Unlocking Transitional Chinese: Word Segmentation in Modern Historical Texts

Baptiste Blouin[a], **Hen-hsen Huang**[b], **Christian Henriot**[a], **Cécile Armand**[a]
[a] IrAsia, Aix-Marseille University
first.last@univ-amu.fr
[n] Institute of Information Science, Academia Sinica
hhhuang@iis.sinica.edu.tw

Abstract

This research addresses Natural Language Processing (NLP) tokenization challenges for transitional Chinese, which lacks adequate digital resources. The project used a collection of articles from the *Shenbao*, a newspaper from this period, as their study base. They designed models tailored to transitional Chinese, with goals like historical information extraction, large-scale textual analysis, and creating new datasets for computational linguists. The team manually tokenized historical articles to understand the language's linguistic patterns, syntactic structures, and lexical variations. They developed a custom model tailored to their dataset after evaluating various word segmentation tools. They also studied the impact of using pre-trained language models on historical data. The results showed that using language models aligned with the source languages resulted in superior performance. They assert that transitional Chinese they are processing is more related to ancient Chinese than contemporary Chinese, necessitating the training of language models specifically on their data. The study's outcome is a model that achieves a performance of over 83% and an F-score that is 35% higher than using existing tokenization tools, signifying a substantial improvement. The availability of this new annotated dataset paves the way for refining the model's performance in processing this type of data.

Introduction

Previous studies of the Chinese language based on NLP methods have focused on either modern Chinese — today's Chinese for which there exists a wealth of digital resources such as Wikipedia, Baidu, etc. — or on classical or ancient Chinese based mostly on collections of literary, religious, or medical texts. In this paper, we address the issue of word segmentation for the Chinese language that emerged and developed between the end of the 19th century and the early 1950s. We shall label this language "transitional Chinese".

Despite the huge amount of publications that appeared in China during that century in the form of newspapers, periodicals, encyclopaedias, books, reports, etc., almost no work has been done to address the challenges that transitional Chinese raises for the implementation of NLP methodologies. The major reason is the absence of available digital resources. Although the major libraries and private companies in China have digitized a great number of newspapers or periodicals, the access to the digital versions, when they exist, has been limited to a web interface, most of the time with severe restrictions on downloading the text files or even on copying sections of the text. These databases are designed for text display, consultation, and reading, not for text analysis through computational methods.

Chinese companies and institutions as a rule refrain from providing the text files that would allow researchers to fully implement NLP methodologies. Despite repeated attempts to negotiate TDM rights (Text and data mining) with several mainland companies, we hit a wall when it came to obtaining the text files, even under the strictest terms of confidentiality and corpus protection. To put it bluntly, there is an abyssal gap between the extraordinary effort made at digitizing historical sources in China, especially the vast reservoir of periodicals and newspapers at the Beijing National Library or the Shanghai Library, and the possibility to use these resources to advance research.

The ENP-China project was designed from its inception with NLP methodologies as a key component of its methodology to process historical sources. The press, in particular, was considered as crucial not just because it could provide rich historical information, but also because this was the very

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 92–101

place where the modern Chinese language developed. We were able to acquire the entire collection of the *Shenbao* in full-text files thanks to a private provider based in Taiwan. With this treasure trove, we were in a unique position to finally address the challenge of implementing NLP methodologies and algorithms trained on modern texts on a vast corpus of pre-1949 transitional Chinese as found in the press.

