Summarising Historical Text in Modern Languages

Conference or Workshop Item

How to cite:


For guidance on citations see FAQs.

© 2021 ACL

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Accepted Manuscript

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
Summarising Historical Text in Modern Languages

Xutan Peng* Yi Zheng Chenghua Lin* Advaith Siddharthan*

Department of Computer Science, The University of Sheffield, UK
School of Computer Science and Engineering, Beihang University, China
Knowledge Media Institute, The Open University, UK
{x.peng, c.lin}@shef.ac.uk zhengyi53@buaa.edu.cn
advaith.siddharthan@open.ac.uk

Abstract

We introduce the task of historical text summarisation, where documents in historical forms of a language are summarised in the corresponding modern language. This is a fundamentally important routine to historians and digital humanities researchers but has never been automated. We compile a high-quality gold-standard text summarisation dataset, which consists of historical German and Chinese news from hundreds of years ago summarised in modern German or Chinese. Based on cross-lingual transfer learning techniques, we propose a summarisation model that can be trained even with no cross-lingual (historical to modern) parallel data, and further benchmark it against state-of-the-art algorithms. We report automatic and human evaluations that distinguish the historic to modern language summarisation task from standard cross-lingual summarisation (i.e., modern to modern language), highlight the distinctness and value of our dataset, and demonstrate that our transfer learning approach outperforms standard cross-lingual benchmarks on this task.

1 Introduction

The process of text summarisation is fundamental to research into history, archaeology, and digital humanities (South, 1977). Researchers can better gather and organise information and share knowledge by first identifying the key points in historical documents. However, this can cost a lot of time and effort. On one hand, due to cultural and linguistic variations over time, interpreting historical text can be a challenging and energy-consuming process, even for those with specialist training (Gray et al., 2011). To compound this, historical archives can contain narrative documents on a large scale, adding to the workload of manually locating important elements (Gunn, 2011). To reduce these burdens, specialised software has been developed recently, such as MARKUS (Ho and Weerdt, 2014) and DocuSky (Tu et al., 2020). These toolkits aid users in managing and annotating documents but still lack functionalities to automatically process texts at a semantic level.

Historical text summarisation can be regarded as a special case of cross-lingual summarisation (Leuski et al., 2003; Orăscan and Chiorean, 2008; Cao et al., 2020), a long-standing research topic whereby summaries are generated in a target language from documents in different source languages. However, historical text summarisation poses some unique challenges. Cross-lingual (i.e., across historical and modern forms of a language) corpora are rather limited (Gray et al., 2011) and therefore historical texts cannot be handled by traditional cross-lingual summarisers, which require cross-lingual supervision or at least large summarisation datasets in both languages (Cao et al., 2020). Further, language use evolves over time, including vocabulary and word spellings and meanings (Gunn, 2011), and historical collections can span hundreds of years. Writing styles also change over time. For instance, while it is common for today’s news stories to present important information in the first few sentences, a pattern exploited by modern news summarisers (See et al., 2017), this was not the norm in older times (White, 1998).

In this paper, we address the long-standing need for historical text summarisation through machine summarisation techniques for the first time. We consider the German|DE and Chinese|ZH languages, selected for the following reasons. First, they both have rich textual heritage and accessible (monolingual) training resources for historical and modern language forms. Second, they serve as outstanding representatives of two distinct writing systems (DE...
The contributions of our work are three-fold: (1) we propose a hitherto unexplored and challenging task of historical text summarisation; (2) we construct a high-quality summarisation corpus for historical DE and ZH, with modern DE and ZH summaries by experts, to kickstart research in this field; and (3) we propose a model for historical text summarisation that does not require parallel supervision and provides a validated high-performing baseline for future studies. We release our code and data at https://github.com/Pzoom522/HistSumm.

2 Related Work

Processing historical text. Early NLP studies for historical documents focus on spelling normalisation (Piotrowski, 2012), machine translation (Oravecz et al., 2010), and sequence labelling applications, e.g., part-of-speech tagging (Rayson et al., 2007) and named entity recognition (Sydow et al., 2011). Since the rise of neural networks, a broader spectrum of applications such as sentiment analysis (Hamilton et al., 2016), information retrieval (Pettersson et al., 2016), and relation extraction (Opitz et al., 2018) have been developed.

We add to this growing literature in two ways. First, much of the work on historical text processing is focused on English [EN], and work in other languages is still relatively unexplored (Piotrowski, 2012; Rubinstein, 2019). Second, the task of historical text summarisation has never been tackled before, to the best of our knowledge. A lack of non-EN annotated historical resources is a key reason for the former, and for the latter, resources do not exist in any language. We hope to spur research on historical text summarisation and in particular for non-EN languages through this work.

Cross-lingual summarisation. The traditional strands of cross-lingual text summarisation systems design pipelines which learn to translate and summarise separately (Leuski et al., 2003; Oråsan and Chiorean, 2008). However, such paradigms suffer from the error propagation problem, i.e., errors produced by upstream modules may accumulate and degrade the output quality (Zhu et al., 2020). In addition, parallel data to train effective translators is not always accessible (Cao et al., 2020). Recently, end-to-end methods have been applied to alleviate this issue. The main challenge for this
research direction is the lack of direct corpora, leading to attempts such as zero-shot learning (Duan et al., 2019), multi-task learning (Zhu et al., 2019), and transfer learning (Cao et al., 2020). Although training requirements have been relaxed by these methods, our extreme setup with summarisation data only available for the target language and very limited parallel data, has never been visited before.

