and word classification task by a unified framework model. Inspired by these successful studies, this study also incorporated ideas from the multi-task framework, and jointly trained the CWS task and

[1] http://github.com/jiayan/Jiayan/

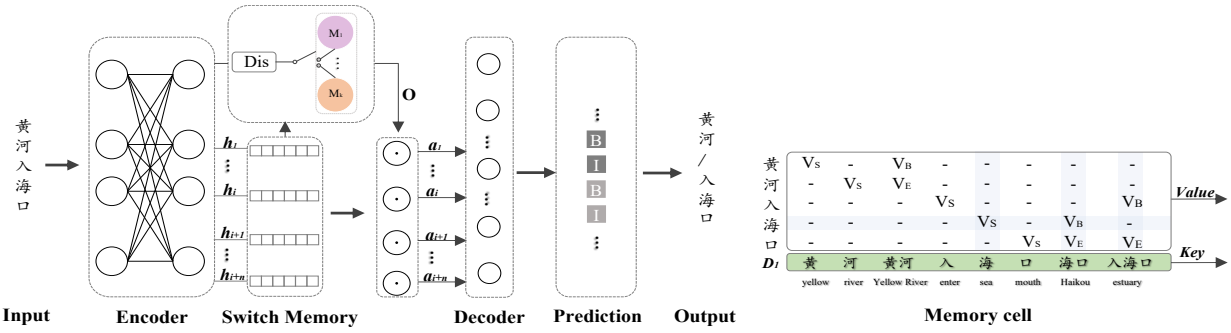

Figure 1: CROSSWISE for cross-era Chinese word segmentation. "Dis" represents the discriminator, specially, the sentence classifier. "M_1" is the first memory cell; its internal structure is shown on the right of the figure. For each character, the first memory cell extracts all candidate words from the input sentence and only retains ones that appeared in the first dictionary (candidate words as keys, words' boundary information as value).

the sentence classification task to enhance the performance of cross-era CWS.

Recently, some studies have noticed the linguistic gap due to the differences between eras. Ceroni et al. (2014) proposed a time-aware re-contextualization approach to bridge the temporal context gap. Chang et al. (2021) reframed the translation of ancient Chinese texts as a multi-label prediction task, then predicted both translation and its particular era by dividing ancient Chinese into three periods.

The use of key-value memory networks were introduced to the task of directly reading documents and answering questions by Miller et al. (2016), which helped to bridge the gap between direct methods and the use of human-annotated or automatically constructed Knowledge Bases. Tian et al. (2020c) applied this mechanism to incorporate n-grams into the neural model for CWS.

Encouraged by the above works, this study designed a multi-task model for cross-era CWS by jointly training the sentence classification task and CWS through the use of a unified framework model. Key-value memory networks are used to integrate era-specific knowledge into the neural network, as was done in research by Tian et al. (2020c).

3 The Proposed Framework

3.1 BERT-CRF model for Chinese word Segmentation

Chinese word segmentation is generally viewed as a character-based sequence labeling task. Specifically, given the sentence $X = \{x_1, x_2, ...x_T\}$, each character in the sequence is labeled as one of $\mathcal{L} = \{B, M, E, S\}$, indicating the location of the character as at the beginning, middle, or end of a word, or that the character is a single-character word. CWS aims to determine the ground truth of labels $Y^* = \{y_1^*, y_2^*, \dots y_T^*\}$:

$$Y^* = arg\ max\ P\left(Y|X\right)_{Y \in \mathcal{L}^T} \quad (1)$$

The universal end-to-end neural CWS architecture usually contains an encoder and a decoder. The framework used in this study is shown in Figure 1; the functions of each part are explained below.