Our work represents the first attempt to break through the limits of existing models for Chinese and to develop models adapted to transitional Chinese through campaigns of annotations to create training sets with verifiable data. A major challenge was to design a model that proved robust across the various genres of texts found in the *Shenbao* and across the whole period under study (1872-1949). The central objective of our experiments and developments was to adjust models and to tailor them to this evolving Chinese — from administrative Classical Chinese to modern common and literary Chinese — with multiple purposes:

- to extract as completely and as accurately as possible the relevant historical information on our research topics
- to enable textual, discourse analysis at a scale heretofore unreachable
- to deliver new datasets for corpus and computational linguists

Newspapers are a most relevant source for analyzing long-term patterns of linguistic and conceptual changes (Hengchen et al., 2021). The newspaper that we used as a core resource represents a huge collection of more than 2.2 million articles published between March 1872 and May 1949. The *Shenbao* was the first daily newspaper published in Chinese in Shanghai. Originally a local publication, it became at once a national newspaper read throughout the empire. It also set the matrix for the subsequent newspapers that appeared at the turn of the century. For almost thirty years, the *Shenbao* set the tone, the pace and the model of news-writing, thereby creating a language in itself, the same language found in later publications (Mittler, 2004).

When the Shenbao was established in 1872 in Shanghai, there was no previous history of mass print media in China. The Chinese state and its local agencies produced ”official gazettes” that printed and circulated official decrees, edicts and lists of appointed officials. These texts were written according to the conventions of administrative classical Chinese which, despite some evolution, remained basically the same for the few centuries before the advent of the modern press. The establishment of privately-run newspapers, however, raised language issues:

- the newspapers needed to reach a wider educated audience than just the officials and literati and make the language of news reporting more accessible.
- the newspapers reported on a much wider range of issues that touched on new topics and notions, especially when it came to international news.
- the newspapers developed *sui generis* from classical Chinese a new language through successive shifts, both in vocabulary, grammatical structure, use of punctuation, layout, and typography.

The classical Chinese language differs considerably from modern Chinese. It consists mostly in single-character words. There are of course exceptions, which include the name of institutions, titles, proper names, etc. Yet, in most classical Chinese texts a character by and large equates a word. This feature generates a phenomenon of polysemy of characters that only the context in which they appeared and the ingrained knowledge acquired by the literati in the major genres (Confucian classics, poetry, memorials, etc.) of literary writing could disambiguate.

Newspapers faced several challenges in adapting the classical language to the constraints of news reporting.

- First, they had to introduce a large range of new expressions to cover in enough precise terms all the concepts and objects that came to China through its interaction with the outside world. Japan, that had been exposed earlier on to concepts imported from the West, became a major supplier of character-based neologisms (Wang, 1998; Chen, 2014). It consisted mostly in two-character expressions that eventually became the norm in modern Chinese. Whereas classical Chinese was quite strictly monosyllabic, modern Chinese became mostly disyllabic.

- Second, the grammatical structure of classical Chinese changed seamlessly in various ways, including a change of some of the basic elements such as pronouns, demonstratives, verbs, etc., while punctuation was introduced incrementally and sometimes a bit haphazardly (Mullaney, 2017). This process of change did not follow any guidelines. The Chinese language reinvented itself under the collective movement and innovation of the literati. Several initiatives by the state sought to define rules for the creation of a modern language, which probably had a certain impact, but it was not until 1922 that a new national language was officially adopted, to be taught throughout the entire school system (Kaske, 2008). Nevertheless, previous writing habits persisted until 1949, which make the newspapers of the Republican era a kaleidoscope of Chinese writing styles and languages.

- Third, the early newspapers included in their pages various genres of texts written in very different styles, from extracts from official gazettes, to literary texts, including poetry, to translations from fiction or telegrams, to news reporting and advertisements. Although some sections lost in importance with time (official gazettes), newspapers generally formed a mosaic of writing styles. The *Shenbao* presented such a kaleidoscope of overlapping genres (Mittler, 2004).

- Fourth, the period from 1872 to 1949 is one of progressive but constant language change. Even in 1872, except for the excerpts from the *Beijing Gazette* (京報) written in administrative classical Chinese, the language used in the *Shenbao* for news reporting was already a different language from the start. Through a data-driven analysis of the content of the *Shenbao*, Magistry has identified six main periods based on clustering. It is consistent with previous studies that defined 1904, 1911 and 1937 as clear-cut shifts. Magistry's study, however, points to other shifts around 1890-1892 and another one in 1922 (Magistry, 2021).

