Creating Biographical Networks from Chinese and English Wikipedia

Baptiste Blouin, Nora van den Bosch, Pierre Magistry

Blouin, Baptiste
Bosch, Nora Van den
Magistry, Pierre

Creating Biographical Networks from Chinese and English Wikipedia

Journal of Historical Network Research x (202x) xx-xx

Keywords

Wikipedia, biography, deep learning, historical network analysis, Wikidata, BERT, NER

Abstract

With the rise of digital humanities, historians explore how to intellectually engage with textual sources given the available computational tools of today. The ENP-China project employs Natural Language Processing methods to tap into sources of unprecedented scale with the goal to study the transformation of elites in Modern China (1830–1949).[1] One of the subprojects is extracting various kinds of data from biographies and, for that, we created a large corpus of biographies automatically collected from the Chinese and English Wikipedia. The dataset contains 228,144 biographical articles from the offline Chinese Wikipedia copy and is supplemented with 110,713 English biographies that are linked to a Chinese page. We also enriched this bilingual corpus with metadata that records every mentioned person, organization, geopolitical entity and location per Wikipedia biography and links the names to their counterpart in the other language. This data structure allows the researcher to analyze the relationships between biographies via shared contents and compare networks in different language settings. In this paper we will describe our methodology for building this new dataset. The first step was to use automatic text classification for extracting Chinese biographies. We trained a binary classifier to detect biographies on manually classified examples and used a subset of unseen texts to assess its accuracy. The second step used Named Entity Recognition to generate metadata and extract relations from the links in Wikipedia. Furthermore, we will delve into the method for building networks from this dataset. We argue that depending on the specific research question, different networks may be built. Using the metadata, researchers can create various kinds of networks to suit their needs. On top of releasing this dataset as an enriched bilingual corpus, we will provide an online interface to query and explore it. Our interface benefits from the bipartite graph structure (it can be seen as a network of documents and entities) and applies the same exploration and clustering strategy as in Cillex. [2]

[1] https://enepchina.hypotheses.org/
[2] https://www.istex.fr/cillex/

1. Introduction

The project "Elites, Networks and Power in Modern China" (ENP - China) aims to study the history of elites in Modern China (1830–1949) using computational techniques, and one of the subprojects is extracting various kinds of data from elites' biographies. Biographies are an important source in historical research, as they not only contextualize past lives and embed them in larger narratives and various social networks, but they are also a time capsule for the way people's pasts were evaluated. For this subproject, we turn to one of the biggest sources of the web, Wikipedia, to create a collection of biographies. We employed automatic text classification to obtain biographies from the Chinese Wikipedia and supplemented the collection with English Wikipedia biographies from another project. Furthermore, we added an extensive index that helps us navigate through the corpus.

Although the biographies were collected as part of the subproject, the corpus presented in this article is not limited to articles about Chinese elites. We started by selecting all articles from the Chinese and English Wikipedia that met the criteria of being a biography and will at a later stage proceed to filter out the relevant texts on Chinese elites from the pool of documents. As a result, other researchers (Chinese- or English-speaking) can use our preliminary corpus to explore biographies of historical figures from different time periods. With the corpus, they can examine how the lives of specific people are presented in the online public space of today and, by combining the articles with the index, they can map and analyze the relationships between biographical texts. Since the documents originate from two different language and cultural communities, one can also use the same index to gain insight into how biographies are linked differently depending on the language/cultural context in which they were created. In other words, Scholars can exploit the structure of this dataset to produce and study various kinds of biographical networks.

The paper is divided into three parts. The first part describes the content, structure, and size of the dataset. The second part focuses on the compilation of the corpus and metadata. It will discuss the extraction and update of Chinese biographies, the selection of English biographies, and the named entity recognition. The last part briefly presents a use case of the dataset.