Encoding layer. According to Fu et al. (2020), although BERT-based (Devlin et al., 2019) models for CWS are imperfect, BERT is superior, in many aspects, to models that have not been pre-trained. For example, BERT is more suitable for dealing with long sentences; therefore, this study utilizes BERT released by Google Devlin et al. (2019) as the shared encoder, which is pre-trained with a large amount of unlabeled Chinese data.

$$\{\mathbf{h}_1...\mathbf{h}_i...\mathbf{h}_T\} = Encoder(\{x_1...x_i...x_T\}) \quad (2)$$

where $\mathbf{h}_i$ is the representation for x_i from the encoder.

Decoding layer. This study is able to use a shared decoder for samples from different eras because era-aware representation have been combined for each character by the Switch-memory module. There are various algorithms that can be implemented as decoders, such as conditional random fields (CRF) (Lafferty et al., 2001) and softmax. According to (Tian et al., 2020c), CRF performs better in word segmentation tasks. Therefore, considering the framework of this study, CRF is used as the decoder.

In the CRF layer, $P(Y|X)$ in Eq. 1 can be represented as:

$$P(Y|X) = \frac{\emptyset(Y|X)}{\sum_{Y' \in \mathcal{L}^T} \emptyset(Y'|X)} \quad (3)$$

where, $\emptyset(Y|X)$ is the potential function, and only interactions between two successive labels are considered.

$$\emptyset(Y|X) = \prod_{i=2}^{T} \sigma(X, i, y_{i-1}, y_i) \quad (4)$$

$$\sigma(\mathbf{x}, i, y', y) = exp(s(X, i)_y + b_{y'y}) \quad (5)$$

where $b_{y'y} \in \mathbf{R}$ is trainable parameters respective to label pair (y', y). The score function $s(X, i) \in \mathbb{R}^{|\mathcal{L}|}$ calculate the score of each lable for i_{th} character:

$$s(X, i) = \mathbf{W}_s^\top \mathbf{a}_i + b_s \quad (6)$$

where $\mathbf{a}_i$ is the final representation for i_{th} character. $\mathbf{W}_s \in \mathbb{R}^{d_a \times L}$ and $b_s \in \mathbb{R}^{|\mathcal{L}|}$ are trainable parameters.

3.2 Switch-memory mechanism

The Switch-memory consists of d memory cells and a switcher. For an input sentence, there are d memory cells for each character. The switcher governs how much information in each cell will be integrated into the network. And the state of the switcher depends on the sentence classifier.

3.2.1 Memory cells

The dictionary has been a useful external source to improve the performance of CWS in many studies.(Yang et al., 2017; Liu et al., 2018; Zhang et al., 2018). However, the ability to incorporate the dictionary into previous research has been limited by either concatenating candidate words and character embeddings or the requirement of handcrafted templates. In this study, key-value memory networks are utilized to incorporate dictionary information, which is initially applied to the Question Answering (QA) task for improved storage of prior knowledge required by QA. Furthermore, this network structure can also be used to store the existing knowledge that is required by cross-era CWS.

Ancient Chinese is not a static language but is instead a diachronic language. Ancient Chinese has three development stages: Ancient, Middle Ancient, and Near Ancient. Each stage has a specific lexicon and word segmentation granularity. Therefore, this research has constructed four dictionaries $\mathcal{D} = \{D_0, D_1, D_2, D_3\}$, associating with the four development stages of Chinese, respectively, and each dictionary is era-related. When input a sentence, four memory cells are generated for each character in the sentence according to the four dictionaries, and each memory cell maps candidate words and word boundary information to keys and values.

Rule	$V_{i,j}$
x_i is the beginning character of $w_{i,j}$.	V_B
x_i is the ending character of $w_{i,j}$.	V_E
x_i is a single word, $w_{i,j}$.	V_S

Table 2: the rules for assigning different values to x_i according to its position in word $w_{i,j}$.

