A Dataset for Named Entity Recognition and Entity Linking in Chinese Historical Newspapers

Baptiste Blouin, Cécile Armand, Christian Henriot

A Dataset for Named Entity Recognition and Entity Linking in Chinese Historical Newspapers

Baptiste Blouin, Cécile Armand, Christian Henriot
Aix-Marseille University, IrAsia, Aix-en-Provence, France
{baptiste.blouin, cecile.armand, christian.henriot}@univ-amu.fr

Abstract

In this study, we present a novel historical Chinese dataset for named entity recognition, entity linking, coreference and entity relations. We use data from Chinese newspapers from 1872 to 1949 and multilingual bibliographic resources from the same period. The period and the language are the main strength of the present work, offering a resource which covers different styles and language uses, as well as the largest historical Chinese NER dataset with manual annotations from this transitional period. After detailing the selection and annotation process, we present the very first results that can be obtained from this dataset. Texts and annotations are freely downloadable from the GitHub repository.

Keywords: Named Entity Recognition, Entity Linking, Historical Newspapers, Chinese

1. Introduction

Named-entity recognition (NER) is the Natural Language Processing (NLP) task consisting in identifying and classifying mentions of entities in texts. These mentions belong to a set of predefined categories, among which people, locations, and organizations are the most common.

In this paper, we present a new historical Chinese dataset [1] with named entities belonging to 6 main classes, entity linking from two databases, co-reference and biographical relationship. These annotations were made on two sources of particular interest for digital humanities research, newspapers, and biographies.

Historical newspapers are rich sources of textual information, offering invaluable insights into culture, society, and politics. These records provide a unique vantage point for understanding the transformation of one of the world's oldest civilizations, particularly in the context of China. However, accessing and deciphering content from historical Chinese newspapers presents numerous challenges. This is primarily due to the evolution of the Chinese language, its scriptural nuances, and the inherent complexities of named entities within them. Traditional Named Entity Recognition systems, predominantly designed for contemporary texts, falter when applied to historical content due to variations in language usage, orthography, and the presence of now-archaic terminologies. Moreover, the task is further complicated when these entities need to be linked contextually to their historical significance or to other related entities.

Simultaneously, Latin scripts have seen progress in historical NER and Named Entity Linking (NEL). This is partially facilitated by the abundance of well-preserved records, a continuity in script forms, and an active academic community that leverages these texts for socio-historical analyses. This work has provided significant insights into Latin-based languages and their historical trajectories

Yet, delving into historical Chinese newspapers, especially those from the 19th and 20th centuries, is laden with challenges. Firstly, preservation concerns are paramount. Many newspapers from this period were printed on materials susceptible to deterioration, making them ephemeral in nature. Given the turbulent historical events in China during these two centuries, including the Opium Wars, the fall of the Qing Dynasty, the Republican era, and the rise of the People's Republic, many publications were lost to time, conflict, or were intentionally destroyed. Such geopolitical upheavals and sociocultural revolutions often led to the systematic obliteration of records or negligent preservation.

Secondly, the linguistic landscape of China has undergone significant evolution. The newspapers of the 19th and early 20th centuries witnessed a diverse range of linguistic styles, from Classical Chinese to various vernacular forms, and regional dialects. This multitudinal linguistic representation, while culturally rich, poses considerable challenges in interpretation and understanding, especially when compounded by the evolution of Chinese characters and the eventual simplification movement in the mid-20th century.

Additionally, access to these newspapers is limited due to their rarity and the specialized nature of collections that house them. Many exist only in select archives, private collections, or specialized institutions, often with restricted access.

Within this context, NER in historical Chinese

[1] https://gitlab.com/enpchina/ENP-NER

newspapers becomes even more complex. Contemporary NER systems, optimized for present-day language, are ill-equipped to deal with the aforementioned linguistic diversity and the nuances of historical references.

Recognizing this gap, we present a meticulously curated historian-annotated dataset designed for the identification and contextual understanding of named entities in historical Chinese newspapers. This dataset, which combines the insights of historical experts with linguistic expertise, not only captures named entities but also offers contextual annotations linking them to their historical significance. The introduction of this dataset aims to pave the way for improved machine learning models and enhanced scholarly research in the realm of historical Chinese texts.

We begin this paper by contextualizing our research within the broader landscape of historical NER datasets studies, emphasizing the gaps in existing research pertaining to historical Chinese texts.

