Integration of Named Entity Recognition and Sentence Segmentation on Ancient Chinese based on Siku-BERT

Sijia Ge
University of Colorado-Boulder
Sijia.Ge@colorado.edu

Abstract

Sentence segmentation and named entity recognition are two significant tasks in ancient Chinese processing since punctuation and named entity information are important for further research on ancient classics. These two are sequence labeling tasks in essence so we can tag the labels of these two tasks for each token simultaneously. Our work is to evaluate whether such a unified way would be better than tagging the label of each task separately with a BERT-based model. The paper adopts a BERT-based model that was pre-trained on ancient Chinese text to conduct experiments on *Zuozhuan* text. The results show there is no difference between these two tagging approaches without concerning the type of entities and punctuation. The ablation experiments show that the punctuation token in the text is useful for NER tasks, and finer tagging sets such as differentiating the tokens that locate at the end of an entity and those are in the middle of an entity could offer a useful feature for NER while impact negatively sentences segmentation with unified tagging.

1 Introduction

The Chinese classics is the invaluable legacy for both public and academia. The study of ancient classic texts involves various disciplines such as religion, arts, literature, politics, etc. The public can learn about the diverse aspects of ancient China and enhance their knowledge of history; the ancient Chinese texts provide evidence for research areas like the diachronic evolution of the Chinese characters and so on. Besides, handling languages like ancient Chinese is also significant to multi-language processing as the grammar, vocabulary and other linguistic categories are strongly different from those in other languages, including modern Chinese. For example, “之” (zhi) was used to express the possessive in ancient Chinese while “的” (de) replaces it to become the major word to express possessive in modern Chinese.

Two related downstream tasks are sentence segmentation and named entity recognition since most ancient texts were not punctuated originally and the entity information would be the foundation for further research, only if we know the names of the official positions we could know the hierarchy of the administrative systems in ancient China.

For the above two tasks, we can consider both of them as sequence labeling tasks (i.e. token classification tasks), which means the model would classify each token in the input sequence to a predesigned category. For the NER task, the model would identify which tokens would be components of a named entity, and identify the relative position to the entity. In other words, it would determine the span of the entity from the input sequence. In the sentence “**As of now, Twitter is still a publicly-traded company on the New York Stock Exchange.**”, there are two named entities: “**Twitter**”, we can tag it as “S” to represent this is an entity consists one single token, but an entity can also consist of several tokens, like “**New York Stock Exchange**”. We can tag “New” in “New York Stock Exchange” as “B”, which signifies that “New” is part of an entity and is the beginning of that entity; for sentence segmentation, we can tag the token which is followed by punctuation as a special tag and tag the rest as the other label. So in the previous instance, “now” and “exchange” can be tagged as special tokens to represent there is punctuation followed by these tokens.

Considering these two tasks as token classification makes it possible to combine them and tag each token for the two tasks simultaneously, which is efficient. The advancement of deep learning especially the development of BERT-based models improved the performance of token classification tasks (Devlin et al., 2018). Some scholars have tried pre-trained the BERT-based model on raw ancient Chinese texts and fine-tuned them on the specific downstream tasks such as word segmen-

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 167–173
November 20, 2022. ©2022 Association for Computational Linguistics

tation, part-of-speech tagging, and so on, which verified the feasibility of applying the BERT-based model to ancient Chinese datasets (YU Jingsong, 2019; Hu et al., 2021; Chang et al., 2021). However, most current works just focus on one specific sub-task resulting in a lack of comparison between separate tagging and unified tagging.

In our project, we want to answer the question: whether unified tagging on ancient Chinese texts would be better than separate tagging in terms of NER and sentence segmentation when training with a BERT-based model. we adopt the siku-BERT model as our pre-trained model which was pre-trained on the unlabeled ancient Chinese raw text and then fine-tuned on the *Zuozhuan*, comparing the performance of the separated tagging scheme and the integrated tagging scheme.

2 Related Work

Sentence segmentation on ancient Chinese: Tang et al. (2021) applied the incremental training approach to the pre-trained model and got an improvement of 1.83% and 2.21% respectively compared to the model without incremental training on sentence segmentation and punctuation tasks. Hu et al. (2021) developed a BERT+CNN model to perform sentence segmentation on poems, lyrics, and prose text, which was 10% higher than the Bi-GRU model on all of these three text styles. An LSTM-CRF model was employed (Xu et al., 2019), with the assistance of a radical embedding, the performance of this model improved compared to the typical LSTM-based model in sentence segmentation, and the result from the epitaph text of the Tang dynasty arrived at a F-1 score of 81.34%.