Candidate words as keys. Following Tian et al., for each x_i in the input sentence, each dictionary has many words containing x_i, we only keep the n-grams from the input sentence and appear in each dictionary, resulting $w_i^d = \{w_{i,1}^d, w_{i,2}^d, ...w_{i,j}^d...w_{i,m_i}^d\}$, x_i is a part of word $w_{i,j}^d \in D_d$, $d \in [0, 3]$. We use an example to illustrate our idea. For the input sentence show in Figure 1, there are many n-grams containing x_3 = "海(sea)", we only retain ones that appear in D_0 for the first memory cell, thus, w_3^0 = {"海口(HaiKou)", "入海口(estuary)", "海(sea)"}. Similarly, we can generate w_3^1, w_3^2, w_3^3 for the second, third and fourth memory cell according to D_1, D_2, D_3. Then, the memory cell compute the probability for each key (which are denoted as $e_{i,j}^w$ for each $w_{i,j}^d$), here $\mathbf{h}_i$ is the embedding for x_i, which is encoded by the encoder.

$$p_{i,j}^d = \frac{exp(\mathbf{h}_i \cdot e_{i,j}^w)}{\sum_{j=1}^{m_i} exp(\mathbf{h}_i \cdot e_{i,j}^w)} \quad (7)$$

Word boundary information as values. As we know, CWS aims to find the best segment position. However, each character x_i may have different positions in each $w_{i,j}^d$. For example, x_i may be at the beginning, middle, ending of $w_{i,j}^d$, or x_i maybe a single word. Different positions convey different information. Therefore, we use the boundary information of candidate words as values for key-value networks. As shown in Table 2, a set of word boundary values $\{V_B, V_E, V_S\}$ with embeddings $\{e_{V_B}, e_{V_E}, e_{V_S}\}$ represent the x_i's different positions in $w_{i,j}^d$, and we map x_i

to different value vectors according to its positions. As a result, each w_i^d for x_i has a values list $\mathcal{V}_i^d = [v_{i,1}^d, v_{i,2}^d, ...v_{i,j}^d, ...v_{i,m_i}^d]$. In Figure 1, x_3 = "海(sea)", for the first memory cell, we can map candidate word boundary information to the value list $\mathcal{V}_3^0 = [V_S, V_B]$. Four cells for x_i has a values list $\mathcal{V}_i = [v_i^0, v_i^1, v_i^2, v_i^3]$. Then the d_{th} memory cell embedding for x_i is computed from the weighted sum of all keys and values as follow.

$$\mathbf{o}_i^d = \sum_{j=1}^{m_i} p_{i,j}^d e_{i,j}^{v^d} \tag{8}$$

where $e_{i,j}^{v^d}$ is the embedding for $v_{i,j}^d$. Next, the final character embedding is the element-wise sum of $\mathbf{o}_i$ and $\mathbf{h}_i$, or their concatenation, passing through a fully connected layer as follow:

$$\mathbf{a}_i = \mathbf{W}_o \cdot (\mathbf{o}_i \odot \mathbf{h}_i) \tag{9}$$

where $\odot$ operation could be sum or concatenate, $\mathbf{W}_o \in \mathbb{R}^T$ is a trainable parameter and the output $\mathbf{a}_i \in \mathbb{R}^T$ is the final representation for the i_{th} character. $\mathbf{o}_i$ is the final memory embedding for the i_{th} character, and can be calculated as follow.

$$\mathbf{o}_i = Switcher([\mathbf{o}_i^0, \mathbf{o}_i^1, \mathbf{o}_i^2, \mathbf{o}_i^3]) \tag{10}$$

where *Switcher* is used to control how much information in each memory cell will be combined with the output of the encoder.

3.2.2 Switcher

Inspired by the benefits of multi-task, a classifier has been added on top of the encoder to predict the era label of the input sentence. The discriminator predicts the probability of the correct era label, z, conditioned on the hidden states of the encoder, $\mathbf{H}$, which is the output of "[CLS]" from BERT. The loss function of the discriminator is $\mathcal{J}_{disc} = -logP(z|\mathbf{H})$, through minimizing the negative cross-entropy loss to maximizes $P(z|\mathbf{H})$.