Word segmentation constitutes a prerequisite for many tasks of textual analysis such as topic modeling, word frequencies, semantic network analysis, etc. Whereas the task of word segmentation for languages based on Latin characters has become almost trivial, it remains a challenge for Chinese.

Whether in modern or in classical Chinese, characters are aligned vertically or horizontally next to one another. In the case of classical Chinese, all the way to the 1920s one also faces the near or complete absence of punctuation (Galambos, 2021). Although modern punctation was introduced in the last decade of the 19th century, its usage remained very uneven and unstable until the early 1930s (Hamm, 2021). While punctuation does not separate tokens per se, the presence of punctuation introduces a significant element of sentence segmentation that helps for tokenization. Modern Chinese has received a lot of attention, with a constant flow of papers in the major conferences on various methods to produce accurate word segmentation. Although less rich, works have also focused on classical (and even ancient) Chinese (Chen and Tai, 2009; Huang and Wu, 2018; Han et al., 2018). While both strands of research provide models, pretrained resources, and conceptual frameworks, neither procure readily replicable and usable models for transitional Chinese. The main challenge in the ENP-China project therefore has been to design a robust model that can adapt to the various stages of language development in transitional Chinese. In the next sections, we present the context of the study by reviewing previous appraoches to Chinese word segmentation and available datasets for training models. We introduce our new dataset, the experiments we have made with existing models and the model that we have trained.

1 Chinese word segmentation

Chinese word segmentation (CWS) has been a subject of prolonged debate within the field of NLP. For specific tasks and datasets, adhering to the character level proves to be a more suitable approach, yielding higher-performance results and simplifying the process. Thus, in certain scenarios, it is deemed less essential to perform word segmentation. For instance, a study by (Li et al., 2019) strongly asserts that with the advent of neural methods in NLP, CWS is gradually becoming an irrelevant or even detrimental step in the NLP pipeline.

In the context of computational methods in the humanities, the task of addressing CWS continues to be an area of significant interest. In fact, this crucial task enables multiple analyses of Chinese text, empowering researchers and practitioners to extract valuable insights and knowledge.

To accomplish this, it becomes essential to have a tokenizer that is specifically tailored and adapted to the data being processed. The effectiveness of the tokenizer greatly impacts the accuracy and efficiency of subsequent NLP tasks on Chinese texts.

Over time, CWS has been extensively studied, leading to the development of sophisticated systems capable of achieving near-perfect results. Some of these systems boast an impressive F-score close to 100%, indicating a high level of precision and recall in identifying word boundaries within Chinese texts. These remarkable achievements in segmentation accuracy further enhance the overall performance of NLP applications when working with Chinese language data.

During the early stages of CWS, the predominant methods employed were lexicon- and rule-based approaches, such as forward maximal matching, reverse maximal matching, and least word cut. While these techniques were relatively straightforward to implement, they suffered from low accuracy in segmenting words effectively.

To address this challenge, from a machine learning perspective, CWS started to be approached as a sequence labeling task. This method involved predicting whether each input character should be separated from its neighboring characters (Xue, 2003). In the 2000s, many researchers turned to conditional random fields or maximum entropy Markov models to tackle this task.

As the field progressed, approaches utilizing supervised or unsupervised features emerged (Zhao and Kit, 2008), contributing to the state-of-the-art performance in CWS. These techniques made significant strides in improving segmentation accuracy and efficiency.

A major turning point came around 2013 when researchers began incorporating neural networks into their work, revolutionizing the research landscape for CWS. Whether adopting feed-forward, recurrent, or convolutional neural network architectures, these approaches offered substantial benefits, particularly in reducing the need for labor-intensive data engineering tasks (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015).

Despite the promise of neural network-based methods, early trials did not consistently outperform non-neural systems. Refining these neural approaches became necessary to achieve their full potential in CWS tasks. However, over time, researchers made remarkable progress, and neural network-based methods eventually surpassed the capabilities of their non-neural counterparts, leading to significant advancements in the field.