2. The Dataset

2.1. Content

The dataset consists of two major components: the English- and Chinese-language biographies and the metadata. We started building the corpus from the Chinese Wikipedia. Although English is the dominant language of Wikipedia (with its over six million articles), that does not mean that the Chinese Wikipedia is merely a translated version of the English entries. The two sites are edited independently from each other, which results in a number of Chinese articles not connected to an English page. Therefore, we used Machine Learning techniques to extract Chinese articles that are likely to be a biography and added their English counterparts at a later stage. The collection of the English biographies was done in a previous project.[3] However, we

[3] Lebret, Remi, David Grangier, and Michael Auli. "Neural Text Generation from Structured Data with Application to the Biography Domain." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin Texas, USA, November, 2016. http://arxiv.org/abs/1603.07771.

selected only those English articles that have a link to the Chinese page in the text. Although the English articles from that project were collected in 2016, we used their page identifiers to obtain the most recent versions of the pages.

Gathering the biographical entries was only the first step in creating the dataset. We also enriched them with additional information that creates interlinks within the whole corpus. Two types of metadata connect the articles. One is the named entities. Named entities refer to all things with a proper noun that are mentioned in the text. In this dataset, we recorded every mention of a person, an organization, a geopolitical entity (GPE) or location in an article. As a result, we have an index of the corpus that allows us to group biographies based on where they intersect in terms of content and observe the presence or absence of relationships between the documents' content. The other metadata is inter-language links, i.e., links to an article on the same subject but in another language. We used those links as a bridge between the Chinese and English biographies and also used them to establish connections between mentioned names, given an URL is provided in the biography text for that name. The latter helps track the name in both languages, as well as to differentiate instances with the same name.

The dataset released on Zenodo[4] consists of a very large set of files (see below). To provide a more convenient way to browse it and get an overview of what can be found in it, we decided to design an online exploration tool.[5] This tool was inspired by our previous work on Cillex[6] and combines a full-text search index (Solr) and a network exploration interface (based on Padagraph[7]). It is available as a web application operating in three steps. The first step allows the user to run a query against the Solr index to retrieve a set of Wikipedia biographies with associated named entities. The query can be written in English or Chinese, and it is possible to expand the results by also retrieving the corresponding pages in the other language (whenever an inter-wiki link was found). The second step helps the user to build a table that displays results in a format ready for visualization in Padagraph. The user can freely edit the table or proceed directly to the third step: graph exploration. Our interface dynamically creates networks, where the nodes represent Wikipedia pages and entity mentions. Edges between two pages are drawn when there is an inter-language link between the two pages. There are also edges that connect entities with the pages in which they are mentioned. The exploration tool allows the user to specify a query as a starting point of a random walk in the graph (to explore the neighborhood of a node) or to view a "global" graph made of the most central nodes in the result set. When a node is selected, the interface displays its properties, such as a link to the actual wiki page or a picture if we could find one.

2.2. Structure

The data of each article is saved in a folder that carries the Wikipedia page identifier, which can be found in the html content, as its name. In every folder one finds an xml file and a csv file, both having the same name. The xml file contains all the information of the article: the raw text, the Wikipedia ID, the URL, the article title, the identifier of the corresponding page in the other

[4] https://zenodo.org/record/4059194
[5] https://pdg.enpchina.eu/wiki-cillex
[6] https://www.istex.fr/cillex/
[7] https://www.padagraph.io/

language, as well as the URL of that other page. The csv file is the place where all named entities are stored. Every row in the file records a name mentioned in the article, the position of that name in the sentence, and the type of the mention (i.e., 'person,' 'geo-political entity,' 'location,' and 'organization').

When one applies named entity recognition to documents, there is some degree of information loss, in the sense that it does not distinguish different entities with the same name. For example, all places with the name 'Paris' (whether in the USA or France) would be treated as one single instance. However, we reduced this problem by means of entity linking. We found that sometimes a name of a person, an organization, a GPE or location is followed by a URL link in the Wikipedia text and that link leads to the article about that mentioned subject. This extra data could help distinguish various entities that share the same name. Another reason to incorporate this information is that this data provides an indirect way to connect entities that have both a Chinese and an English name. That is why some named entities in the file are accompanied by a URL of the article, the Wikipedia ID, as well as the link and identifier of the corresponding page in the other language. However, we did not go any further in entity linking. We only relied on the links proposed in Wikipedia pages, so it is possible that in some cases, the same entity points to two different pages or, conversely, that two different entities point to the same page. It is also possible that the same entity is linked in a biography in one language but not in another.