Subsequently, we delve into an extensive overview of the sources and databases that have formed the foundation of our study.

We then proceed to elaborate on the guidelines we have established for annotation, meticulously crafted to ensure consistency, precision, and historical contextual relevance.

Following this, we offer an overview of the dataset, elucidating its structure, the diverse range of named entities it encompasses, and the potential it holds for researchers and linguists in their endeavors.

In the final phase of our research, we present our preliminary findings based on the dataset, highlighting its effectiveness in enhancing the accuracy of NER systems tailored for historical Chinese texts. Furthermore, we provide baseline insights that may pave the way for future research directions.

The ultimate goal of this research is to shed light on the untapped potential of historical Chinese newspapers, equipping researchers with more robust tools to facilitate their investigations in this domain.

2. Related Work

Years of intensive digitization endeavors have resulted in an unparalleled quantity of historical documents now accessible in digital formats, complete with text that machines can process. This technological advancement has significantly enhanced the preservation and accessibility of historical records. However, it has concurrently created fresh opportunities for mining the content contained within these digitized archives. Consequently, the foremost challenge at present lies in the development of technology adept at swiftly and effectively searching, retrieving, and navigating the wealth of information contained within this expansive repository of historical data.

In pursuit of this objective, the recognition of named entities assumes a central position. However, it is noteworthy that despite its considerable significance, the majority of research efforts within NLP have primarily focused on contemporary data and text.

While contemporary data presents its own set of challenges and complexities, historical documents introduce a unique dimension to named entity recognition. Historical texts often contain language variations, outdated terminology, and context-specific references that differ significantly from modern language. This poses a distinctive set of challenges for developing and fine-tuning NER systems that are capable of effectively handling historical content.

In light of this, there is a growing need for specialized research and technology development in the field of named entity recognition for historical documents.

Several research initiatives have recognized this necessity and established annotated datasets based on historical sources, aiming to gain a better understanding of how current NLP methods perform. Nevertheless, these endeavors, primarily centered around the development of datasets from newspapers of that era, have been relatively scarce and primarily confined to languages of Latin origin, as seen in the following datasets.

Corpus	Period	Lang.	# NEs
Quaero	19C	fr	147,682
Europeana	19C	fr, de, nl	40,801
Finnish	19C-20C	fi	26,588
Czech Hist	19C	cz	4,017
HIPE	18C-21C	de, en, fr	19,848
NewsEye	19C-20C	de, fr, fi, sv	30,580

Table 1: Historical newspapers NER datasets.

The Quaero Old Press corpus (Rosset et al., 2012) was the first and biggest dataset, which comprises 295 pages from French newspapers dating back to December 1890. It has reasonably good OCR quality and annotators corrected inaccuracies, making it useful for testing NER systems in the presence of OCR errors. The Europeana NER corpus (Neudecker, 2016) is a substantial collection of NE-annotated historical newspaper articles in Dutch, French, and German, primarily from the 19th century. They were selected randomly from the Europeana newspaper collection, with a focus on pages having at least 80% word-level accuracy. While three entity types were considered (person, location, organization). The Finnish NER

Corpus (Kettunen et al., 2016) consists of digitized journals and newspapers published between 1836 and 1918. It features manual OCR correction by librarians and manual or semi-automatic NE annotations. The Czech Historical Corpus (Hubková et al., 2020) is a smaller dataset created from the year 1872 of the Czech title Posel od Čerchova. It encompasses annotations for six entity types and was manually annotated by only two individuals. The HIPE Corpus (Ehrmann et al., 2020), covering approximately 200 years (1798-2018) of historical news in French, German, and English, was sourced from Swiss, Luxembourgish, and American newspapers. OCR quality varies, and the corpus follows Impresso guidelines. The NewsEye Dataset (Hamdi et al., 2021) is a substantial collection of articles from newspapers published between the mid-19th and mid-20th centuries in French, German, Finnish, and Swedish. It covers four entity types and uses guidelines similar to Impresso.

These datasets, summarized in Table 1, are just a part of the broader landscape of annotating historical documents. Various other sources, including literary works, medical journals, travelogues, and more, have undergone entity annotation as well (Ehrmann et al., 2021).