In the realm of NLP, the emergence of transformers marked another significant milestone. Recent advancements in the field utilized a customized Transformer model for sequence tagging, resulting in the achievement of state-of-the-art performance (Duan and Zhao, 2020; Chou et al., 2023).

However, a common issue with many NLP models is that they are often trained and evaluated on contemporary datasets, limiting their applicability to historical or ancient texts. CWS is a field with a long history in Chinese NLP, resulting in the creation and continual evaluation of numerous datasets, like the ones listed below.

- SIGHAN Bakeoff 2005 (Emerson, 2005): It comprises diverse types of Chinese text, including news articles, fiction, and academic papers from various sources, such as CKIP (Academia Sinica) and City University of Hong Kong for traditional Chinese, and Beijing University and Microsoft Research (China) for simplified Chinese.
- Chinese Penn Treebank (CTB) [1]: It is one of the most widely used datasets for CWS research, and it exists in three versions: CTB6, CTB7, and CTB9.
- PKU Corpus (Yu et al., 2018): This dataset was collected from the People's Daily website and contains various text types, including news articles and editorials.
- Universal Dependency (UD) [2] for Mandarin Chinese: Similar to many other widely studied languages in NLP, Mandarin Chinese also has a Universal Dependency Annotation Scheme, which provides a standardized framework for dependency-based syntactic analysis.

Recently, to address the need for models capable of handling ancient Chinese texts, the EvaHan 2022 dataset [3] was created. It consists of annotated ancient Chinese text developed as part of the EvaHan project, a collaborative effort by several Chinese universities to build resources for ancient Chinese NLP. The Ancient Chinese language dates back to around 1000BC-221BC.

[1] https://www.cs.brandeis.edu/ clp/ctb/
[2] https://universaldependencies.org/treebanks/zh_pud
[3] https://circse.github.io/LT4HALA/2022/EvaHan

Thanks to the impressive performance of the models trained on these contemporary data, many off-the-shelf solutions are widely adopted in the context of digital humanities research, facilitating the exploration and analysis of Chinese texts from various time periods and genres:

- LAC (Jiao et al., 2018) is a joint lexical analysis tool developed by Baidu's NLP Department, which realizes the functions of Chinese lexical segmentation, lexical annotation, and proper name recognition.
- Jieba [4] is a module that is specifically used for CWS.
- Stanza (Qi et al., 2020) : This library was created by the Stanford NLP Group. It contains different tools for linguistic analysis such as POS tagging, lemmatization, segmentation and handles 66 languages including Chinese.
- SnowNLP [5] is a class library written in python inspired by TextBlob. It was created specifically to process Chinese.
- THULAC (Maosong Sun et al., 2016) is a set of Chinese lexical analysis toolkit developed and launched by the Laboratory of NLP and Social Humanities Computing of Tsinghua University, with the functions of Chinese lexical segmentation and lexical annotation.

These tools were designed with a strong emphasis on achieving high performance and user-friendliness. Nevertheless, it remains crucial to evaluate their performance when applied to specific datasets and Chinese language variants.

2 Dataset creation

Within the scope of our project, our primary objective was to evaluate various tokenizers on our specific dataset. The aim was twofold: first, to estimate the performance of off-the-shelf tokenizer tools, and second, to address the possibility that these readily available solutions might not yield suitable results for our unique data. In such a scenario, we planned to develop a custom model tailored to the requirements of our dataset.

In order to delve deeper into the development of the language during the particular period under examination, we engaged in manual tokenization of sentences extracted from the *Shenbao*. This involved carefully segmenting the text into individual words. We proceeded in two steps: for the first round, we aimed at annotating a large select of texts published between the years 1872 and 1947. Three annotators were trained on a sample of the selected corpus, after which they annotated separately 741 articles. For the second round, we selected 52 articles published between the years 1872 and 1907 that were annotated, curated, and used for evaluating our model.