id	entity	type	start_pos	end_pos	link_zh	id_zh	link_en	id_en
2278	周恩來	PER	0	5	None	None	None	None
...	...	...	...	...	...	...	...	...
2278	中国工农红军第一方面军	ORG	69	90	https://zh.wikipedia.org/wiki/中国工农红军第一方面军	361551	None	None
2278	中共中央革命军事委员会	ORG	101	122	https://zh.wikipedia.org/wiki/中共中央革命军事委员会	9620	https://en.wikipedia.org/wiki/Central_Military_Commission_(China)	214345
...	...	...	...	...	...	...	...	...
2278	中國	GPE	1	4	None	None	None	None
...	...	...	...	...	...	...	...	...
2278	江苏淮安城内驸马巷	LOC	60	77	None	None	None	None

Figure 1. Sample of different named entities retrieved from the Chinese biography of Zhou Enlai. It shows the document ID, name of entity, type of entity, the numbers indicating its place in the sentence, the URL and ID of its Wikipedia page and, finally, the link and page identifier of the page in English.

2.3. Size

Documents

We retrieved 228,601 biographical articles from the Chinese Wikipedia dump using a text classifier and collected 728,321 English articles from a project by Lebret et al.[8] In the end, we only kept the Chinese pages that still existed in the current version of the Chinese Wikipedia and the English pages that had a corresponding page in Chinese. We had a total of 338,857 documents, 228,144 of which are Chinese and 110,713 English. Among them, 110,958 pages from the Chinese corpus have an English inter-language link and 110,713 pages from the English corpus have a Chinese one. The difference in the number of links between the two languages is

[8] Lebret, Remi, David Grangier, and Michael Auli. "Neural Text Generation from Structured Data with Application to the Biography Domain." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin Texas, USA, November, 2016. http://arxiv.org/abs/1603.07771.

either due to the fact that several Chinese pages are linked to the same English page or that some pages are linked to pages that no longer existed on the day the corpus was created. For instance, Lady Guan (關氏 or 關羽女, daughter of Guan Yu) and Guan Yu (關羽) are both linked to a single page in English, the one on Guan Yu.

Named Entities

This section presents statistics on the entities named in the corpus. The results are divided into three tables, table 1 for the Chinese corpus, table 2 for the English corpus, and table 3 for the above two corpora combined.

These three tables each contain five columns. The second column gives the sum of occurrences of named entities for each category, and the third column gives the number of distinct annotated named entities. The second part of the table (columns 5 and 4) provides the number of distinct links associated with these entities and their occurrences respectively. As for columns 2 and 3, column 4 contains all the values, and column 5 only the distinct values. The links are coming from Wikipedia and each connects a named entity to its corresponding page. More information on how we obtained these links can be found in section 3.1.

Type	Count	Distinct count	Link	Distinct Link
Persons	3,600,226	802,035	378,904	104,142
Organizations	1,319,972	407,601	146,179	29,972
GPE's	1,782,456	139,939	27,396	8,313
Locations	224,140	63,773	21,398	10,838
TOTAL	6,926,794	1,413,348	573,877	153,265

Table 1. Chinese NER stats

Type	Count	Distinct count	Link	Distinct Link
Persons	4,801,680	1,504,807	331,838	159,727
Organizations	1,896,158	744,388	103,938	48,177
GPE's	1,717,476	260,366	13,347	7,714
Locations	198,509	81,119	53,918	23,244
TOTAL	8,613,823	2,590,680	503,041	238,862

Table 2. English NER stats

Type	Count	Distinct count	Link	Distinct Link
Persons	8,401,906	2,306,842	710,742	263,869
Organizations	3,216,130	1,151,989	250,117	78,149
GPE's	3,499,932	400,305	40,743	16,027
Locations	422,649	144,892	75,316	34,082
TOTAL	15,540,617	4,004,028	1,076,918	392,127

Table 3. Total NER stats

3. Creation Process

In this section, we will discuss the creation process of the dataset. This process involves two steps. In the first part, we will dive into the composition of the corpus. We will outline the method for extracting biographies and talk briefly about the selection of English articles. The second part will focus on named entities and language linking.