Nonetheless, as of now, there has been a notable absence of annotated Chinese newspapers from this specific historical period using this particular form of annotation. In essence, such annotated resources for this era have yet to be made accessible to the wider public or research community. This gap underscores the need for more comprehensive efforts in this area to enhance our understanding of historical Chinese documents and facilitate broader academic research.

3. Description of the Sources

3.1. Text sources

In our pursuit of conducting thorough annotations, we have taken a deliberate approach in selecting two specific types of sources that hold great significance in the field of historical research. These two sources include a newspaper, providing valuable contemporary insights into events and perspectives of the past, and a multilingual biographical dictionary, a comprehensive reference tool containing information about notable individuals from various backgrounds and regions.

We will provide further information and context about these sources, describing their relevance and importance.

ShenBao, also known as Shun Pao or Shanghai News, was a leading newspaper in Shanghai from 1872 to 1949. Established by British businessman Ernest Major, it stood out as one of China's pioneering modern newspapers, renowned for its independence and reliability, navigating through significant political and social changes in China.

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 1872 and 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).

Using ShenBao for a Named Entity Recognition (NER) dataset offers several advantages:

- Historical Depth: Spanning over seven decades, ShenBao has a lot of information about events, personalities, and societal changes, offering a diverse range of entities for annotation.
- Cultural and Political Significance: Given its role in shaping public opinion during pivotal moments, such as the anti-Japanese sentiment and political shifts, the newspaper can provide contextually rich sentences, essential for high-quality NER datasets.
- Diverse Content: As a major newspaper, ShenBao covered various topics, from politics and economics to culture and society, ensuring a broad spectrum of named entities.
- Language Evolution: The long publication span of ShenBao can help capture the evolution of the Chinese language and terminologies over time, making the dataset comprehensive and linguistically diverse.

In addition, when we compare this source to others within the same volume from the same time period, it is important to note that this particular source did not undergo Optical Character Recognition (OCR) processing. Instead, the textual content of this newspaper was retyped entirely by human effort.

This manual retyping process aimed to reproduce the original text faithfully, resulting in a remarkable degree of fidelity. While there may be occasional human errors in recognizing specific characters, the overall textual content closely mirrors the content of the original source.

This meticulous transcription approach distinguishes this source, ensuring that it retains the nuances and authenticity of the original material. It provides researchers with a valuable resource that closely resembles the historical text, with only minor imperfections stemming from the human element of transcription.

Who's Who is an American publisher of a number of directories containing short biographies. The books are usually entitled Who's Who in… followed by some subject, such as in our case, Who's Who in China. Basic information, including their date and place of birth, their educational background, and the past and current positions they held in various institutions. Since the individuals were still living at the time of publication, such sources did not cover their entire life. Despite their incompleteness, their added value compared to dictionaries lies in the fact that they provide a highly detailed list of positions, often with their exact date, including participation in social clubs and associations. Efficiently extracting this information, including named entities and their contextual relationships, and seamlessly linking these entities across different languages is a matter of paramount significance for historians. It empowers researchers and scholars to gain a comprehensive understanding of historical events, relationships, and networks that transcend linguistic and geographical boundaries.

In our selection process, we randomly selected a total of 181 newspaper articles to construct our dataset. These articles were chosen to span a wide timeframe, ranging from 1872 to 1947. We ensured an even distribution across the years within this specified period, ensuring that our dataset represents a comprehensive view of historical content over time.

Furthermore, we extended our selection to encompass 15 bilingual biographies. Within this subset, we allocated 5 biographies for each of the specific years: 1917, 1925, and 1944, which are our only years with the bilingual version.

3.2. Databases

Wikidata is an open and structured knowledge base containing extensive information about various entities such as people, places, events, and concepts. It can be used to cross-reference historical entities, facilitating a more comprehensive understanding of historical narratives and contexts. Its multilingual support is particularly useful for dealing with historical texts written in different languages or involving figures and events from various regions.

However, it is important to note that while Wikidata strives for accuracy and comprehensiveness, there might be instances where certain historical entities are missing or incomplete. Nevertheless, having access to partial information on Wikidata is often better than having no reference at all. The collaborative nature of Wikidata means that data is continually refined and expanded, potentially filling gaps in historical knowledge over time.

At the time of writing, Wikidata includes over 100 million entries edited by more than 23,000 contributors.