In this study, $\mathbf{H}$ is fed into a fully-connected layer and let it pass through a softmax layer to obtain probabilities for each era label.

Switch mode. For the switcher, we propose two switcher modes, *hard-switcher* and *soft-switcher*. *Hard-switcher* switches memory cells according to the predicted final result from the discriminator. For the input sentence in Figure 1, if the predicted result is the modern era, then the switcher will switch to the memory cell associated with modern Chinese, and $\mathbf{o}_i = \mathbf{o}_i^d$. *Soft-switcher* switches memory cells according to the predicted probability, which is calculated by the weight of each memory cell. *Soft-switcher* means that the information from all four dictionaries may be fused into the current character's representation. For example, the predicted probability list is $[0.1, 0.2, 0.1, 0.6]$; therefore, the final memory representation for the i_{th} character is $\mathbf{o}_i = \mathbf{o}_i^0 * 0.1 + \mathbf{o}_i^1 * 0.2 + \mathbf{o}_i^2 * 0.1 + \mathbf{o}_i^3 * 0.6$.

3.2.3 Objective

In this framework, the discriminator is optimized jointly with the CWS task, which both share the same encoding layer. Different weights are assigned to the loss of the two tasks, the final loss function is:

$$\mathcal{J} = \alpha \mathcal{J}_{cws} + (1 - \alpha)\, \mathcal{J}_{disc} \tag{11}$$

where α is the weight that controls the interaction of the two losses. $\mathcal{J}_{cws}$ is the negative log likelihood of true labels on the training set.

$$\mathcal{J}_{cws} = -\sum_{n=1}^{N} log(P(Y_n|X_n)) \tag{12}$$

where N is the number of samples in the training set, and Y_n is the ground truth tag sequence of the n_{th} sample.

4 Experiment

4.1 Datasets

The model proposed in this study has been evaluated on four CWS datasets from Academia Sinica Ancient Chinese Corpus[2] (ASACC) and SIGHAN 2005 (Emerson, 2005). The statistics of all the datasets are listed in Table 3. Among these datasets, PKIWI, DKIWI, AKIWI from ASACC, correspond to near ancient Chinese, middle ancient Chinese, ancient Chinese, respectively, and MSR from SIGHAN 2005 is a modern Chinese CWS dataset. It should be noted that PKIWI, DKIWI, and AKIWI are traditional Chinese and were translated into simplified Chinese prior to segmentation.

For PKIWI, DKIWI, and AKIWI, 5K examples were randomly picked as a test set; then, 10% of examples were randomly selected from training set as the development set. Similar to previous work (Chen et al., 2017), all datasets are pre-processed by

[2] http://lingcorpus.iis.sinica.edu.tw/ancient

Datasets			Words	Chars	Word types	Char Types	Sents	OOV Rate
ASACC	AKIWI	Train	2.8M	3.2M	65.3K	7.5K	59.7K	-
		Test	0.2M	0.3M	15.7K	4.4K	5K	4.35%
	DKIWI	Train	2.2M	2.8M	44.3K	6.0K	50.1K	-
		Test	0.2M	0.3M	13.0K	3.8K	5K	4.91%
	PKIWI	Train	6.4M	7.8M	117.0K	7.2K	144.1K	-
		Test	0.2M	0.3M	18.6K	4.4K	5K	1.71%
SIGHAN05	MSR	Train	2.4M	4.1M	88.1K	5.2K	86.9K	-
		Test	0.1M	0.2M	12.9K	2.8K	4.0K	2.60%

Table 3: Detail of the four datasets.

replacing Latin characters, digits, and punctuation with a unique token.

In the cross-era learning scenarios, all of the training data from four eras corpora were used as the training set. Then, all of the test data from four corpora were used as the cross-era test set to evaluate the model. Finally, F1 and OOV recall rates (R_{oov}) were computed according to the different eras.