3.1. Compilation of the Chinese and English subcorpus

In the first step, we downloaded an offline copy of the Chinese Wikipedia on dumps.wikimedia.org, in an xml format.[9] The offline copy, also called a *wikidump*, has around two million pages (one-third the size of the English Wikipedia). However, not all pages in the wikidump are articles describing a subject. Some pages are meant to redirect visitors to the relevant page, some are lists of subjects with similar names, and others are lists of subjects under the same category, etc. So, after removing these non-articles with the python tool *WikiExtractor*, we reduced the size of the corpus to 1,046,744 pages.[10]

By inspecting the XML files, we concluded that there was no metadata that identifies the biographies and, therefore, we had to rely on the unstructured textual data of the pages. We did some experiments on what method to use for classifying articles into biographies and non-biographies. At first, we tried to select articles by detecting predetermined keywords in the text, such as 'born in' ((*chu*)*sheng zai*... (出)生在) combined with 'family background' (*chushen*出身), etc. Such a method is called rule-based classification. We did some experiments in order to assess the performance of this method. However, after a few iterations, we found out that the sheer number and variety of articles complicated the process of finding the "right" keywords.

This is why we decided to rely on deep learning for text classification. Text classification is an important problem in natural language processing (NLP). The task is to assign a document to one or more predefined categories, in our case, "biography" or "non-biography." It has been used in a wide range of applications, such as sentiment analysis,[11] topic categorization,[12] and

[9] https://dumps.wikimedia.org/zhwiki/
[10] https://github.com/attardi/wikiextractor
[11] Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis," n.d., 94.
[12] Lewis, David D, Yiming Yang, Tony G Rose, and Fan Li. "RCV1: A New Benchmark Collection for Text

email filtering,[13] and the methods for this task have changed significantly over the years. Early machine learning approaches for text classification were based on the extraction of bag-of-words features followed by a supervised classifier such as naïve Bayes[14] or a linear Support Vector Machine.[15] Later, better word representations were introduced, such as latent semantic analysis,[16] skipgram,[17] fastText,[18] and today contextualized word embeddings,[19] which improved classification accuracy. For our extraction, we used one of the most widely used contextualized word representations to date, BERT[19], combined with the neural network's architecture, BiLSTM. BiLSTM is state of the art for many NLP tasks, including text classification. In our case, we trained a model[20] with examples of Chinese biographies and non-biographies so that it relies on specific semantic features of each type of entry in order to predict its category. Therefore, once trained, given a new entry never seen before by the model, the model, based on the representations of the words of this biography and the weights learned by our neural networks, will predict the most probable class to which this entry belongs. Below we will delve deeper into the process of selecting our training data.

In order to train and test a binary classifier, we need to have a collection of examples and counterexamples for the model to process. Not only does the number of articles have to be high enough for the algorithm to "understand the differences," but the types of content need to be varied to minimize potential bias. So, instead of creating a list of articles on our own, we turn to Wikidata to generate it. Wikidata is a central repository that holds data of all kinds of subjects from various sources such as Wikipedia. Every subject is represented as an item, with its unique identifier, a label, and a description, and is further described by triple statements, each consisting of the item identifier, the property, and the value (which is usually the identifier of another item). This way of storing and linking data is highly structured, in the sense that the types of properties are standardized according to community guidelines, and computers are able to infer other statements from triple statements based on a schema that maps relations between properties. For example, 'A is the daughter of B' can be interpreted as 'B is a parent of A.'[21] Such structuring allows for very powerful and specific queries. Using the Wikidata SPARQL endpoint, we obtained lists containing the titles, the Wikidata ID, and the Wikipedia

---

Categorization Research," n.d., 37.