The Modern China Biographical Database (**MCBD** (2021)) constitutes a core initiative to establish a long-term publicly accessible resource for historical research in the China field. The temporal coverage of the database is 1830 to 1949, namely it includes all the individuals born between 1800 and 1930 who were active in China during this period, regardless of their origin, nationality and the duration of their presence in China. This period covers the newspaper period of our study. By using it, we cover entities not present in contemporary databases and also open up the possibility of entity linking to other databases smaller than Wikidata. At the time of writing, MCBD holds data on more than 153,000 individuals, with information on 165,000 positions and 36,000 curricula/degrees, as well as 18,000 institutions and 16,000 companies.

4. Annotation Process

In order to efficiently and comprehensively annotate the 196 documents in our dataset, we organized the annotation process into three distinct groups, each involving three annotators. These groups were responsible for different subsets of the documents based on the type of content to be annotated.

Two of these groups were tasked with annotating the newspapers, with one group handling 101 articles and the other focusing on 80 articles. The third group of three annotators was specifically assigned to annotate the 15 multilingual biographies contained within the dataset.

To ensure consistency and uniformity across the annotations, each of these three groups employed the same annotation tool and strictly adhered to a shared set of guidelines.

Furthermore, to confirm that all annotators had a clear and consistent understanding of the guidelines, a set of five articles outside the primary batch was collectively annotated. This process helped establish a common baseline of understanding

among the annotators, ensuring the quality and reliability of the annotations throughout the dataset.

4.1. Annotation Guidelines

For this annotation campaign, we relied heavily on the Impresso (Ehrmann et al., 2019) guideline, which we adjusted to suit our needs and the requirements of the Chinese language.

The Impresso guideline outlines various named entity types for annotation. For individuals, there are specific categories to denote a single person, a group of people like a musical band, and authors of newspaper articles. Locations are categorized based on administrative divisions such as towns and nations, physical features like geographical and hydrological entities, and other classifications like facilities and addresses. Organizations are divided into administrative bodies and entities that offer products or services. There are also categories for media products, doctrines, and absolute dates. Additionally, certain flags can be applied to these entities, such as ”unresolvable” for ambiguous entities, ”noisy entity” for unclear ones, and ”literal” for those with a direct meaning. The guidelines emphasize using specific subtypes for annotation, and in cases of uncertainty, an ”unknown” subtype can be used, especially for locations in the annotation project.

In comparison with this guideline, we have differentiated between two types of dates: absolute and relative. Absolute dates are dates that appear in the following format: YYYY or YYYY-MM or YYYY-MM-DD. Relative dates are dates that can be calculated in reference to a specific point in time (foundation of the Republic, imperial reigns). The point of reference may or may not be explicit in the text.

In addition, we have also decided to annotate named events. Events refer to named events such as wars, political movements, meetings of organizations, etc.

We also added entity linking, which consists in linking the named entity found in the text with its avatar defined by a unique identifier in our two external knowledge base. This is a crucial task for disambiguating names. A given entity can appear under different names. Conversely, different persons can share the same names. As for the Impresso guideline, entity components and nested entities are excluded from the linking.

We have also added a coreference annotation. Coreference is the task of finding all expressions that refer to the same entity in a text. We link the named entities that are mentioned in a text with all their alternative forms in the same text, without annotating them as named entities.

We have also added an annotation layer reserved for biographies, that of relationships with functions. The objective is to connect entities that are interrelated, such as a position (comp.func) that a person held in an organization (org.) at a given time (time) and place (loc).

Finally, the nature of the Chinese language makes it possible to detach the names of various entities from the entity they refer to. We then decided to annotate these discontinuous entities as a relationship between the various parts.

The complete guideline will be available in the data repository.

4.2. Annotation Tool

We used, INCEpTION (Klie et al., 2018), an interactive platform designed for linguistic annotation. It provides a rich set of features to support the annotation, curation, and management of linguistic data. In INCEpTION, users log in to access their annotation projects from a central dashboard. Once a document is opened for annotation, they can navigate and create various annotations, with the sidebar offering tools like search and statistics. Additionally, there is a curation phase where annotations from multiple annotators are merged using different strategies, and workload management tools help in tracking annotator progress.