4.2 Experimental configurations

In our experiments, for the encoder BERT, we follow the default setting of the BERT (Devlin et al., 2019). The key embedding size and value embedding size are the same as the output of the encoder, and they have been randomly initialized. For the baseline model Bi-LSTM, the character embedding size is set to 300, and the hidden state is set to 100. For the transformer, the same settings as Qiu et al. (2020) were followed. The loss weight coefficient α is a hyper-parameter that balances classification loss and segmentation loss; the model achieves the best performance when α is set to 0.7, which was identified by searching from 0 to 1 with the equal interval set to 0.1.

The words from the training set are used as the internal dictionary, and each training set generates its own dictionary. The simplified Chinese dictionary sourced from jieba [3] is used as the external dictionary for MSR, and words from *The ErYa* (an ancient dictionary) and ancient Chinese textbooks were extracted as the external dictionary for AWIKI. For PWIKI and DWIKI, high-frequency bi-grams and tri-grams were extracted from the corresponding period corpus [4] as external dictionaries.

[3] github.com/fxsjy/jieba/tree/master/jieba/dict.txt

[4] http://core.xueheng.net/

4.3 Overall results

To begin, in this section, the experimental results of the proposed model on the test sets from the four cross-era CWS datasets are provided, which can be seen in Table 4.

Several observations can be made from the data provided in Table 4. First, BERT-CRF in single-era scenarios (ID:1 in Table 4) and cross-era learning without the SM module (ID:6) are compared. As can be seen in the table, when mixing four datasets, the average F1 value of all datasets decreases slightly. Single-era dataset learning has an average F1 value of 97.61, while cross-era learning without the Switch-memory module has a 97.32 average F1 value. This indicates that performance cannot be improved by merely mixing several datasets.

Second, the models with the SM mechanism (ID:3,5,7) outperformed the baseline models (ID:2,4,6) in terms of F1 value and R_{oov} on all datasets. For example, the average F1 score for BERT-CRF with SM module (ID:7) improved by 0.92% when compared to BERT-CRF (ID:6), and the average R_{oov} went from 76.15 to 82.37. This indicates that the Switch-memory can help improve segmentation and R_{oov} performance by integrating era-specific knowledge.

Third, among different encoders, the improvement of the pre-trained encoder BERT on the F1 value is still significant. When using Bi-LSTM as the encoder (ID:2,3), the average F1 value and the R_{oov} are 89.15 and 90.66, respectively. When using BERT as the encoder (ID:6,7), the F1 value improves by approximately 8%. The reason for this may be that the pre-training processing supplements some effective external knowledge.

To further illustrate the validity and effectiveness of this model, the best results from this study are compared to works that have been previously

NO.	En-De		AWIKI	PWIKI	DWIKI	MSR	Avg.
Single-era learning							
1	BT-CRF	F	97.62	97.58	97.19	98.03	97.61
		R_{oov}	68.85	76.58	74.80	**86.85**	76.77
Cross-era learning							
2	BL-CRF	F	89.78	85.98	87.04	93.81	89.15
		R_{oov}	45.55	46.43	37.51	58.06	46.89
3	BL-CRF+SM	F	90.66	87.41	89.18	95.42	90.66
		R_{oov}	43.48	44.40	32.78	68.74	47.35
4	TR-CRF	F	95.89	95.43	95.87	92.68	94.97
		R_{oov}	57.87	58.01	47.07	72.24	58.80
5	TR-CRF+SM	F	96.69	97.04	96.87	96.71	96.82
		R_{oov}	64.22	57.23	50.42	71.34	60.80
6	BT-CRF	F	97.04	97.51	96.96	97.75	97.32
		R_{oov}	68.78	75.39	73.94	86.48	76.15
7	CROSSWISE	F	**98.46**	**98.04**	**98.42**	98.04	**98.24**
		R_{oov}	**83.88**	**81.86**	**77.25**	86.50	**82.37**

Table 4: Experimental results of the proposed model on the tests of four CWS datasets with different configurations. "+SM" indicates that the model uses the Switch-memory module. There are two blocks. The first block is results of the baseline model (BERT-CRF) on the single-era dataset. The second block consists of the results of cross-era learning model with different encoders ("BL" for Bi-LSTM, "TR" for Transformer, "BT" for BERT, "CROSSWISE" for BERT-CRF+SM). Here, F, R_{oov} represent the F1 value and OOV recall rate respectively. The maximum F1 values are highlighted for each dataset.