Named entity recognition for ancient Chinese: Wu et al. (2015) developed a deep neural network to generate word embeddings and conducted a named entity recognition task. The results showed that this model performed better than the state-of-the-art CRF model, arriving at the highest F-1 score of 92.80%. A Bi-LSTM-CRF model was proposed and applied to the traditional Chinese medicine patents' named entity recognition problems (Deng et al., 2021). The paper verified that context semantic information can be learned without feature engineering, and the performance was better than the baseline methods, arriving at an F-1 score of 94.48%. Chang et al. (2021) applied a BERT-Bi-LSTM/IDCNN-CRF model to the NER task and performed better than the Bi-LSTM-CRF benchmark model, which was 4.79% higher in F-1 score when CLUENER dataset was applied. A radical-level-based Bi-LSTM-CRF model with a self-attention mechanism was employed (Yin et al., 2019), solving the problem of how to deal with hidden information due to the properties of Chinese characters, and arriving at a F-1 score of 93.00% in CCKS_2017 dataset and 86.34% in TP_CNER dataset.

Our work is different from previous literature as we integrate the two sequence labeling tasks into one by merging labels and comparing the performance of the integrated tagging approach with the performance of the separated tagging approach.

Unified char-based tagging for ancient Chinese: YU Jiangde (2015) developed a Max-Entropy model with a unified character-based label set to perform word segmentation, part-of-speech tagging, and named entity recognition tasks. The experimental results showed that training with the unified char-based label set would be better than training in three separate turns. Cheng et al. (2020) developed a Bi-LSTM-CRF model to conduct word segmentation, part-of-speech tagging, and sentence segmentation tasks with a unified char-based label set, which also proved that labels integration with mixed corpus would get better performance. Qi et al. (2021) constructed a model unifying the word segmentation and part-of-speech (POS) tagging tasks and got the F-1 scores of 95.98% in the word segmentation task and 88.97% in the POS tagging task.

Compared to these works, although we also adopt the unified char-based tagging, our work adopts the BERT-based model to perform on the NER and sentence segmentation tasks, which is expected to extract the features better and get better performance.

3 Methods

Model: we adopt the Siku-BERT (Wang et al., 2021) as the pre-trained model, it was trained on SiKuQaunShu (a collection of ancient classics, including 536,097,588 tokens, all characters were written in traditional Chinese). Such a huge corpus will cover most ancient Chinese characters so that it would be sufficient for training. The pre-trained model architecture is based on the Chinese BERT-base model. Besides the BERT model, we add one CRF (Conditional Random Field) layer to get the global optimized label sequence.

Hardware: all experiments are trained on a Tesla V80 GPU.

Pre-processing on tagging scheme: as mentioned in the introduction part, we convert the NER task and sentence segmentation into sequence labeling tasks. To do that, we adopt the char-based labeling (Ng and Low, 2004) as the annotation scheme. That means, for the NER task, we would annotate each token (character) with one label, to signify whether this character is a part of an entity, and if so, what is the position of the token concerning the entity. We use "B" (beginning), "I" (internal), "S" (single token as an entity), and "O" (outside) to represent the position of the current character concerning an entity.

For example, "Twitter" and "New York Stock Exchange" are two named entities in the sentence "**As of now, Twitter is still a publicly-traded company on the New York Stock Exchange.**" Since "New York Stock Exchange" is an entity, we would tag the sequence as "B I I I", the "B" corresponds to "New" and mirrors that it is the beginning of an entity, and "I" means the corresponding token inside an entity, "Twitter" is a single token entity, so it is tagged as "S". Other characters that are not a part of an entity are tagged as "O" (outside of the entity).

An ancient Chinese example is

九O 月O 晉B 惠I 公I 卒O 懷B 公I 立O

(In September, Jin Hui Gong died and Huai Gong inherited the throne.)

"晉惠公"（Jin Hui Gong）is a person's name, so it is tagged as B I I .

Since our purpose for the sentence segmentation task is exactly to find out the position where punctuation occurs, we wipe out all tokens that are punctuation in our experiments, otherwise, the model just needs to tag the special label for the token that is followed by punctuation to get a good result, but this is too easy for this task. We tag the character which is followed by punctuation as "P" and those are not with "L", so like the above sentence:

As of now, Twitter is still a publicly-traded company on the New York Stock Exchange .