3 HISTSUMM Corpus

3.1 Dataset Construction

In history and digital humanities research, summarisation is most needed when analysing documentary and narrative text such as news, chronicles, diaries, and memoirs (South, 1977). Therefore, for DE we picked the GerManC dataset (Durrell et al., 2012), which contains Optical Character Recognition (OCR) results of DE newspapers from the years 1650–1800. We randomly selected 100 out of the 383 news stories for manual annotation. ForZH, we chose 『万历邸抄』 (Wanli Gazette) as the data source, a collection of news stories from the Wanli period of Ming Dynasty (1573–1620). However, there are no machine-readable versions of Wanli Gazette available; worse still, the calligraphy copies (see Appendix B) are unrecognisable even for non-expert humans, making the OCR technique inapplicable. Therefore, we performed a thorough literature search on over 200 related academic papers and manually retrieved 100 news texts.\(^1\)

To generate summaries in the respective modern language for these historical news stories, we recruited two experts with degrees in Germanistik and Ancient Chinese Literature, respectively. They were asked to produce summaries in the style of DE MLSUM (Scialom et al., 2020) and ZH LCSTS (Hu et al., 2015), whose news stories and summaries are crawled from the Süddeutsche Zeitung website and posts by professional media on the Sina Weibo platform, respectively. The annotation process turned out to be very effort-intensive: for both languages, the experts spent at least 20 minutes in reading and composing a summary for one single news story. The accomplished corpus of 100 news stories and expert summaries in each language, namely HISTSUMM (see examples in Tab. 1), were further examined by six other experts for quality control (see details in § 6.2).

\(^1\)Detailed references are included in the ‘source’ entries of ZH HISTSUMM’s metadata.

Figure 1: Publication time of HISTSUMM stories.

Figure 2: Topic composition of HISTSUMM.

<table>
<thead>
<tr>
<th>Year</th>
<th>DE</th>
<th>ZH</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>1650</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>1700</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>1750</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>1800</td>
<td>28</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 2: Comparisons of mean story length (Lstory), summary length (Lsummary), and compression rate (CR = Lsummary / Lstory) for summarisation datasets.

3.2 Dataset Statistics

Publication time. As visualised in Fig. 1, the publication time of DE and ZH HISTSUMM stories exhibits distinguished patterns. Oldness is an important indicator of the domain and linguistic gaps (Gunn, 2011). Considering news in ZH HISTSUMM is on average 137 years older than its DE counterpart, such gaps can be expected to be greater. On the other hand, DE HISTSUMM stories cover a period of 150 years, compared to just 47 years for ZH, indicating the potential for greater linguistic and cultural variation within the DE corpus.

Topic composition. For a high-level view of HISTSUMM’s content, we asked experts to manually classify all news stories into six categories (shown in Fig. 2). We see that the topic compositions of DE and ZH HISTSUMM share some similarities. For instance, Military (e.g., battle reports) and Politics (e.g., authorities’ policy and personnel changes) together account for more than half the stories in both languages. On the other hand, we also have language-specific observations. 9% DE stories are about Literature (e.g., news about book publications), but this topic is not seen in ZH HISTSUMM. And while 14% DE stories are about
Sovereign (e.g., royal families and Holy See), there are only 2 examples in ZH (both about the emperor; we found no record on any religious leader in Wanli Gazette). Also, the topics of Society (e.g., social events and judicial decisions) and Natural Disaster (e.g., earthquakes, droughts, and floods) are more prevalent in the ZH dataset.

**Story length.** In news summarisation tasks, special attention is paid to the lengths of news stories and summaries (see Tab. 2). Comparing DE HIST-SUMM with the corresponding modern corpus DE MLSUM, we find that although historical news stories are on average 53% shorter, the overall compression rate (CRs) is quite similar (6.8% vs 5.8%), indicating that key points are summarised to similar extents. Following LCSTS (Hu et al., 2015), the table shows character-level data for ZH, but this is somewhat misleading. While most modern words are double-character, single-character words dominate the historical vocabulary, e.g., the historical word ‘朋’ (*friend*) becomes ‘朋友’ in modern ZH. According to Che et al. (2016), this leads to a character length ratio of approximately 1:1.6 between parallel historical and modern samples. Taking this into account, the CRs for the ZH HISTSUMM and LCSTS datasets are also quite similar to each other.

When contrasting DE with ZH (regardless of historical or modern), we notice that the compression rate is quite different. This might reflect stylistic variations with respect to how verbose news reports are in different languages or by different writers.

### 3.3 Vicissitudes of News

Compared with modern news, articles in HISTSUMM reveal several distinct characteristics with respect to writing style, posing new challenges for machine summarisation approaches.

**Lexicon.** With social and cultural changes over the centuries, lexical pragmatics of both languages have evolved substantially (Gunn, 2011). For DE, some routine concepts from hundreds of years ago are no longer in use today, e.g., the term ‘Brachmonat’ (№41), whose direct translation is *fallow month*, actually refers to *June* as the cultivation of fallow land traditionally begins in that month (Grimm, 1854). We observe a similar phenomenon in ZH HISTSUMM, e.g., ‘贡市’ (№24 and №31) used to refer to markets that were open to foreign merchants, but is no longer in use. For ZH, additionally, we notice that although some historical words are still in use, their semantics have changed over time, e.g., meaning of ‘聞’ has shifted from *hear to smell* (№53), and that of ‘走’ has changed from *run to walk* (№25).