[13] Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz. "A Bayesian Approach to Filtering Junk E-Mail," n.d., 8.

[14] McCallum, Andrew, et Kamal Nigam. « A Comparison of Event Models for Naive Bayes Text Classification », s. d., 8.

[15] Joachims, Thorsten. « Text Categorization with Support Vector Machines: Learning with Many Relevant Features ». In *Machine Learning: ECML-98*, édité par Claire Nédellec et Céline Rouveirol, 1398:137‑42. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. doi:10.1007/BFb0026683.

[16] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, et Richard Harshman. « Indexing by Latent Semantic Analysis ». *Journal of the American Society for Information Science* 41, nº 6 (1990): 391‑407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[17] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, et Jeff Dean. « Distributed Representations of Words and Phrases and Their Compositionality », s. d., 9.

[18] Joulin, Armand, Edouard Grave, Piotr Bojanowski, et Tomas Mikolov. « Bag of Tricks for Efficient Text Classification ». *arXiv:1607.01759 [cs]*, 9 août 2016. http://arxiv.org/abs/1607.01759.

[19] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. « BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding ». *ArXiv:1810.04805 [Cs]*, 24 mai 2019. http://arxiv.org/abs/1810.04805.

[20] https://github.com/flairNLP/flair

[21] https://www.wikidata.org/wiki/Wikidata:Introduction

links of articles.

While obtaining a list of potential biographies was unproblematic, it was a challenge to create an effective sample of non-biographies. Technically, every non-person page could serve as a counterexample. But after carrying out an error analysis, we found that a completely random sample could not prepare the model to reject pages of fictional characters, movies, films, manga, bands, etc. As a result, we recomposed the collection of non-biographies, which consists of:

- 2,860 *fake* persons (items categorized in Wikidata as "fictional characters," "fictional humans," "literary characters," "comics characters," "video game characters," etc.)
- 3,040 media examples (categorized as "films," "television series," "literary works," etc.)

This collection was further supplemented with a list of 2,984 random examples generated by the Wikipedia API. We removed person pages from the latter by accessing the Wikidata page of every list item using its Wikidata ID and rejecting it based on the presence of the attribute 'human' on that page.

To obtain the full texts of example and counterexample articles, we inserted the page titles from the lists on the Wikipedia Special:export page and downloaded the articles in .xml files. After that, we filtered out potential non-articles from the files. The algorithm was given the first three sentences of each (counter)example page and was set to distinguish between the language of a biography and that of a non-biography. To test its performance during the training phase, we used a randomized 10 percent of the training sample. In the end, we did a final test. We presented the algorithm with a set of 415 manually labeled unknown articles. As an outcome, ten articles of the test samples were wrongly classified as biographies, one biography was skipped, and one Chinese article was excluded because it was completely written in the English language. The model reached an accuracy of 97.5%.

With the text classifier, we obtained a list of 228,601 articles that are likely to be biographies. We made a short survey of the list, and it seems the detection of biographies was successful. However, there are still a few false positives, which are mostly pages containing lists, such as awards. This collection is then supplemented with 4,502 extra Chinese Wikipedia articles. These extra articles are Chinese versions of pages present in the English Wikipedia biography dataset[22]. Some of these English articles contained links to Chinese pages that were missing in the Chinese subcorpus created from our automatic extraction. This is a relatively low number considering that in all the biographies of the Lebret et al. project that contained a link to a Chinese page, 61,229 were already detected during our extraction.

We ran the extraction on the February 2020 version of the Chinese wikidump (zhwiki-2020-02-01). We used the collected page identifiers to retrieve the html content from the website around June 17, 2020, in order to have the most up-to-date version of the Wikipedia content since the identifiers are invariant. During this phase, 37 Chinese pages were lost in the process due to the fact that the pages did not exist anymore in the latest version. After that, we

[22] Lebret, Remi, David Grangier, and Michael Auli. "Neural Text Generation from Structured Data with Application to the Biography Domain." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin Texas , USA, November, 2016. http://arxiv.org/abs/1603.07771.

repeated the process of extracting named entities and inter-language links.