When using INCEpTION to annotate named entities based on the Impresso guidelines, the primary goal is to identify and annotate all named mentions within the text. The guidelines provide specific categories, such as persons, organizations, locations, productions, and time, each with its distinct subtypes. For instance, a person can be an individual, a collective, or an article author. Annotators should always opt for the most specific label available, using only subtypes for manual annotation. If there is uncertainty about the exact subtype of an entity, the ”unknown” label can be applied, though this is mainly reserved for locations in the Impresso context.

5. Results

5.1. Curated dataset

Upon the conclusion of the annotation process, our dataset comprises 196 fully annotated documents, collectively featuring a total of 6,491 identified entities distributed across 121,971 characters from the three annotation batches.

Across three distinct batches of data, an average inter-annotator agreement of 73% (measured using Cohen’s kappa[2]) was observed. Specifically, there was a higher level of agreement, reaching 77%, in the annotation of biographies. In contrast, the agreement was slightly lower, at 72%, in the annotation of newspapers. While this score

[2]An agreement refers to a similar annotation across all tasks for a particular entity.

is somewhat below the levels seen in other corpora listed in Table 1, where agreements generally reach or exceed 80%, it remains a notably strong level of agreement.

It is worth highlighting that achieving high inter-annotator agreement can be particularly challenging when dealing with historical Chinese texts from this specific period. The language nuances, variations, and intricacies present in these documents pose unique difficulties for annotators. Therefore, even though the agreement is not as high as some other corpora, it is a commendable accomplishment given the complexities inherent in annotating historical Chinese text from this era.

Finally, it is important to note that obtaining these 6,491 entities involved a curation process. This step was conducted by a historian who meticulously reviewed and refined the annotations across all three distinct batches of data. This curation ensured the accuracy, consistency, and historical relevance of the identified entities, enhancing the overall quality of our dataset for research and analysis.

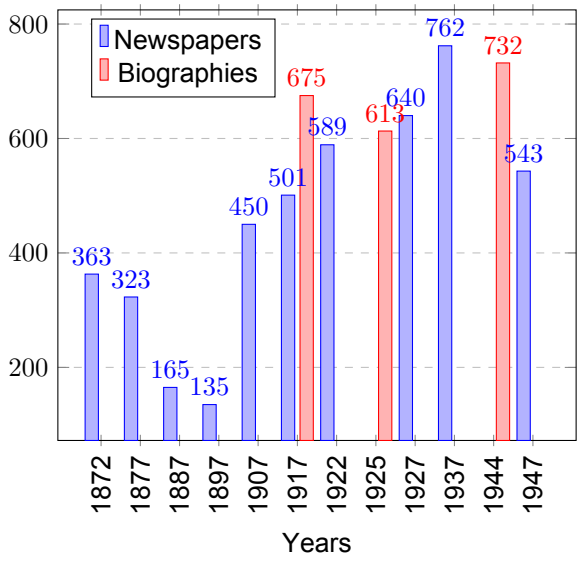

Figure 1: Number of entities annotated by year in our dataset.

Our annotation statistics, outlined in Table 2 for named entities and Table 3 for linking, provide a comprehensive overview of our dataset's composition and the effectiveness of our annotation process. One noteworthy aspect is the well-balanced distribution of entities across the three major classes: location (loc), organization (org), and person (pers). This balance reflects our commitment to inclusivity and ensuring that our dataset covers a diverse range of entity types, enriching its utility for a variety of research purposes.

An interesting observation emerges when we consider the inclusion of the MCDB linking database alongside Wikidata. This addition has proven highly valuable, as it has enabled us to establish a total of 547 links that would have otherwise remained inaccessible if we had relied solely on Wikidata for entity linking. This underscores the significance of leveraging multiple resources to enhance the completeness and richness of our dataset, allowing researchers to access a more extensive network of linked information and facilitating more robust analyses and insights.

Types	Subtypes	Count
event		84
loc		1579
	adm.town	723
	adm.reg	285
	adm.nat	264
	oro	106
	fac	80
	add.phys	50
	phys.hydro	38
	phys.geo	20
	adm.sup	8
	unk	5
org		1524
	ent	514
	busi	340
	adm	263
	edu	193
	asso	185
	ent.pressagency	29
pers		2192
	ind	2153
	coll	39
prod		58
	creation	38
	media	17
	doctr	3
time		1054
	abs.year	328
	abs.day	109
	abs.month	104
	rel.day	204
	rel.year	132
	rel.month	98
	rel.ref	79
all		6491

Table 2: NER statistics of our dataset.