**Syntax.** Another aspect of language change is that some historical syntax has been abandoned. Consider ‘daß derselbe noch länger allda/ biß der Frantz. Abgesandter von dannen widerrum abreisen möge/ verbleiben soll’ (the same should still remain there for longer, until the France Ambassador might leave again) (№33). We find the subordinate clause is inserted within the main clause, whereas in modern DE it should be ‘daß derselbe noch länger allda verbleiben soll, biß der Frantz. Abgesandter von dannen widerrum abreisen möge’. For ZH, inversion is common in historical texts but becomes rare in the modern language. For example, sentence ‘王氏之女成仙者’ (Ms. Wang’s daughter who became a fairy) (№65) where the attributive adjective is positioned after the head noun, should be ‘王氏之成仙（的）女’ according to modern ZH grammars. Also, we observe cases where historical ZH sentences without constituents such as subjects, predicates, objects, prepositions, etc. In these cases, contexts must be utilised to infer corresponding information, e.g., only by adding ‘居正’ (*Juzheng*, a minister’s name) to the context can we interpret the sentence ‘已, 又为私书安之云’ (№20) as ‘after that, (Juzheng) wrote a private letter to comfort him’. This adds extra difficulty to the generation of summaries.

**Writing style.** To inform readers, a popular practice adopted by modern news writers is to introduce key points in the first one or two sentences (White, 1998). Many machine summarisation algorithms leverage this pattern to enhance summarisation quality by incorporating positional signals (Edmundson, 1969; See et al., 2017; Gui et al., 2019). However, this rhetorical technique was not widely used in HISTSUMM, where crucial information may appear in the middle or even the end of stories. For instance, the keyword ‘Türck’ (*Turkish*) (№33) first occurs in the second half of the story; in article №7 of ZH HISTSUMM (see Tab. 1), only after reading the last sentence can we know the final outcome (i.e., the authority’s seal had been saved from fire).

### 4 Methodology

Based on the popular cross-lingual transfer learning framework of (Ruder et al., 2019), we propose
a simple historical text summarisation framework (see Fig. 3), which can be trained even without supervision (i.e., parallel historical-modern signals).

**Step 1.** For both DE and ZH, we begin with respectively training modern and historical monolingual word embeddings. Specially, for DE, following the suggestions of Wang et al. (2019), we selected subword-based embedding algorithms (e.g., FastText (Joulin et al., 2017)) as they yield competitive results. In addition to training word embeddings on the raw text, for historical DE we also consider performing text normalisation (NORM) to enhance model performance. This orthographic technique aims to convert words from their historical spellings to modern ones, and has been widely adopted as a standard step by NLP applications for historical alphabetic languages (Bollmann, 2019). Although training a normalisation model in a fully unsupervised setup is not yet realistic, it can get bootstrapped with a single lexicon table to yield satisfactory performance (Ljubešić et al., 2016; Scherrer and Ljubešić, 2016).

For ideographic languages like ZH, word embeddings trained on stroke signals (which is analogous to subword information of alphabetic languages) achieve state-of-the-art performance (Cao et al., 2018), so we utilise them to obtain monolingual vectors. Compared with simplified characters (which dominate our training resources), traditional ones typically provide much richer stroke signals and thus benefit stroke-based embeddings (Chen and Sheng, 2018), e.g., traditional ‘葉’ (leaf) contains semantically related components of ‘木’ (plant) and ‘木’ (wood), while its simplified version (‘叶’) does not.

Therefore, to improve the model performance we also conduct additional experiments on enhanced corpora which are converted to the traditional glyph using corresponding rules (CONV) (see § 5.3 for further details).

**Step 2.** Next, we respectively build two semantic spaces for DE and ZH, each of which is shared by historical and modern word vectors. This approach, namely cross-lingual word embedding mapping, aligns different embedding spaces using linear projections (Artetxe et al., 2018; Ruder et al., 2019). Given parallel supervision is very limited in real-world scenarios, we mainly consider two bootstrapping strategies: in a fully unsupervised (UspMap) style and through identical lexicon pairs (IdMap). While the former only relies on topological similarities between input vectors, the latter additionally takes advantage of words in the intersected vocabulary as seeds. Although their historical and current meanings can differ (cf. § 3.3), in most cases they are similar, providing very weak parallel signals (e.g., ‘Krieg’ (war) and ‘Frieden’ (peace) are common to historical and modern DE; ‘天’ (universe) and ‘人’ (human) to historical and modern ZH).

**Step 3.** In this step, for each of DE and ZH we use a large monolingual modern-language summarisation dataset to train a basic summariser that only takes modern-language inputs. Embedding weights of the encoder are initialised with the modern partition of corresponding cross-lingual word vectors in Step 2 and are frozen during the training process, while those of the decoder are randomly initialised and free to update through back-propagation.

**Step 4.** Upon convergence in the last step, we directly replace the embedding weights of the encoder with the historical vectors in the shared vector space, yielding a new model that can be fed with historical inputs but output modern sentences. This entire process does not require any external parallel supervision.

5 Experimental Setup

5.1 Training Data

Consistent with § 3.1, we selected DE MLSUM and ZH LCSTS as monolingual summarisation training sets. For monolingual corpora for word embedding training, to minimise temporal and domainal variation, we only considered datasets that were similar.
to articles in MLSUM, LCSTS, and HistSUMM, i.e., with text from comparable periods and centred around news-related domains.

For modern DE, such resources are easy to access: we directly downloaded the DE News Crawl Corpus released by WMT 2014 workshops (Bojar et al., 2014), which contains shuffled sentences from online news sites. We then conducted tokenisation and removed noise such as emojis and links. For historical DE, besides the already included GerManC corpus, we also saved Deutsches Textarchiv (Nolda, 2019), Mercurius-Baumbank (Ulrike, 2020), and Mannheimer Korpus (Mannheim, 2020) as training data. Articles in these datasets are all relevant to news and have topics such as Society and Politics. Note that we only preserved documents written in 1600 to 1800 to match the publication time of DE HistSUMM stories (cf. § 3.2). Apart from the standard data cleaning procedures (tokenisation and noise removal, as mentioned above), for historical DE corpora we replaced the very common slash symbols (/) with their modern equivalents: commas (,) (Lindemann, 2015). We also lower-cased letters and deleted sentences with less than 10 words, yielding 505K sentences and 12M words in total.