For obtaining English biographies, we only kept the English articles that were directly linked to a Chinese counterpart. As a result, based on our 228,144 articles in Chinese, we extracted 110,713 English biographies.

3.2. Named Entities

Named Entity Recognition (NER) is a typical sequence labeling task in the natural language processing field. The objective of this task is to determine entity boundaries and classify them into predefined categories such as persons, organizations and location names. Named Entity Recognition forms a core subtask to build knowledge from semi-structured and unstructured text sources. Because we process data from Wikipedia, which provides hyperlinks to other pages, we broke our entity extraction into two steps. First, by using a pre-trained model, we extracted all the named entities present in a page; then, in a second step, we linked the entities to their Wikipedia pages in Chinese and/or English, if these pages exist. To do so, after extracting an entity we looked in the html source to check if it had a Wikipedia page. If a Wikipedia page identifier was found, we attached this link to the corresponding entity in the current language and then opened the page to check if there was an inter-language link in the other language. We used two models, one for the Chinese page and one for the English page. Both models were trained on OntoNotes corpora[23] in their own language. For both of the models, we used a bidirectional recurrent neural network with a subsequent conditional random field decoding layer proposed by Flair[24] and trained them on OntoNotes.[25] For the Chinese part, we trained our own model, which we have discussed in a different paper,[26] and for the English part we used the pre-trained model proposed by Flair. Extracting named entities from each biography and linking them to their own Wikipedia pages, whenever possible, establishes links between biographies by way of the entities present in their contents. Moreover, since most of the biographies in our corpus as well as the entities are linked with their counterparts in the other language, it is also possible to link, for example, two Chinese biographies by means of the same entities present in their corresponding English pages.

4. Potential use

[23] Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. "OntoNotes: The 90% Solution." In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60. New York City, USA: Association for Computational Linguistics, 2006. https://www.aclweb.org/anthology/N06-2015.

[24] Akbik, Alan, Duncan Blythe, and Roland Vollgraf. "Contextual String Embeddings for Sequence Labeling," In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August, 2018.

[25] Weischedel, Ralph , Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. "Ontonotes release 5.0." Linguistic Data Consortium, October 2013. https://catalog.ldc.upenn.edu/LDC2013T19.

[26] Blouin, Baptiste, Pierre Magistry. "Contextual characters with segmentation representation for named entity recognition in Chinese.' In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, Hanoi, Vietnam (held online), October, 2020.

As mentioned before, the data is the basis for building a complex network incorporating two types of nodes: biographical texts and named entities. The edges either indicate an instance of a mention between page and named entity, or represent an inter-language correspondence between a biographical text (or named entity) in one language and its counterpart in the other language (although the inter-language link of an entity node depends on whether a URL is provided in the biographical text).

To give a use case example for this network data, one may ask whether different language communities in Wikipedia emphasize different types of personal relationships in biographies. For example, a prominent CCP figure such as Zhou Enlai (1898-1976). One tackles the question by drawing comparisons between the egocentric of Zhou Enlai in English and Chinese. For each language, one may construct the network in which the page node, the biography of Zhou Enlai, is linked to different entity nodes, specifically the named entities of the type 'person.' One may then assess the importance of each person's name using its edge weight (number of its occurrences in the text). Finally, one may check whether there is an inter-language link for every entity node and use these links to detect whether some persons occur in both biographical networks. This allows one to discover which nodes are mentioned most frequently in English and Chinese respectively and which nodes receive special attention in both linguo-cultural communities. Of course, the method might be relatively crude as not every mention of a person indicates a significant relationship to the subject of the biography. In a certain way, one has to add another layer of data to evaluate and categorize the type of relationships in the biographical texts. One can also flip the approach by examining which biographies mention Zhou Enlai and what relationship Zhou Enlai has to the subjects of those biographies.