By tokenizing these historical articles, we sought to gain relevant insights into the linguistic patterns, syntactic structures, and lexical variations present in the Chinese language during that specific historical timeframe. This approach allowed us to address the challenges posed by the absence of modern linguistic resources and the particularities that might arise in historical Chinese texts.

Through this comprehensive evaluation and analysis of word segmentation methods on our well-curated dataset, we anticipated obtaining a clearer picture of the tokenizer's efficacy in handling historical Chinese texts. Ultimately, this process played a vital role in shaping our subsequent decisions regarding the selection of the most appropriate tokenizer or the development of a custom model best suited to our research objectives and historical data.

To achieve this goal, we engaged two groups of three distinct annotators who were tasked with tokenizing the documents based on our guidelines. Each annotator was asked to add basic punctuation to mark the end of sentences. A blank space was introduced between tokens when their part of speech was different. Only place names, job titles, proper names of persons were left unspeparated.

Additionally, for the second round, a separate individual — a historian with high skills with classical and modern Chinese — was entrusted with curating the annotated data. This systematic approach ensured that the word segmentation process was carried out consistently and according to the specified guidelines, while also guaranteeing the quality and accuracy of the curated dataset.

Upon the successful completion of the annotation and curation processes, we obtained a carefully curated dataset, which we will call ENP-TOK [6], documented in Table 1. The agreement between annotators of the training set, calculated as the F1-score, is 75%, and the agreement between annotators of the evaluation set is 78%. These results

[4] https://github.com/fxsjy/jieba
[5] https://github.com/isnowfy/snownlp
[6] https://gitlab.com/enpchina/enp-tok

highlight the difficulty of this task during this period.

	Train	Eval
# Articles	741	52
# Sentences	11 132	396
# Characters	867 474	36 707
# Words	350 631	15 498

Table 1: ENP-TOK dataset statistics

ENP-TOK serves as a valuable resource for evaluating the performance and quality of various tokenizers, providing crucial observations for our research.

3 Experiments

By using this new annotated dataset, we have effectively assessed the performance of off-the-shelf tokenizers, particularly when applied to texts from the period of interest.

We conducted a thorough evaluation of the five most commonly used word segmentation tools. The results of this evaluation are summarized and presented in Table 2, providing measured references on the efficacy and suitability of each tokenizer for our specific historical texts.

Model	Precision	Recall	F-score
Jieba	42,94 %	53,61 %	47,61 %
Stanza	49.87 %	52.35 %	51.09 %
LAC	59.17 %	**69.09 %**	**63.74 %**
SnowNLP	**63.04 %**	52.45 %	57.26 %
thulac	47,24 %	61,42 %	53,41 %

Table 2: Off-the-shelf tools results on ENP-TOK dataset

In comparison to the performance achieved on contemporary datasets, the results obtained from these off-the-shelf models are significantly lower. It is imperative to verify and validate these models before employing them for larger-scale analyses to ensure the reliability of their outcomes. Although these off-the-shelf models offer convenience and quick implementation, we need to note that they may not represent the cutting-edge in terms of performance at the present moment.

To explore the potential for achieving better results without the need for additional annotation, we sought to evaluate more advanced and up-to-date models available in the HanLP library. HanLP is a NLP toolkit designed for production environments, focusing primarily on Chinese language processing. It provides pre-trained models specifically tailored for various datasets used in CWS evaluations.

Through the utilization of HanLP's (He and Choi, 2021) top-performing models for each dataset, we aimed to showcase the benefits of employing more recent and computationally heavier models, while emphasizing the crucial role of selecting the appropriate training set.

The outcomes of this experiment, detailing the performance of HanLP's best models, are documented in Table 3. These results highlight the potential advances that can be achieved in CWS using contemporary, state-of-the-art models without the need for additional manual annotations.

Based on these experiments, HanLP's models demonstrate significantly improved performance compared to off-the-shelf models in most cases. However, it is essential to acknowledge that the disparity in language models utilized—such as Electra, Roberta, and Bert—along with their distinctive learning regimes and training data, makes it challenging to predict which models would yield superior results on different types of datasets.