For modern ZH, we further collected news articles in the corpora released by He (2018), Hua et al. (2018), and Xu et al. (2020) to train better embeddings. For historical ZH, to the best of our knowledge, there is no standalone Ming Dynasty news collection except Wanli Gazette. Therefore, from the resources released by Jiang et al. (2020), we retrieved Ming Dynasty articles belonging to categories of Novel, History/Geography, and Military. Raw historical ZH text does not have punctuation marks, so we first segmented sentences using the Jiayan Toolkit. Although Jiayan supports tokenisation, we skipped this step as the accuracy is unsatisfactory. Given that a considerable amount of historical ZH words only have one character (cf. § 3.2 and § 3.3), following Li et al. (2018) we simply treated characters as basic units during training. Analogous to historical DE, we removed sentences with less than 10 characters. The remaining corpus has 992k sentences and 28M characters.

5.2 Baseline Approaches
In addition to the proposed method, we also consider two strong baselines based on the Cross-lingual Language Modelling paradigm (XLM) (Lample and Conneau, 2019), which has established state-of-the-art performance in the standard cross-lingual summarisation task (Cao et al., 2020). More concretely, for DE and ZH respectively, we pretrain baselines on all available historical and modern corpora using causal language modelling and masked language modelling tasks. Next, they are respectively fine-tuned on modern text summarisation and unsupervised machine translation tasks. The former becomes the (XLM-E2E) baseline, which can be directly executed on HistSUMM in an end-to-end fashion; the latter (XLM-Pipe) is coupled with the basic summariser for modern inputs in Step 3 of § 4 to form a translate-then-summarise pipeline.

5.3 Model Configurations
Normalisation and convention. We normalised historical DE text using cSMTiser (Ljubešić et al., 2016; Scherrer and Ljubešić, 2016), which is based on character-level statistical machine translation. Following the original papers, we pretrained the normaliser using RIDGES corpus (Odebrecht et al., 2017). As for the ZH character convention, we utilised the popular OpenCC project which uses a hard-coded lexicon table to convert simplified input characters into their traditional forms.

Word embedding. As discussed in § 4, when training DE and ZH monolingual embeddings, we respectively ran subword-based FastText (Joulin et al., 2017) and stroke-based Cw2Vec (Cao et al., 2018). For both languages, we set the dimension at 100 and learned embeddings for all available tokens (i.e., minCount = 1). Other hyperparameters followed the default configurations. After training, we preserved the most frequent 50K tokens in each vocabulary (NB: historical ZH only has 13K unique tokens). To obtain aligned spaces for modern and historical vectors, we then utilised the robust VecMap framework (Artetxe et al., 2018) with its original settings.

Summarisation model. We implemented our main model based on the robust Pointer-Generator Network (See et al., 2017), which is a hybrid framework for extractive (to copy source expressions

---

Footnotes:

2Following the topic taxonomy of Jiang et al. (2020).
3Sampling inspection confirmed that their domains are similar to those of Wanli Gazette.
4https://github.com/jiaeyan/Jiayan
5https://github.com/BYVoid/OpenCC
We assessed all models with the standard ROUGE summarisation models. After setting up (XLM (Lample and Conneau, 2019) is based on ZH (57M sentences) for modern standard Table 4: ROUGE F1 scores (%) of yield their highest possible performance, we trained via pointing) and abstractive (to produce novel ROUGE-1, ROUGE-2, and ROUGE-L. Following precision due to moderate computational power ac-

rameters), we trained them from scratch with FP16 the encoder and decoder (cf. in Step 3 of § 4), we trained them on the enhanced corpora, i.e., normalised DE (NORM) and converted ZH (CONV).

6 Results and Analyses

6.1 Automatic Evaluation

We assessed all models with the standard ROUGE metric (Lin, 2004), reporting F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. Following Hu et al. (2015), the ROUGE score of ZH outputs are calculated on character-level.

As shown in Tab. 3, for DE, our proposed methods are comparable to the baseline approaches or outperform the baselines by small amounts; for ZH, our models are superior by large margins. Given that XLM-based models require a lot more training resources than our model, we consider this a positive result. For comparison of the strengths and weaknesses of the models, we show their performance for a modern cross-lingual summarisation task in Tab. 4. To heighten the contrast we chose two languages (ZH and EN) from different families and with minimal overlap of vocabulary. As shown in Tab. 4, the XLM-based models outperform our method on this modern language cross-lingual summarisation task by large margins.

The difference in the performance of models on the modern and historical summarisation tasks illustrate key differences in the tasks and also some of the shortcomings of the models. Firstly, the great temporal gap (up to 400 years for DE and 600 years for ZH) between our historical and modern data hurts the XLM paradigm, which relies heavily on the similarity between corpora (Kim et al., 2020). In addition, Kim et al. (2020) also show that inadequate monolingual data size (less than 1M sentences) is likely to lead to unsatisfactory performance of XLM, even for etymologically close language pairs such as EN–DE. In our experiments we only have 505K and 992K sentences for historical DE and ZH (cf. § 5.1). On the other hand, considering the negative influence of the error-propagation issue (cf. § 2), the poor performance of XLM–Pipe is not surprising and is in line with observations of Cao et al. (2020) and Zhu et al. (2020). Our model instead makes use of cross-lingual embeddings, including bootstrapping from identical lexicon pairs. This approach helps overcome data sparsity issues for the historical summarisation tasks and is also successful at leveraging the similarities in the language pairs. However, its performance drops when the two languages are as far apart as EN and ZH.