Models	AWIKI		PWIKI		DWIKI		MSR	
	F	R_{oov}	F	R_{oov}	F	R_{oov}	F	R_{oov}
Chen et al. (2017)	-	-	-	-	-	-	96.04	71.60
Gong et al. (2019)	-	-	-	-	-	-	97.78	64.20
Luo et al. (2019b)	91.25	56.32	97.01	48.09	97.00	43.18	97.09	75.19
Ye et al. (2019b)	-	-	-	-	-	-	98.40	84.87
Qiu et al. (2020)	96.44	65.06	95.83	63.75	96.31	57.03	98.05	78.92
Huang et al. (2020a)	98.16	78.97	97.70	75.69	98.12	74.28	98.29	81.75
Tian et al. (2020c)	-	-	-	-	-	-	98.28	**86.67**
CROSSWISE	**98.46**	**83.88**	**98.04**	**81.86**	**98.42**	**77.25**	**98.04**	86.50

Table 5: Performance (F1 value) comparison between CROSSWISE and previous state-of-the-art models on the test sets of four datasets.

identified as state-of-the-art. Various aspects of multi-domain and multi-criteria Chinese word segmentation are very similar to the tasks in this study; therefore, this study reproduced experiments on several previous word segmentation models using the four datasets identified in this research (Luo et al., 2019b; Qiu et al., 2020; Huang et al., 2020a). For the multi-domain segmenter PKUSeg (Luo et al., 2019b), four datasets were trained with the pre-trained mixed model. The comparison is shown in Table 5; the model from this study outperforms previous methods.

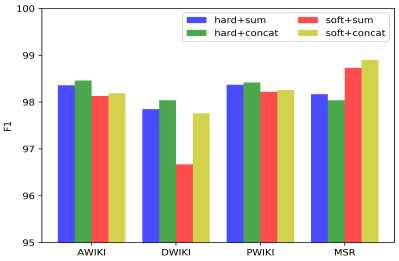

Figure 2: The F1 values of CROSSWISE using four pair settings; "hard+sum" means hard-switcher and the sum of the memory embedding and the character embedding from the encoder as the final character representation.

ID	Switcher	Memory	AWIKI		DWIKI		PWIKI		MSR	
			F	R_{oov}	F	R_{oov}	F	R_{oov}	F	R_{oov}
1	✓	×	98.00	80.62	97.87	80.69	97.52	74.69	98.01	86.48
2	×	✓	98.28	76.58	97.85	74.80	98.32	74.85	98.63	**86.85**
3	✓	✓	**98.46**	**83.88**	**98.04**	**81.86**	**98.42**	**77.25**	**98.04**	86.50

Table 6: Ablation experiments.