KB	Count	Unique
Wididata	1827	620
MCDB	1421	613
Both	874	285

Table 3: NEL statistics of our dataset. Both: The entity is linked in both databases. Unique : count of unique identifiers

5.2. Evaluation

To initiate an initial assessment of our freshly annotated dataset, we have divided it into two distinct segments. The first segment constitutes the training set, encompassing 90% of the articles, and it is accessible without restrictions. The second segment, which makes up the test set, comprises the remaining 10% of the dataset and the gold annotations will be released at a later time.

To ensure a balanced representation of our dataset over time, we used a random selection approach. Specifically, we randomly sampled 10% of the articles from each year within our corpus. This meticulous process was carried out to guarantee that our test set maintains an even distribution across different time periods.

Subsequently, we proceeded to train multiple models for the named entity task, aiming to establish an initial performance baseline using our dataset. Our primary goal in this evaluation was to assess how different word representations influence the models' performance.

Given that we are operating within a transitional phase of the Chinese language, it became crucial to gauge the significance of contemporary language models (LM) compared to their counterparts trained on ancient Chinese. We sought to determine whether applying state-of-the-art language models designed for modern Chinese would yield comparable results to those trained specifically for historical Chinese, recognizing the evolving nature of the language and its impact on NLP tasks.

Our approach involved leveraging a conventional fine-tuning process, which included modifying a transformer-based language model and incorporating an additional classification layer specifically designed for the NER task. To ensure robustness and accuracy in our NER model, we conducted a comprehensive evaluation.

The evaluation of each language models specifically adapted for the nuances of Chinese language processing is based on the average of 20 separate training runs.

We then selected 6 different LMs, 3 trained on contemporary data and 3 on historical data :

- **Bert** (Devlin et al., 2019) the *de facto* language model transformers who were trained mainly on the Chinese Wikipedia (which was translated into simplified and traditional Chinese).

- **Bert-wwm** (Cui et al., 2019) uses the whole word masking strategy during training, as it is more suitable for processing Chinese. Also trained on the Chinese Wikipedia, they use both Simplified and Traditional Chinese in this dump and do not convert the Traditional Chinese portion into simplified one. They also use in-house collected extended data contains encyclopedia, news, and question answering web.

- **MacBert** (Cui et al., 2020) based on the same data and strategy as (Cui et al., 2019) they change the learning task by correcting words in the sentence.

- **SIKU** (Wang et al., 2021) is a Bert model trained on the "Siku Quanshu" dataset, allowing a model with a larger vocabulary than the bert-base (+8663 entries).

- **Bert-ancient** (Wang and Ren, 2022) Trained on a larger-scale dataset from the same period as SIKU, which results from a larger vocabulary than SIKU. (+8417 entries)

- **Guwen** [3] a RoBERTa model pre-trained on the daizhige dataset which contains 15,694 books in Classical Chinese.

LMs	Precision	Recall	F1
Bert	54.10%	59.61%	56.72%
Bert-wwm	51.24%	58.69%	54.70%
MacBert	55.53%	61.27%	58.26%
Siku	53.70%	60.36%	56.83%
Bert-ancient	43.65%	54.95%	48.64%
Guwen	41.50%	46.58%	43.89%

Table 4: Results obtained by the different LMs on the Chinese part of our dataset.

The results from this experiment, as detailed in Table 4, shed light on the complexities involved in working with documents from this transitional period. Determining the linguistic characteristics of the Chinese language during our specified period of interest (1872-1947) is not a straightforward task. Contrary to initial assumptions, the language used during this era does not necessarily lean closer to contemporary Chinese than it does to ancient Chinese.

Additionally, the results raise intriguing observations about the language models, which were not specifically trained on press-related content, we employed. Our dataset seems to align more closely with the linguistic patterns found in Wikipedia-like language than with the language typically encountered in literary works or historical news articles.

These outcomes underscore the imperative need to employ specialized approaches when working with data from this specific historical period. One potential avenue for improvement could involve

[3] https://huggingface.co/ethanyt/guwenbert-base

training language models on a dedicated corpus of historical press articles. Such an approach would better capture the unique nuances, vocabulary, and language dynamics characteristic of the transitional period, thus enhancing the accuracy and effectiveness of language models in handling this type of historical data. This highlights the importance of tailoring NLP techniques to the unique challenges posed by historical documents, ensuring a more accurate and insightful analysis of this rich historical context.