When analysing the ablation results of the proposed method, on DE and ZH we found different trends. For DE, scores achieved by all the four setups show minor variance. To be specific, models bootstrapped with identical word pairs outperformed the unsupervised ones, and models trained on normalised data yielded stronger performance.
Among all tested versions, UspMap+Norm got the best score in ROUGE-2 and IdMap+Norm led in ROUGE-1 and ROUGE-L, indicating that the normalisation enhancement does benefit DE historical text summarisation models. For ZH, as predicted, with richer glyph information encoded, the stroke-based embedding method can better learn word semantics. We find that UspMap+CONV outperforms UspMap and IdMap+CONV outperforms IdMap. Adding identical words during mapping initialisation brings substantial benefits too: 3.58% and 2.52% ROUGE-L improvement for IdMap over UspMap and IdMap+CONV over UspMap+CONV, respectively.

6.2 Human Judgement

To gain further insights, we invited six experts to conduct human evaluations. Like the annotators in § 3.1, they also held degrees in Germanistik or Ancient Chinese Literature. Beyond the standard dimensions of summarisation evaluation (Informativeness, Conciseness, and Fluency), we added ‘Currentness’ as the fourth, which focuses on measuring ‘to what extent a summary follows current rather than early linguistic styles’. We used a five-point Likert scale, with 1 for worst and 5 for best. For each language, experts were only asked to rate the gold-standard human summary and the summaries generated by the XLM-E2E baseline and the best two setups in § 6.1. For each of the 100 news stories in each language, 3 experts independently each rated the three model outputs and the human summary.

The final results are given in Tab. 5. When comparing different systems, we report statistical significance as the p-value of two-tailed t-tests with Bonferroni correction (Dror et al., 2018). We found that in all aspects the scores for the gold-standard summaries were always above 4 points, indicating the high quality of the gold-standard summaries. Across both languages, our models outperform the baseline for informativeness and conciseness (p<0.01) and achieve comparable levels of fluency and currentness. Summaries generated by XLM-E2E were slightly more fluent than our approach for both DE and ZH (p<0.05), indicating that the baseline has merit with respect to its language modelling abilities. However, it tended to make errors in understanding historical inputs and locating key points; e.g. the human reference for ZH article No57 is focused on the commander’s decision of bursting the river to beat the rebel army (‘宁夏之役中，魏学曾为了击溃叛乱部，决定决河灌城’), but XLM-E2E summarises it as 黄河大堤水。比塔顶还高几丈’ (the surface of the river is several feet higher than the tower top), which is fluent but irrelevant.

As for different setups of the proposed algorithm, for DE, in dimensions of Informativeness, Conciseness and Fluency, the performance of UspMap+Norm and IdMap+Norm was almost equally good. The improvement from utilising identical word pairs for cross-lingual word embedding mapping seems more evident for Currentness, i.e., the average score was 0.08 higher (p<0.05). For ZH, while IdMap and IdMap+CONV achieved close Informativeness scores, the latter outperforms the former in other three aspects by 0.08, 0.12, and 0.09 respectively (p<0.01). This observation indicates that when the lexical encoding is improved with enriched stroke-level information, the model is less likely to include redundant information in the summaries (i.e., conciseness score is higher), and the produced sentences are more fluent in terms of modern ZH grammars (see output examples in Appendix A).

6.3 Error Analysis

We further analysed model inputs with the lowest scores in § 6.2, and found that they were mostly for stories whose content was dissimilar to any sample in modern training sets. For instance, five ZH
texts in HISTSUMM are on themes not seen in modern news (i.e., witchcraft (№46), monsters (№35 and №46), and abnormal astromancy (№8 and №28)). On these texts, even the best-performing IdMap+CONV model outputs a large number of [UNK] tokens and can merely achieve average Informativeness, Conciseness, Fluency, and Correctness scores of 1.41, 1.67, 1.83, and 1.60 respectively, which are significantly below its overall results in Tab. 5. This reveals the current system’s shortcoming when processing inputs with theme-level zero-shot patterns. This issue is typically ignored in the cross-lingual summarisation literature due to the rarity of such cases in modern language tasks. However, we argue that a key contribution of our proposed task and dataset is that they together indicate new improvement directions beyond standard cross-lingual summarisation studies, such as the challenges of zero-shot generalisation and historical linguistic gaps (cf. § 3.3).

7 Conclusion and Future Work

This paper introduced the new task of summarising historical documents in modern languages, a previously unexplored but important application of cross-lingual summarisation that can support historians and digital humanities researchers. To facilitate future research on this topic, we constructed the first summarisation corpus for historical news in DE and ZH using linguistic experts. We also proposed an elegant transfer learning method that makes effective use of similarities between languages and therefore requires limited or even zero parallel supervision. Our automatic and human evaluations demonstrated the strengths of our method over state-of-the-art baselines. This paper is the first study of automated historical text summarisation. In the future, we will improve our models to address the issues highlighted in this study (e.g. zero-shot patterns and language change), add further languages (e.g., English and Greek), and increase the size of the dataset in each language.

Acknowledgements

This work is supported by the award made by the UK Engineering and Physical Sciences Research Council (Grant number: EP/P011829/1) and Baidu, Inc. Neptune.ai generously offered us a team license to facilitate experiment tracking.

We would like to express our sincerest gratitude to Qirui Zhang, Qingyi Sha, Xia Wu, Yu Hu, Silu Ding, Beiye Dai, Xingyan Zhu, and Juecheng Lin, who are all from Nanjing University, for manually annotating and validating the HISTSUMM corpus. We also thank Guanyi Chen, Ruizhe Li, Xiao Li, Shun Wang, Zhiang Chen, and the anonymous reviewers for their insightful and helpful comments.