We will tag "now" and "exchange" as "P" and others as "L" since only these two precede punctuation.

For our ancient Chinese example:

九月, 晉惠公卒, 懷公立.

We tag the sequence as followings:

九L 月P 晉L 惠L 公L 卒P 懷L 公L 立P

That means "月""卒""立" are followed by punctuation. If there are two continuous punctuation in the original raw text, we just tag "P" once for the token before this punctuation since we just need to segment the sentence instead of recovering each punctuation.

Besides tagging these two tasks separately, we also merge the labels of each token for these two tasks into one so that we can train two tasks in one experiment. If we join the label for NER and sentence segmentation tasks with "-", the above example would be like

九O-L 月O-P 晉B-L 惠I-L 公I-L 卒O-P 懷B-L 公I-L 立O-P

So "晉 B-L" signifies that "晉" locates at the beginning of an entity but this character is not followed by punctuation.

The purpose is to compare the performance of these two tagging approaches (one is tagging for specific tasks separately and the other is tagging for both two tasks by merging the label set) with the Siku-BERT model.

Training data: Our training data is *Zuozhuan*, a chronicle of general history records during the Spring and Autumn (770-476 BC) and Warring States (475-221) periods in China. The number of tokens in the training data is 244,345, and including 25,005 entities, 33,775 punctuation. The split ratio for the training and validation sets is 8 : 2.

Hyperparameters setting: Our main purpose is not the absolute score of each task but the relative difference between the two tagging approaches, thus we keep all settings the same across experiments.

We set up 10 epochs for the training in total and the batch size for each epoch is 8, the learning rate is 0.001, and the scheduler step is 600. We adopt cross entropy as our loss function and AdamW as the optimizer; the drop-out rate is 0.2.

4 Experimental Design

4.1 The separate training for NER and sentence segmentation

First, we conduct experiments for NER and sentence segmentation separately to compare the result with the training through the unified tagging approach. Noticed that in our experiments we ignore the type of the entities (person names, place names, etc) and the type of the punctuation (colon, comma, period, etc) to make the experiment simpler while it would be worthwhile to explore with different types of entities and punctuation on a finer granularity level.

The evaluation metric for both the NER task and the sentence segmentation task would be the F-1 score mainly since the distribution of the frequency of the labels is imbalanced, but for the NER task, we count the number of samples that the model hits based on the entity level rather than based on the token level. To make it clearer, for the above examples, if the correct label sequence is

New B York I Stock I Exchange I

And the model predicts it as :

New B York I Stock O Exchange B

We don't count this as one correct prediction even if half tokens are labeled correctly.

The evaluation metric for the task of sentence segmentation is the F-1 score as well, for this task, we count the correct samples just based on the token so that we count the number of "P" and "L" labels in the gold labels and the model predictions.

4.2 Integration training for NER and sentence segmentation

To compare the effect of the integrated tagging scheme on both tasks, we train with the integrated label tagging on these two tasks together but evaluate each task separately as in 4.1. All settings are the same as 4.1 except for the tagging approach.

It's impossible to feed all texts as one into the model and thus we have to segment samples. The length of a sample cannot be the original sentence length that is segmented by punctuation, because if we do so, the model would know that only if it tags a punctuation marker "P" for the token at the end of each sample (that is a segmentation point), then the model can perform well. We don't allow the model to "cheat" in this way, so we segment each sample into a fixed length of 128, and apply it to both tasks and both tagging approaches for the sake of fairness.

Besides the major experiments, we add two groups of ablation experiments to explore the role of other factors. One of the two main factors is the pre-process approach for the NER task specifically, in this factor we can divide into two sub-factors, one is whether preserve the punctuation on the data, and the other is whether segment the sample based on a fixed number or the original sentence length, we conduct such experiments only for the NER task side without impacting the sentence segmentation; the second main factor is the label set for NER, besides "BIOS", we can also use "BIOES" to tag the entity. The only difference is that we use "E" (end) and "I" (internal) to differentiate whether the current character is the end of an entity or just inside an entity but not at the end, this increases the number of the classes. We compare the two label sets for both the NER and sentence segmentation tasks.