Model	Dataset	F-score
HanLP	SIGHAN2005	78.06 %
HanLP	CTB6	61.99 %
HanLP	PKU	63.68 %
HanLP	MSRA	76.91 %
HanLP	UD	68.94 %
FastHan	Multi	74.98 %

Table 3: Models results on ENP-TOK dataset

To address this challenge, FastHan (Geng et al., 2021) emphasizes the paramount importance of Generalization as a crucial attribute for any successful NLP toolkit. To achieve this, they developed a model trained on a diverse range of corpora specifically for CWS. The resulting model demonstrated improved performance across various sources.

In Table 3, we observe the results of applying this model to our dataset, confirming that it does outperform the majority of off-the-shelf tools. However, it falls slightly short compared to some models trained on more specific corpora.

Based on the findings, it becomes evident that employing corpora that align more closely with our specific requirements is the key to obtaining consistent and reliable results. Customizing the training data to suit the unique characteristics of our target

text allows us to achieve optimal performance, ensuring that the model is better suited to handle the specificities of historical Chinese texts.

Given the lack of annotated data aligned with our specific period, it seemed wise to use at least annotated data from traditional Chinese or ancient Chinese.

4 Model training

To validate this hypothesis, we conducted extensive model training using three distinct datasets: CKIP, EvaHan, and ENP-TOK. Our aim was to assess the significance of both the training and inference datasets in shaping model performance. Beyond the influence of the CWS training data, we also delved into the impact of the language models employed to train these models.

All results presented in this section are averaged over 10 random initializations.

Initially, we used the BERT (Devlin et al., 2019) language model for Chinese. Because of its ability to deal with simplified and traditional Chinese (automatically translated), and its popularity.

Dataset	Precision	Recall	F-score
CKIP	71,25 %	78,75 %	74,81 %
EvaHan	78,12 %	78,19 %	78,16 %
ENP-TOK	**82.93 %**	**80.91 %**	**81.91 %**

Table 4: Result obtained on ENP-TOK when training from BERT-base-chinese[7], on several datasets

In view of the results presented in Table 4, the use of a training dataset aligned with our inference dataset yields better results, even if the quantity of data remains smaller than the other dataset. What is more, it seems that the Chinese we deal with is closer to ancient Chinese than to contemporary Chinese. Using data from EvaHan gives better results than CKIP.

Based on the results, presented depending on the quality of the original document in Table 4, using data from older Chinese sources leads to better outcomes. Specifically, utilizing data from EvaHan yields superior results. However, it is worth noting that the training of these models was initially initialized with word representations from contemporary language models. Therefore, to further investigate this, we repeated the previous experiment, but this time, we employed a language model trained on older Chinese data.

[7]https://huggingface.co/bert-base-chinese

The findings suggest that incorporating data from historical Chinese sources can enhance the performance of the models for the task at hand. However, to fully explore the potential benefits, it is essential to consider the impact of using language models pre-trained on historical data (Wang and Ren, 2022), which may provide a more contextually relevant starting point for the training process.

Dataset	Precision	Recall	F-score
CKIP	65,68 %	76,62 %	70,73 %
EvaHan	83,55 %	81,11 %	82,31 %
ENP-TOK	**84.17 %**	**82.04 %**	**83.09 %**

Table 5: Result obtained on ENP-TOK when training from BERT-ancient-Chinese[8], on several datasets

The results, in Table 5, demonstrate that using language models aligned with the languages used in the sources leads to superior performance. These findings assert that the transitional Chinese that we are processing is more closely related to ancient Chinese than contemporary Chinese, which substantiates the need to train language models specifically on our data.

Additionally, it is interesting to examine the results obtained on CKIP. The misalignment between the language model and the annotated data negatively impacts performance when transferring to another dataset. This indicates that using a language model that is not well-aligned with the target data can lead to a decrease in performance when applying the model to different datasets. It highlights the importance of using language models that are tailored to the characteristics of the target data.