References


Hou Leong Brent Ho and Hilde De Weerdt. 2014. MARKUS. Text Analysis and Reading Platform. Funded by the European Research Council and the Digging into Data Challenge.


Eva Pettersson, J. Lindström, B. Jacobsson, and Rosemarie Fiebranz. 2016. Histsearch - implementation and evaluation of a web-based tool for automatic information extraction from historical text. In *Histoinformatics@DH*.


Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai
Sun, Yechen Xu, Yiming Cui, Cong Yu, Qian-qian Dong, Yin Tian, Dian Yu, Bo Shi, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark.


Die Arbeiten im hiesigen Arsenal haben schon seit langer Zeit nachgelassen, und seitdem die Perser so sehr von den Russen geschlagen worden sind, hört man überhaupt nichts mehr von Kriegsrüstungen in den türkischen Provinzen. Die Pforte hatte nicht geglaubt, daß Rußland eine so starke Macht nach den Ufern des kaspischen Meeres abschicken, und daß der Krieg mit den Persern sobald eine so entscheidende Wendung nehmen würde. Alle kriegerischen Nachrichten, die wir jetzt aus den türkischen Provinzen erhalten, erstrecken sich blos auf die bewaffneten Räuber-Korps, die in der Gegend von Adrianopel noch immer ihren Unfug fortsetzen, der auch wohl nicht eher aufhören wird, bis die Pascha’s selbst bestraft worden sind, die die Räuber beschützen. - Im Anfange dieses Monats erschien eine russische Fregatte am Eingange des schwarzen Meeres, ward durch Sturm vor den türkischen Forts vorbei in den Kanal getrieben, ohne daß die Kommandanten dieser Forts ihr den geringsten Widerstand entgegen stellen konnten, und legte sich, Bujukdere gegenüber, vor Anker. Sobald der Kapitän-Pasha dies erfuhr, verfügte er, daß jene Kommandanten abgesetzt werden sollten, und beschwerte sich bei dem hiesigen russischen Minister darüber, daß jenes Kriegsschiff sich unterstanden habe, wider alle Stipulationen der Traktaten in den Kanal einzulaufen. Nachdem aber der Zufall, wodurch dies geschehen ist, näher aufgeklärt war, widerrief der Kapitän-Pascha die Befehle, die gegen die Kommandanten der an dem Kanal gelegenen Forts erlassen wurden. Auch ward auf Ansuchen des russischen Gesandten der gedachte Fregatte aller mögliche Beistand geleistet, um sich repariren, und nach der Krimm, woher sie gekommen war, zurückkehren zu können. - Die Gesandten, welche die Pforte schon seit 2 Jahren nach Wien und Berlin bestimmt hat, sind noch immer hier; dies beweiset, daß alle Schwierigkeiten in Rücksicht dieser Missionen noch nicht gehoben sind. Der nach Paris bestimmte türkische Gesandte wird aber, wie es heißt, bald abreisen. - Zwei sehr angesehene französische Offiziers, die in türkischen Dienst getreten waren, sind wieder aus demselben entlassen worden.

Wie es zwischen Russland und der Türkei lief, war noch unsicher. (How things would go between Russia and Turkey, was still uncertain.)
Story


(A few days ago there was a member of parliament incognito here from the Royal Family of Sweden, but still unconsciously in some business. On the elapsed Monday, a Currier was sent from their Royal M to the emissaries to monks, Grafen von Königsegg, as people hear, that I. Chur-Fürstl. Durchl. is contesting against I. Royal M, that they help to promote the Royal and Roman Empire interests in every possible way, and that Royal Abgesander who is thinking of such a resolution, wants to continue his journey in the applying Legations Commission, that the same should still remain there for longer, until the Franz. Abgesander might leave again, so that I. Chur-Fürstl. Durchl through the first envisaged delegate does not want to come to other thoughts. The day before yesterday Käys. new Türk interpreter, Mr. Miniski, who was sent to the Vezier afterwards, has come here, from whom people heard that the intended Vezier, like the bases of Erlau and Waraden, were for the accused hiding of the rebels very excited, and denied that they had so far knowingly tolerated their promise against the rebels in their territories, but that such a thing would have happened much more from the Abassi, and thought Vezier had made his previous promise against the IKM. at most let them contest again: but regardless of this sinceration, people know for sure, the contemplated rebels are not only tolerated by the Türken in their areas, but also been armed, and in the recent action the Turks were themselves there on the side of the rebels, so it can be viewed now as a real rupture, which is why at court many patents on new recruitments are issued.)

Expert

Der Kaiser versuchte, durch Verhandlungen seine Interessen gewährzuleisten. Inzwischen boten die Türken wider Versprechen den Rebellen Unterstützung.

(The emperor tried to safeguard his interests through negotiations. Meanwhile Turks broke the promise and provided support to the rebels.)

IdMap+NORM

es befindet sich schon etliche tage her von der crone schweden ein abgeordneter incontum allhier , aber noch unbewusst in was negotio . am verwischen montage ist von ihrer käysr. allda gegen i .

(A few days ago there was a member of parliament incognito here from the Royal Family of Sweden, but still unconsciously in some business. On the elapsed Monday is from their Royalty all against I.)