metric	sep tagging	uni tagging
test_accuracy	0.934	0.907
test_f1	0.869	0.867
test_precision	0.881	0.901
test_recall	0.894	0.86

Table 1: The result of NER task on separate tagging and unified tagging

5 Results

5.1 The NER task results on two tagging approaches

We conduct the NER task first and compute the F-1 score of tagging NER only and that of unified tagging. The result in table 1 ("sep tagging" means the data only includes the NER label, "uni tagging" means the data includes both NER and sentence segmentation tags) illustrates that there is no big difference between the two tagging approaches and both get an F-1 score of 0.87, while the separate tagging is a little bit better than the unified tagging one in terms of accuracy. It seems that the model does not improve with the additional sentence segmentation tags in the unified tagging approach.

5.2 The sentence segmentation task results on two tagging approaches

The result of the sentence segmentation task is the same as the NER task, with the F-1 score on both tagging schemes arriving at 0.97. The reason why the performance on the sentence segmentation task is so high is that we evaluate it on a token level, which would be high since much more non-punctuation markers than punctuation markers. Such a task is a binary token classification, for each token, the model has a 50% chance to hit the correct tag. While for the NER task the model only has a 25% probability to tag the correct token, the result would be even lower when evaluating the performance on the entity level.

5.3 The impact of punctuation on NER task

There is no difference in terms of the NER task when training with separate tagging and unified tagging, it seems that the label of punctuation does not offer any clue for the named entity identification. While we find out that the punctuation token benefits the NER task from the ablation experiment. What we do is conducting two more NER experiments on the same data but pre-process the data in different ways. One is changing the length of the samples, there is no punctuation in the data, but keep the original length of each sentence as the length of each sample, which means each sample is segmented by the original punctuation but removes all punctuation tokens; the other one not only keeps the original sentence length for each sample but also preserves punctuation. Comparing the performance of the former one and the initial experiment can infer the impact of the length of the sample on the NER task and comparing the performance of these two additional experiments can get the impact of the punctuation token on the performance.

The result is shown in table 2, "N" means there are no punctuation tokens in the sample while "Y" means preserving punctuation tokens, so "N+128 length" refers to the initial experiment setting. We can observe that the first one and the second one are almost the same, which shows that the length of the sample would not impact the NER result, while the third one is better than the first two, which shows the role of punctuation tokens for the NER task. It makes sense since there are some frequent structures such as the place names are followed by a period and the person names often occur before a colon when quoting the sentence from the speaker.

We also wonder why does the unified tagging not work better from our results in 5.1 as it at least offers the punctuation information to some extent by the label set. One possible explanation is that in our experiment, the model cannot learn the relationships between the entity labels and the different types of punctuation, in other words, whether a token follow by punctuation makes less sense than what type of punctuation follows it, consequently, it doesn't work if the entity information is related to a type of punctuation specifically.

5.4 The impact of granularity of label sets

Another ablation experiment is comparing the impact of the granularity in terms of different label sets for NER. Besides the "BIOS" label set, the other popular label set to tag the named entity is "BIOES". Compared to the "BIOS" tag set, it has one "E" label to represent the end of the entity. For "New York Stock Exchange", it would be tagged as "B I I I" with "BIOS" label set but tagged as "B I I E" with "BIOES" label set. Such a change makes the model to further figure out whether the current token is the end of the entity or just inside the entity. We reproduce the experiments in the same way as our initial experiments except for the NER tagging labels. The result is shown as table 3.

From the table, we can conclude that the performance improves a bit for both tagging approaches with the "BIOES" label set compared to the "BIOS" tagging set, and the separate tagging approach is slightly better than the unified tagging approach, which is opposite to the previous work (e.g. Cheng et al., 2020). The finer label set is a double-edged sword. On the one hand, finer labels offer more information and features for the models; on the other hand, more labels make it harder for the model to make a correct prediction. The model doesn't need to differentiate the tokens locate at the end of entities and ones locate at the middle of entities with "BIOS" labels.

For the sentence segmentation task, the performance on the separate tagging does not change while the performance decreases obviously on the integrated tagging approach, decreasing from 0.97 to 0.915, which is surprising for us since the previous work reported it improved by 3.5% compared to separated tagging (Cheng et al., 2020).