Aligning the language model with the linguistic properties of the data being processed, especially for historical languages like ancient Chinese, can significantly improve performance on various NLP tasks. Furthermore, understanding the impact of language model alignment and its effects on transfer learning can help researchers and practitioners make more informed decisions when deploying models across different datasets.

As of now, our efforts have culminated in the development of a model that achieves an impressive performance of over 83% . This achievement means a substantial leap in comparison to using the Jieba tokenizer, as our model reaches an F-score that is +35% higher.

While our results approach the levels attainable

[8]https://huggingface.co/Jihuai/bert-ancient-chinese

on contemporary datasets, it is important to acknowledge that there are several potential avenues for further improvement.

The availability of this new annotated dataset opens up a wide array of possibilities for enhancing and fine-tuning the performance of the model in processing this type of data. We can pursue different paths, including linguistic analysis of the annotated dataset, or optimization at the training stage by incorporating specific language models.

With the aid of linguistic analysis, we can gain valuable insights into the unique characteristics and linguistic patterns present in modern historical Chinese texts. These results can help us refine the model to better capture and handle these nuances, ultimately improving its overall performance.

On the other hand, exploring various language models and implementing the most suitable one can significantly impact the model's ability to handle historical texts effectively. By selecting a specific language models that aligns closely with the linguistic traits of the historical period, we can boost its accuracy and adaptability to the complexities of the data.

This dataset is therefore used for continuous exploration and experimentation, empowering us to refine and optimize the model, or even pave the way for the development of advanced models tailormade for processing historical Chinese texts. This newfound dataset opens exciting possibilities for progress in this specialized domain.

5 Conclusion

Our study aimed to identify the most effective methods for CWS, specifically focusing on historical texts written in transitional Chinese. We created a novel, manually annotated dataset, derived from the *Shenbao*. This dataset provided us with a rich linguistic resource, allowing a comprehensive analysis of word segmentation methods for historical Chinese texts.

The evaluation of five commonly used off-the-shelf tokenizers revealed that while these models offer ease of use and quick implementation, their performance on our historical dataset was significantly lower than on contemporary datasets. This finding underscores the importance of validating models before using them for larger-scale analyses, to ensure the reliability of their outcomes. Further investigation with the HanLP library and the FastHan model demonstrated notable performance improvements, suggesting potential advancements in CWS using contemporary, state-of-the-art models without the need for additional manual annotations. However, the disparity between these models and the language models used for training makes it challenging to predict which models would yield superior results on different datasets.

Our exploration of different datasets and language models for training our segmentation model led us to a critical realization: using high-quality, relevant training data closely aligned with our target texts allowed us to achieve the best performance. This insight emphasized the importance of prioritizing data quality over quantity, and that customization of the training data to suit the unique characteristics of the target texts can lead to more accurate and reliable results.

Ultimately, our efforts culminated in the development of a model that achieved an impressive F-score of over 83% , a significant leap compared to using the Jieba tokenizer. While these results approach the levels attainable on contemporary datasets, there are several avenues for further improvement. The availability of ENP-TOK dataset opens up new possibilities for enhancing the model's performance, ranging from linguistic analysis of the annotated dataset to optimization at the training level by incorporating specific language models.

The improved CWS of transitional Chinese has significant implications for historians conducting research on China. The better the CWS, the more accurate and efficient the text analysis becomes. This improved accuracy can help streamline the research process, allowing historians to accurately segment and analyze large volumes of text, thus opening up new avenues of inquiry and enabling the exploration of previously unmanageable datasets. As most of the texts published in the period under consideration — including periodicals, books, diaries, etc. — were written in a style that closely matches the various forms on which we have trained our model, the tokenizer we propose can serve to unlock vast corpora of historical sources on which existing models fail. This can lead to more nuanced understandings of the wide range of historical texts made available through digitization, offering deeper insights into China's history.