Sample from AWIKI (Ancient Chinese): 故上化下，犹风之靡草也。 (Therefore, the superior civilizes and the subordinate, like the winds swept the grass)	
Golds	故/上/之/化/下/，/犹/风/之/靡/草/也/。
	So/superior/zhi/enlighten/subordinate/,/like/wind/zhi/swept/the/grass/.
w/o SM	故/上/之/化/下/，/犹/风/之/靡草/也/。
Ours	故/上/之/化/下/，/犹/风/之/靡/草/也/。
Sample from MSR (Modern Chinese): 天津市“鱼与熊掌兼得”的实践也就分外值得人们重视。 (Tianjin's practice of "getting both the fish and the paw" deserves special attention.)	
Golds	天津市/“/鱼/与/熊掌/兼/得/”/的/实践/也/就/分外/值得/人们/重视/。
	Tianjin/"/fish/and/bear's paw/both/get/"/of/practice/also/then/extraordinary/ worth/people/important/.
w/o SM	天津市/“/鱼与熊掌/兼/得/”/的/实践/也/就/分外/值得/人们/重视/。
Ours	天津市/“/鱼/与/熊掌/兼/得/”/的/实践/也/就/分外/值得/人们/重视/。
Sample from MSR (Modern Chinese): 从大乱走向大治，中经雍正承前启后。 (From chaos to prosperity, through Yongzheng connects the past and the future.)	
Golds	从/大/乱/走/向/大/治/，/中/经/雍正/承前启后/。
	From/big/chaos/go/to/big/prosperity/,/middle/through/Yongzheng/connect/.
w/o SM	从/大/乱/走/向/大/治/，/中经/雍正/承前启后/。
Ours	从/大/乱/走/向/大/治/，/中/经/雍正/承前启后/。

Table 7: Segmentation cases from the test sets of MSR, AWKI and DWIKI datasets.

4.4 Ablation study

Table 6 shows the effectiveness of each component in the SM module.

The first ablation study is conducted to verify the effectiveness of memory cells. In this experiment, the sentence classification task is no longer a switcher but simply a joint training task with word segmentation. We can see that the ancient Chinese datasets (AWIKI, DWIKI, PWIKI) are more sensitive to memory cells than MSR. This may be explained by the fact that the encoder is pre-trained with a large quantity of modern Chinese data, and the memory cells in this study incorporate some ancient era knowledge into the model, which helps to boost the performance of the three ancient Chinese datasets.

The second ablation study is to evaluate the effect of the switcher. For this experiment, the average of four embedded memory cells is used as the final memory representation. The comparison between the second and the third line indicates that the switcher is an important component when integrating era-specific information.

In summary, in terms of average performance, the switcher and the memory cells can both boost the performance of R_{oov} considerably.

4.5 Mode selection

In this section, the effect of the switcher mode and the combination mode (concatenate or sum) of memory embedding and character embedding is investigated.

To better understand the effect of the different configurations, this study examines the four pair settings to train the model on the four datasets in this study; the results are shown in Figure 2, and different color bars represent different datasets. As can be seen, *soft-switcher* significantly improves the F1 value on MSR compared to *hard-switcher*, while the other three datasets prefer *hard-switcher*, which suggests that the forward direction of knowledge dissemination from ancient Chinese to mod-

ern Chinese can help modern Chinese word segmentation, and that the reverse knowledge dissemination will have a negative impact on ancient Chinese word segmentation. Concatenating memory embedding and character embedding from the encoder outperforms the combination of the two; therefore, this study chose the pair of configurations, "hard +concat", to obtain the experimental results in the last row of Table 4 and Table 5.

4.6 Case study

This study further explores the benefits of the SM mechanism by comparing some cases from BERT-CRF and CROSSWISE. Table 7 lists three examples from the test sets of Ancient Chinese and modern Chinese datasets. According to the results, in the first sentence, "靡(swept)" and "草(grass)" are two words in ancient Chinese, BERT-CRF treats these two words as a single word; BERT-CRF gives the second sentence the wrong boundary prediction in "中(middle)" and "经(through)." However, this study's CROSSWISE achieves all exact segmentation of these instances. The third sample is a sentence written in both ancient and modern Chinese,"鱼与熊掌兼得," which is a famous classical sentence in ancient Chinese. CROSSWISE also can split the sentence correctly. Therefore, it can be concluded that the model is flexible for Chinese word segmentation of era-hybrid texts and can produce different segmentation granularity of words according to the era of the sentence. Concurrently, it shows that the SM mechanism is effective in integrating era-specific linguistic knowledge according to different samples.