6 Conclusions

In our project, we adopt a BERT-based model to evaluate the performance of the named entity recognition task and the sentence segmentation task on ancient Chinese text with a unified tagging approach and a separate tagging approach respectively. We find out that there is no difference when we take different tagging strategies, both strategies get an F-1 score of 0.87 on the NER task and 0.97 on the sentence segmentation task, which poses a challenge to the conclusion that unified tagging is always better concluding from the previous works; we also conclude that punctuation marker is im-

process method	N + 128 length	N + original length	Y+ original length
test_f1	0.867	0.869	0.881

Table 2: The result of NER task on different pre-process approaches

tasktagging approach	separate tagging	unified tagging
NER	0.891	0.882
sentence segmentation	0.97	0.915

Table 3: The F1 score compared with two tagging schemes on BIOES label

portant for the NER task from our ablation experiments, training on the data with punctuation would be better on the NER task; moreover, finer tagging set like "BIOES" is better than "BIOS" for the NER task for both separated and unified tagging approaches, but performs worse on the sentence segmentation task if apply unified tagging.

Our experiment shows an inconsistent result to the previous research that also compared the unified tagging scheme with the separate tagging approach (YU Jiangde, 2015; Cheng et al., 2020; Shi et al., 2010); the further exploration we need to do is tagging the types of the entity and punctuation as well and count the performance based on different types of entities or punctuation, in addition, we want to try different model architectures and different pre-trained models, we want to verify whether what we observe from our current experiments is specific for the model and architecture.

Our work pays attention to the ancient Chinese data which is low-resource and being ignored. The shortage of annotated data and the relatively fewer application scenarios make it a minority field in NLP, which is required more research. The usage of advanced deep learning techniques for automatic sentence segmentation and named entity recognition of ancient Chinese not only facilitate readers to read, but also can be of great significance to the arrangement of ancient books, and the intelligent application of ancient Chinese.

References

Yuan Chang, Lei Kong, Kejia Jia, and Qinglei Meng. 2021. Chinese named entity recognition method based on bert. In *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, pages 294–299. IEEE.

Ning Cheng, Bin Li, Liming Xiao, Changwei Xu, Sijia Ge, Xingyue Hao, and Minxuan Feng. 2020. Integration of automatic sentence segmentation and lexical analysis of ancient chinese based on bilstm-crf model. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 52–58.

Na Deng, Hao Fu, and Xu Chen. 2021. Named entity recognition of traditional chinese medicine patents based on bilstm-crf. *Wireless Communications and Mobile Computing*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Renfen Hu, Shen Li, and Yuchen Zhu. 2021. Knowledge representation and sentence segmentation of ancient chinese based on deep language models. *Journal of Chinese Information Processing*, 35(4):8–15.

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 *EMNLP*, pages 277–284.

Zhang Qi, Jiang Chuan, Ji Youshu, Feng Minxuan, Li Bin, Xu Chao, and Liu Liu. 2021. Unified model for word segmentation and pos tagging of multi-domain pre-qin literature. *Data Analysis and Knowledge Discovery*, 5(3):2–11.

Min Shi, Bin Li, and Xiaohe Chen. 2010. Crf based research on a unified approach to word segmentation and pos tagging for pre-qin chinese. *Journal of Chinese Information Processing*, 2(24):39–45.

Xuemei Tang, Qi Su, Jun Wang, Yuhang Chen, and Hao Yang. 2021. (automatic traditional Ancient Chinese texts segmentation and punctuation based on pre-training language model). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 678–688, Huhhot, China. Chinese Information Processing Society of China.

Dongbo Wang, Chang Liu, Zihe Zhu, Jiang, Feng, Haotian Hu, Si Shen, and Bin Li. 2021. Construction and application of pre-training model of "siku quanshu" oriented to digital humanities.

Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624.

Han Xu, Wang Hongsu, Zhang Sanqian, Fu Qunchao, and Liu Jun. 2019. Sentence segmentation for classical chinese based on lstm with radical embedding. *The Journal of China Universities of Posts and Telecommunications*, 26(02):1–8.

Mingwang Yin, Chengjie Mou, Kaineng Xiong, and Jiangtao Ren. 2019. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. *J. of Biomedical Informatics*, 98(C).

YU Zhengtao YU Jiangde, HU Shunyi. 2015. A unified character-based tagging approach to chinese lexical analysis. *Journal of Chinese Information Processing*, 29(6):1.

ZHANG Yongwei YU Jingsong, WEI Yi. 2019. Automatic ancient chinese texts segmentation based on bert. 33(11):57.