This study illuminates the importance of dataset relevance and quality in achieving optimal results

for CWS, particularly for transitional Chinese texts. The insights and methods derived from this study contribute significantly to the field of historical Chinese text analysis and provide valuable tools for historians working with these rich and complex linguistic resources.

Acknowledgment

- This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476).
- The authors wish to acknowledge fruitful discussions with Pierre Magistry. They also thank the annotators, Chang Yu-jun, Chiang Chia-wei, Hsu Wan-lin, and the curator Dr. Sun Huei-min (Institute of Modern History, Academia Sinica)

References

Bing Chen and Xiaoying Tai. 2009. A Hybrid Approach to Chinese Word Segmentation.

Haijing Chen. 2014. A Study of Japanese Loanwords in Chinese. Master's thesis, University of Oslo, Oslo.

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.

Tzu Hsuan Chou, Chun-Yi Lin, and Hung-Yu Kao. 2023. Advancing Multi-Criteria Chinese Word Segmentation Through Criterion Classification and Denoising. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6460–6476, Toronto, Canada. 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. *Proceedings of the 2019 Conference of the North*, pages 4171–4186.

Sufeng Duan and Hai Zhao. 2020. Attention Is All You Need for Chinese Word Segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3862–3872, Online. 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*.

Imre Galambos. 2021. Premodern Punctuation and Layout. In Jack W. Chen, editor, *Literary Information in China: A History*, pages 125–134. Columbia University Press.

Zhichao Geng, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2021. fasthan: A bert-based multi-task toolkit for chinese nlp. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 99–106.

John Christopher Hamm. 2021. Modern Punctuation and Layout. In Jack W. Chen, editor, *Literary Information in China: A History*, pages 135–142. Columbia University Press.

Xu Han, Hongsu Wang, Sanqian Zhang, Qunchao Fu, and Jun S. Liu. 2018. Sentence segmentation for classical chinese based on LSTM with radical embedding. *CoRR*, abs/1810.03479.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Hengchen, Ruben Ros, Jani Marjanen, and Mikko Tolonen. 2021. A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*, 36(Supplement_2):ii109–ii126.

Shilei Huang and Jiangqin Wu. 2018. A pragmatic approach for classical Chinese word segmentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Zhenyu Jiao, Shuyu Sun, and Ke Sun. 2018. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *ArXiv*.

Elisabeth Kaske. 2008. *The politics of language in Chinese education, 1895-1919*. Sinica Leidensia. Brill, Leiden. OCLC: 171268385.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252, Florence, Italy. Association for Computational Linguistics.

Pierre Magistry. 2021. Le(s)? chinois du Shun-pao .

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: An Efficient Lexical Analyzer for Chinese. Original-date: 2016-05-17T05:05:05Z.

Barbara Mittler. 2004. *A newspaper for China?: power, identity, and change in Shanghai's news media, 1872-1912*. Number 226 in Harvard East Asian studies monographs. Harvard University Asia Center ; Distributed by Harvard University Press, Cambridge (Mass.).

Thomas S. Mullaney. 2017. Quote unquote language reform: New-style punctuation and the horizontalization of chinese. *Modern Chinese Literature and Culture*, 29(2):206–250.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-Margin Tensor Neural Network for Chinese Word Segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303, Baltimore, Maryland. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Binbin Wang. 1998. Gezai Zhongxi Zhijian de Riben—Xiandai Hanyu zhong de Riyu Wailaiy Wenti. ——"". Japan Exists between East and West— the Issue of Japanese Loanwords in Modern Chinese. . No.8. . *Shanghai Wenxue. . Shanghai Literature*, (8):71–80.

Pengyu Wang and Zhichen Ren. 2022. The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168, Marseille, France. European Language Resources Association.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Shiwen (Peking University) Yu, Huiming (Peking University) Duan, and Yunfang (Peking University) Wu. 2018. Corpus of Multi-level Processing for Modern Chinese.

Hai Zhao and Chunyu Kit. 2008. Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

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.