6. Discussion

The experiments conducted, and the results obtained using this new dataset are, at this stage, in a preliminary phase of exploration. Given the unique nature of this dataset and the diverse range of annotations it encompasses, there are numerous avenues for research and study. It stands out as an exceptional resource within the field of NLP.

For each of the proposed tasks within NLP, such as named entity recognition, entity linking, coreference resolution and relation extraction, the challenges and methodologies employed vary significantly. This diversity is largely attributed to the complex nature of the dataset, which encompasses a language, Chinese, that is undergoing continuous evolution. Managing these distinct NLP tasks within the context of a rapidly changing language presents several formidable challenges for automated language processing.

Beyond the realm of NLP, this dataset holds immense potential for linguists and, undoubtedly, scholars in the domain of digital humanities. It serves as a valuable asset for delving into the intricacies of a language that is in a constant state of flux. Furthermore, it is a testament to the interdisciplinary nature of digital data, offering an expansive landscape for exploration and analysis that transcends conventional boundaries. This dataset not only provides insights into language and history but also highlights the collaborative synergy between technology, culture, and human scholarship.

To foster collaboration and engagement among these diverse communities, we have a broader vision in mind. We intend to incorporate this dataset into an evaluation campaign dedicated to the processing and analysis of historical documents. This initiative goes beyond the mere provision of data; it aims to create a dynamic platform where researchers, historians, linguists, NLP experts, and digital humanities scholars can collectively contribute, evaluate, and advance the state of the art in historical document analysis. That is why we are keeping the test set gold annotations at this time.

7. Conclusion

Within this research paper, we proudly release a novel dataset meticulously annotated for named entities, entity linking, coreference, and relational information. This rich dataset draws its content from Chinese newspapers and biographies, encapsulating a treasure trove of historical and linguistic insights.

The extensive annotation process, as well as the promising initial results we have presented, instill in us the conviction that this dataset holds substantial potential and utility for a wide array of research endeavors spanning diverse domains. Whether it be in the realms of linguistics, history, natural language processing, or digital humanities, this dataset promises to be a valuable asset, facilitating nuanced analyses and enhancing our understanding of the Chinese language and its historical context.

Building upon this foundation, we are currently engaged in a two-pronged effort. Firstly, we are dedicated to crafting annotations that are intricately designed to capture the nuances of events within the newspaper data. This initiative underscores our unwavering commitment to continually enhance and fine-tune the dataset, aligning it with the evolving requirements of researchers and scholars. We aim to ensure that this dataset maintains its enduring relevance and continues to play a pivotal role in advancing knowledge and scholarship.

Concurrently, we are actively involved in the development of NLP approaches tailored specifically to handle this distinct type of data. This journey begins with the refinement and training of language models customized to the unique characteristics of Chinese as it appears in historical newspapers during our defined period of interest. This proactive approach underscores our dedication to advancing the capabilities of NLP techniques for effectively processing and extracting insights from historical sources, thereby contributing to a deeper understanding of our linguistic and historical heritage.

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 the annotators Chang Yu-jun, Chiang Chia-wei, Hu Yi-fan, Kuo Chih-wen, Tseng Chin-ying, Guo Weiting and Jiang Jie.

8. Bibliographical References

2021. Modern China Biographical Database.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *CoRR*, abs/2109.11406.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 288–310, Cham. Springer International Publishing.

Maud Ehrmann, Camille Watter, Matteo Romanello, and Simon Clematide. 2019. Impresso Named Entity Annotation Guidelines.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Günter Hackl, Jose G Moreno, and Antoine Doucet. 2021. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334, Virtual Event, Canada. ACM.

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.

Helena Hubková, Pavel Kral, and Eva Pettersson. 2020. Czech Historical Named Entity Corpus v 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4458–4465, Marseille, France. European Language Resources Association.

Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2016. Old Content and Modern Tools - Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910. ArXiv:1611.02839 [cs].

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

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.).

Clemens Neudecker. 2016. An Open Corpus for Named Entity Recognition in Historic Newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4348–4352, Portorož, Slovenia. European Language Resources Association (ELRA).

Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn, and Pierre Zweigenbaum. 2012. Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 40–48, Jeju, Republic of Korea. Association for Computational Linguistics.

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.

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.