5 Conclusion

In this study, a flexible model, called CROSSWISE, for cross-era Chinese word segmentation is proposed. This model is capable of improving the performance of each dataset by fully integrating era-specific knowledge. Experiments on four corpora show the effectiveness of this model. In the future, the incorporation of other labeling tasks into CROSSWISE, such as POS tagging and named entity recognition, may prove to be insightful.

Acknowledgments

This research is supported by the NSFC project "the Construction of the Knowledge Graph for the History of Chinese Confucianism" (Grant No. 72010107003). We would like to thank Professor Jun Wang and Hao Yang for their insightful discussion.

Ethical Considerations

The datasets used in this paper are open datasets and do not involve any ethical issues.

References

Andrea Ceroni, Nam Khanh Tran, Nattiya Kanhabua, and Claudia Niederée. 2014. Bridging temporal context gaps using time-aware re-contextualization. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, page 1127–1130. ACM.

Ernie Chang, Yow-Ting Shiue, Hui-Syuan Yeh, and Vera Demberg. 2021. Time-aware Ancient Chinese text translation and inference. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 1–6, Online. Association for Computational Linguistics.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long short-term memory neural networks for Chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206, Lisbon, Portugal. Association for Computational Linguistics.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1193–1203. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. Rethink cws: Is chinese word segmentation a solved task? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5676–5686, Online. Association for Computational Linguistics.

Jingjing Gong, Xinchi Chen, Tao Gui, and Xipeng Qiu. 2019. Switch-lstms for multi-criteria chinese word segmentation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6457–6464. AAAI Press.

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. Association for Computational Linguistics.

Kaiyu Huang, Degen Huang, Zhuang Liu, and Fengran Mo. 2020a. A joint multiple criteria model in transfer learning for cross-domain chinese word segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 3873–3882. Association for Computational Linguistics.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020b. Towards fast and accurate neural Chinese word segmentation with multi-criteria learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2062–2072, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1204–1214. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Junxin Liu, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2018. Neural chinese word segmentation with dictionary knowledge. In *Natural Language Processing and Chinese Computing*, pages 80–91, Cham. Springer International Publishing.

Yang Liu, Yuanhe Tian, Tsung-Hui Chang, Song Wu, Xiang Wan, and Yan Song. 2021. Exploring word segmentation and medical concept recognition for Chinese medical texts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 213–220, Online. Association for Computational Linguistics.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019a. Pkuseg: A toolkit for multi-domain chinese word segmentation.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Xuancheng Ren, and Xu Sun. 2019b. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Xipeng Qiu, Hengzhi Pei, Hang Yan, and Xuanjing Huang. 2020. A concise model for multi-criteria chinese word segmentation with transformer encoder. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 2887–2897. Association for Computational Linguistics.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based joint segmentation and pos tagging for chinese using bidirectional rnn-crf.

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. Association for Computational Linguistics.

Yuanhe Tian, Yan Song, and Fei Xia. 2020b. Joint Chinese word segmentation and part-of-speech tagging via multi-channel attention of character n-grams. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2073–2084, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020c. Improving Chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online. Association for Computational Linguistics.

Jingjing Xu and Xu Sun. 2016. Dependency-based gated recursive neural network for Chinese word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–572, Berlin, Germany. Association for Computational Linguistics.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver,*

Canada, July 30 - August 4, Volume 1: Long Papers, pages 839–849. Association for Computational Linguistics.

Yuxiao Ye, Weikang Li, Yue Zhang, Likun Qiu, and Jian Sun. 2019a. 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, Volume 1 (Long and Short Papers)*, pages 2726–2735, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuxiao Ye, Weikang Li, Yue Zhang, Likun Qiu, and Jian Sun. 2019b. 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, Volume 1 (Long and Short Papers)*, pages 2726–2735, Minneapolis, Minnesota. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 421–431, Berlin, Germany. Association for Computational Linguistics.

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.

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. Association for Computational Linguistics.