UspMap+NORM

es befindet sich schon etliche tage her von der crone schweden ein abgeordneter incontum allhier , aber noch unbewusst in was negotio . am verwischen montage ist von ihrer käysr. . m . an den abgestanden

(A few days ago there was a member of parliament incognito here from the Royal Family of Sweden, but still unconsciously in some business. On the elapsed Monday is from their Royal M to the state ...)
Ihre Königl. Majest. befinden sich noch unweit Thorn/ vnd seynd Cosakische Deputirte vnter Wegs/ jhr factum bey Seiner Majest. zu justificiren, vnd wegen ihrer Treu Versicherung zu thun. Von den Fridens. Tractaten zwischen Pohlen vnd Schweden ist noch wenig zu melden. Seithero die Pohlen bey Marienburg den Schweden eine Schantz/ der Kessel genannt/ Abgenommen/ ist nichts weiters vorgefallen/ auch hiesiger Stadt Völcker vor dem Haupt noch nichts tentirt, jedoch sagt man daß noch disse woche etwas vorgehen werde/ so bald nur alle Battereyen in den 3 Quartieren fertig/ vnd die Mörser darauf gebracht worden/ vmb solches mit Feur zu bezwingen/ weil mit dem Schiessen doch nichts zugewinnen/ vnd der Sturm vnnmöglich zu wagen ist/ daß aber das Brau- vnd Proviant Hauß darin in brand geschossen/ vnd die darinnen befindliche Cavallerie also ruinirt werden/ daß sie keinen Aufßfall mehr thun können/ ist gewiß/ deßgleichen hat der Obriste Zaplizky mit 2000. Mann den Elbingern daß Viehe weggetrieben/ welche darauf mit 500. Mann aufgefallen/ solches wieder zu erobern/ seynd aber mehrentheils niedergemacht/ vnd 6. vernommene Offizierer neben vielen Gemeinen gefangen worden. So ist auch auß Churland über Memmel sichere Zeitung einkommen/ daß Herr General Duglas nur 2000. Mann nach Lifland gebracht/ vnd Pautzke sich mit Accord an die Pohlen ergeben habe/ seynd also von den Schweden in Mittau noch 300. Mann übrig/ deren Ergebung man nechstens zu vernehmen hoffet/ zumahlen selbige formaliter beläget seynd/ vnd keinen Succurs zuvermuthen. Den gefangnen Hertzogen von Churland haben die Schweden wider in Lifland nach Revel gebracht/ dahero zu seiner Erledigung Hoffnung gemacht werden will. (Their Royal Majesties are still not far from Torn, and there are Cossack deputies on the way to their factum to be judged by His Majesty, and for their loyalty. (and for their loyal insurance to do that, from the Fridens. Tracts between Poland and Sweden are still little to be reported. Since the Pohlen near Marienburg took away a Schanz, which is called "boiler", from the Swedes, nothing further has happened, also local city peoples have yet in the first place tented nothing, however, they say that something will happen this week, as soon as all batteries are in the 3 quarters ready, and the mortars were brought to it, in order to defeat it with fire, because by shooting nothing could be gained, and the storm is impossible to be venture, but the brown-known and provisions house was set on fire and in it the cavalry were so ruined that they could no longer do any sorties, is certain. Likewise, Colonel Zaplizki has driven away the cattle from the Elbingers with 2,000 men, who with 500 men failed to conquer such, but were mostly killed, and 6 distinguished officers were captured alongside many common ones. It is also sure to be a newspaper from Churland via Memmel coming in, that General Duglas only brought 2000 men to Lifland, and Paucke has surrendered to the Poles by accord, so from the Swedes in Mittau are still 300 men left, whose surrender is the next that people hoped to hear, as they are formally besieged, and no succurs can be expected. The Swedes have brought the captured Duke of Churland back to Revel in Lifland, therefore desired for his completion hope to be made.)

Der Krieg zwischen Polen und Schweden dauert an. Von einem Friedensvertrag ist noch nicht der Rede. (The war between Poland and Sweden continues. Of the peace treaty is there still no talk.)

ihre königl. maiest. befinden sich noch unweit toren, und sind cosakische deputierte unter weg, ihr factum bei seiner maiest. zu justifizirer, und wegen ihrer treu versicherung zu tun. (Their Royal Majesties are still not far from Torn, and there are Cossack deputies on the way to their factum to be judged by His Majesty, and for their loyal insurance to do that.)

ihre königl. maiest. befinden sich noch unweit toren, und sind cosakische deputierte unter weg, ihr factum bei seiner maiest. zu justifizirer, und wegen ihrer treu versicherung zu tun. von den fridens. (Their Royal Majesties are still not far from Torn, and there are Cossack deputies on the way to their factum to be judged by His Majesty, and for their loyal insurance to do that.)

Der Kaiser ließ einrichtungsbezogene, finanzielle, militärische und personelle Anordnungen vornehmen, um den Krieg weiterzuführen.

heute ist der kaiserl. general-krügs-commissarius, graf von nesselbroe, mit dem wegen angeweisung derer künftigen winter-Quartiere abgefassten plan von wie nach dem kaiserl. erbt. quartier heidelberg, zu des prinzen eugenii hoch-fürstl.
Story


(Danube stream from the 13th Weinm. From Wroclaw comes under the 3rd day the following message: The day before yesterday, a great number of locusts have flown over the local city, and their flight lasted from 10 a.m. to around 4 p.m. A column took up almost the whole breadth of the city, and the height was about 130 to 140 cubits. Also, many other columns spread out in great numbers, and it is reported according to the villi that they had also flown through there in great numbers. This vermin lost on its march many of its companions, who were by the crows, ravens, jackdaws, and other birds busily caught, which eat the belly of a locust and its entrails, and let the rest of them fall to the ground, many of which have been seen on local fields. Yesterday again new swarms have arrived here, who didn’t know what to do, but, like the others, continued their flight, and this migration lasts as long as the sun shines bright and warm. In the night rose up a violent wind, which cooled down our previous warm air quite a bit, which is why we today see few locusts. On the property of the Count of Schweidnitz, at Stephandorf, about four miles from here, this vermin has done great damage, since it has eaten up all the wood for the cattle, and the day before yesterday has another indescribably strong army marched over intended goods, which its flight took against Prochwitz and Lignitz.)