Although this is an example for analyzing links between the biographies and mentioned names , one can also exploit the mapped relations to easily create a subset of biographies and use the subcorpus for other modes of inquiry, such as discourse analysis, or, in case of ENP-China, for data extraction. One is not confined by the given data structure but can repurpose the dataset in accordance to particular needs and goals.

5. Conclusion

We compiled a large pool of biographies from Chinese and English Wikipedia. In order to make this bilingual corpus accessible, we enriched it with an extensive index that lists all mentioned persons, organizations, geopolitical entities and locations per article and also collected inter-language links between the pages and between the mentioned names (given that the latter is accompanied by a link in the text).

Although the ENP - China project uses this corpus to extract data on Chinese elites in the Republican era, this dataset could be repurposed for building a network, in which biographical texts are indirectly connected to each other via the names mentioned in them. One can make use of the index to study the relationships of historical figures presented in popular digital sources, like Wikipedia, and can go further to compare networks of biographies written in different languages by

using the inter-language links.

As mentioned above, such a dataset can be used for various purposes, which is why we decide to make the data available for those interested in analyzing relationships between online biographies. We hope that this dataset can contribute to the historical network research community and provide the scholars with an opportunity to engage with biographical texts in novel ways.

6. Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 788,476).

7. Bibliography

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. "Contextual String Embeddings for Sequence Labeling," In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August, 2018.

Blouin, Baptiste, Pierre Magistry. "Contextual characters with segmentation representation for named entity recognition in Chinese.' In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, Hanoi, Vietnam (held online), October, 2020.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41, no. 6 (1990): 391–407. doi:10.1002/(SICI)1097–4571(199,009)41:6<391::AID-ASI1>3.0.CO;2–9.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT 2019,* 4171–4186, Minneapolis, USA, May 24, 2019. http://arxiv.org/abs/1810.04805.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. "OntoNotes: The 90% Solution." In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60. New York City, USA: Association for Computational Linguistics, 2006. https://www.aclweb.org/anthology/N06-2015.

Joachims, Thorsten. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In *Machine Learning: ECML-98*, edited by Claire Nédellec and Céline Rouveirol, 1398: 137–42. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. doi:10.1007/BFb0026683.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. "Bag of Tricks for Efficient Text Classification." In *EACL (2)*, by Mirella Lapata, Phil Blunsom, and Alexander Koller, 427–31. edited by Mirella Lapata, Phil Blunsom, and Alexander Koller. Association for Computational Linguistics, 2017. http://dblp.uni-trier.de/db/conf/eacl/eacl2017-2.html#GraveMJB17.

Lebret, Rémi, David Grangier, and Michael Auli. "Neural Text Generation from Structured Data

with Application to the Biography Domain." In *EMNLP*, by Jian Su, Xavier Carreras, and Kevin Duh, 1203–13. edited by Jian Su, Xavier Carreras, and Kevin Duh. The Association for Computational Linguistics, 2016. http://dblp.uni-trier.de/db/conf/emnlp/emnlp2016.html#LebretGA16.

Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li. "RCV1: A New Benchmark Collection for Text Categorization Research." *J. Mach. Learn. Res.* 5 (2004): 361--397. http://portal.acm.org/citation.cfm?id=1005332.1005345.

McCallum, Andrew, and Kamal Nigam. "A Comparison of Event Models for Naive Bayes Text Classification." In *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, 41--48, 1998. http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed Representations of Words and Phrases and Their Compositionality." In *Advances in Neural Information Processing Systems 26*, by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, 3111--3119. edited by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, 2013. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Pang, Bo, and Lillian Lee. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2, no. 1–2 (2008): 1--135. doi:10.1561/1500000011.

Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz. "A Bayesian Approach to Filtering Junk E-Mail." In *Learning for Text Categorization: Papers from the 1998 Workshop*. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998. citeseer.ist.psu.edu/sahami98bayesian.html.

Weischedel, Ralph , Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. "Ontonotes release 5.0." *Linguistic Data Consortium*, October 2013. https://catalog.ldc.upenn.edu/LDC2013T19.