Expert
eine große Menge von Heuschrecken sind durchgeflogeen. Ihre Anzahl ist wegen der insektfressenden Vögel und heftigen Wind gesunken.

(A great number of locusts have flown through. Their numbers have decreased due to insectivorous birds and violent winds.)

IdMap+NORM

(Danube stream from the 13th Weinm. From Wroclaw comes under the 3rd day the following message: The day before yesterday, a great number of locusts have flown over the local city, and their flight lasted from 10 a.m. to around 4 p.m. Her evening lasted.)

UspMap+NORM

(Danube stream from the 13th Weinm. From Wroclaw comes under the 3rd day the following message: The day before yesterday, a great number of locusts have flown over the local city, and their flight lasted from 10 a.m. to around 4 p.m.)
(Shaaxi Governor Ye Zhaxiong reported, on March 6th, a yellow colour covered the sky and the day was dark, it slowly turned red, thundered loudly. On the night of the seventh day in the lunar new year, a fire broke down from the sky, big as a basin, with three tails at the back, landed in the northwest. It was discovered that from the 19th year to now, it sometimes thunder, there were many earthquakes, fire and hail, rainstorms and hurricanes, black gas came out of animals’ lips, white gas crisscrossed near buckets, comets flew in the sky, blood gushed from the ground. People all over the country were surprised, there had never been such unusual changes, and in such great number, the report was sent to the emperor.)

(Shaanxi Governor Ye Zhaxiong reported, on March 6th, a yellow colour covered the sky and the day was dark, it slowly turned red, thundered loudly. On the night of the seventh day in the lunar new year, a fire broke down from the sky, big as a basin, with three tails at the back, landed in the northwest. It was discovered that from the 19th year to now, it sometimes thunder, there were many earthquakes, fire and hail, rainstorms and hurricanes, black gas came out of animals’ lips, white gas crisscrossed near buckets, comets flew in the sky, blood gushed from the ground. People all over the country were surprised, there had never been such unusual changes, and in such great number, the report was sent to the emperor.)

(Shaanxi Governor Ye Zhaxiong reported, on March 6th, a yellow colour covered the sky and the day was dark, it slowly turned red, thundered loudly. On the night of the seventh day in the lunar new year, a fire broke down from the sky, big as a basin, with three tails at the back, landed in the northwest. It was discovered that from the 19th year to now, it sometimes thunder, there were many earthquakes, fire and hail, rainstorms and hurricanes, black gas came out of animals’ lips, white gas crisscrossed near buckets, comets flew in the sky, blood gushed from the ground. People all over the country were surprised, there had never been such unusual changes, and in such great number, the report was sent to the emperor.)

(Shaaxi Governor Ye Zhaxiong reported, on March 6th, a yellow colour covered the sky and the day was dark, it slowly turned red, thundered loudly. On the night of the seventh day in the lunar new year, a fire broke down from the sky, big as a basin, with three tails at the back, landed in the northwest. It was discovered that from the 19th year to now, it sometimes thunder, there were many earthquakes, fire and hail, rainstorms and hurricanes, black gas came out of animals’ lips, white gas crisscrossed near buckets, comets flew in the sky, blood gushed from the ground. People all over the country were surprised, there had never been such unusual changes, and in such great number, the report was sent to the emperor.)

(Shaaxi Governor Ye Zhaxiong reported, on March 6th, a yellow colour covered the sky and the day was dark, it slowly turned red, thundered loudly. On the night of the seventh day in the lunar new year, a fire broke down from the sky, big as a basin, with three tails at the back, landed in the northwest. It was discovered that from the 19th year to now, it sometimes thunder, there were many earthquakes, fire and hail, rainstorms and hurricanes, black gas came out of animals’ lips, white gas crisscrossed near buckets, comets flew in the sky, blood gushed from the ground. People all over the country were surprised, there had never been such unusual changes, and in such great number, the report was sent to the emperor.)
(Huangtaiji was granted the title of King Shunyi. Xu Xuemo and others who were officials of the Ministry of Rites said, enemy from the north asked to be offered hereditary peerages, Huangtaiji was Anda’s eldest son, he should inherit the title of king, Sheliketaiji was Huangtaiji’s eldest son, he should inherit the rank of general Longhu. The suggestion was approved.)

(Huangtaiji was Anda’s eldest son Huangtaiji was granted the title of King Shunyi by the Ministry of Rites.)

(Asking to be offered hereditary peerages, Huangtaiji was Anda’s eldest son, do you know?)

(Huangtaiji was Anda’s eldest son and should inherit the rank of general Longhu.)

(The troop in Ningwuguan had an riot which was lead by Lixian.)

(The troop rioted: robbed grain and rice, robbed grain and rice, robbed grain and rice, robbed grain and rice, robbed grain and rice, clamoured to riot!)
論自己無弊國言近日一等浮薄輕進好言喜事之人謹上

己巳，行私公卿大臣勸見掣肘公降。溫旨勉留時行勸以存大

臣之體調用此以安大臣之心各求罷吏科都給事中齊

世臣等都察院左都御史趙錦左副都御史石星等戶書

征免糧田兩湖等處民田如江南各司州縣多著兩湖等處

複行劄下各部院衙門各司州縣著役工

書hud駙馬院馬部院職官俱為當時議者蜂起國是大搖如時待之

翼翼忠勤持巍之侃侃正直中外大小諸臣靡不慕其為