Public Meeting Corpus Construction and Content Delivery

Chenhui Chu[†, a)] / Koji Tanaka[‡] / Haolin Ren[*] / Benjamin Renoust[†, *]
Yuta Nakashima[†] / Noriko Takemura[†] / Hajime Nagahara[†] / Takao Fujikawa[**]
[†] Institute for Datability Science, Osaka University
[‡] Graduate of Information Science and Technology, Osaka University
[*] National Institute of Informatics
[**] Graduate School of Letters, Osaka University

In this paper, we propose a full pipeline of analysis of a large corpus about a century of *public meeting* in historical Australian news papers, from construction to visual exploration. The corpus construction method is based on image processing and OCR. We digitize and transcribe texts of the specific topic of public meeting. Experiments show that our proposed method achieves a F-score of 71.5% for corpus construction. As a result, we built a content search tool for temporal and semantic content analysis.

1. Introduction

Large-scale text corpora are essential for natural language processing. Most existing corpora are created from text that has already been digitized. For instance, the benchmark syntactic parsing dataset *Penn Treebank* [1] is created by labelling part-of-speech tags and syntactic information on the digitized text from the Wall Street Journal newspapers. The parallel corpus *Europarl* [2] that has been used for the machine translation shared task workshop WMT, is created by aligning parallel sentences from the digitized multilingual European Parliament data.

In various fields including literature and humanities, many materials to be studied are not digitized, which are stored in a physical medium such as paper or just scanned but not transcribed into text. By digitizing and transcribing such materials into text, and structuring them via extracting specific topics, we can apply many natural language processing techniques to analyzing them automatically.

In research fields such as literature, digitization, text transcript and structure can significantly increase the value of the original materials.

In this paper, we work on the historical newspaper database Trove[1] (Trove covers major Australian daily newspapers and local newspapers). We propose a corpus construction method based on image processing and OCR, which achieves a high F-score of 71.5%.

We first identify the rule lines in newspaper images and trim the images into newspaper articles. Next, we apply Optical Character Recognition (OCR) to the trimmed articles, and extract the articles with specific topic words. Evaluation conducted on manually annotated golden data indicates that the proposed method can extract 14.9% more articles without excess and deficiency, compared to a baseline that is based on linguistic features extracted from beginning and ending sentences of articles.

We extract articles about the specific topic of *public meeting*[2], which covers 120 years spanning from 19th to 20th centuries. As a result, we develop a tool for content delivery so that we can search and visualize the content of the public meeting articles in semantics and time. Although the proposed method is focused on newspaper data, it is independent from periods and languages and thus can be applied to the corpus construction for historical newspaper data other than Trove.

2. Corpus Construction

The overview of our proposed corpus construction method is shown in Figure 1. We first identify the rule lines in newspaper images, and then trim the rule lines to extract images for articles. Next, we apply OCR to the extracted article images to extract text for the articles. Finally, we filter the articles with a query phrase to filter the articles and thus extract only the target articles that we are interested.

2.1. Trimming

We use OpenCV[3] for identifying the rule lines in newspaper images and trimming. Firstly, we binarize the newspaper images using the method proposed by Ohtsu [3]. The binarization method

a) chu@ids.osaka-u.ac.jp

[1] https://trove.nla.gov.au

[2] Public meeting is the main pillar of public opinion formation for Western Europe in the 19th century.

[3] https://opencv.org/

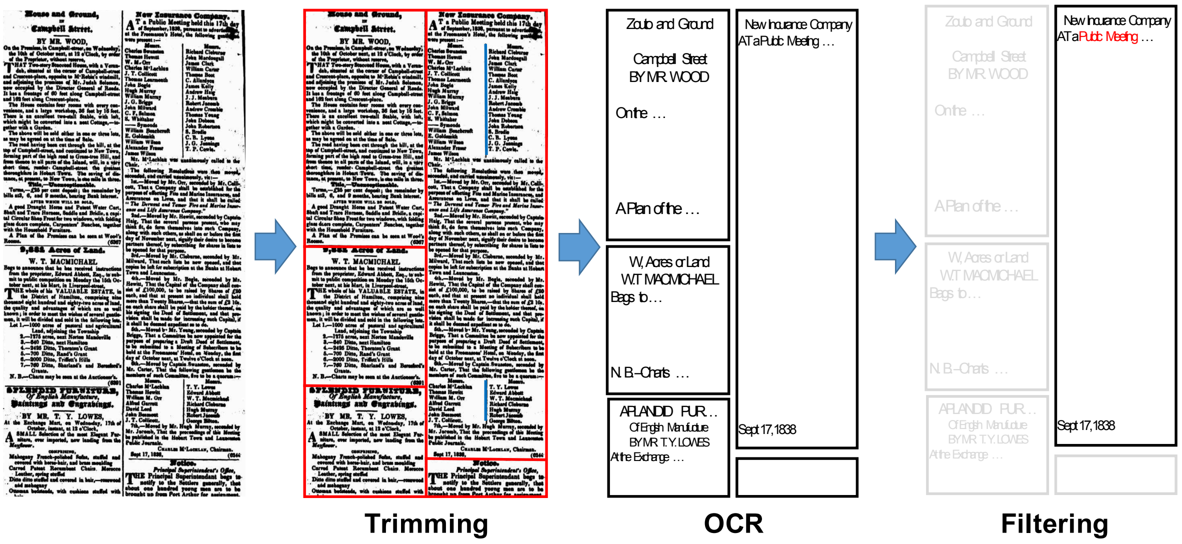


Figure 1 Overview of the corpus construction method.

transfers grayscale images to white-black images by calculating the threshold that maximizes the separation degree from the histogram of picture element numbers. Next, we apply the contour tracking processing algorithm of Suzuki [4] to extract the contours from the binarized images. In order to identify the contours, this algorithm calculates the boundary of the binarized images and sequentially detects the pixels that are the contour counterclockwise. Areas with a height above a threshold and a width below a threshold is identified as a column, and areas with a width above a threshold and a height below a threshold is identified as an article split in the newspaper image. The thresholds are tuned manually. We can finally trim the article images accordingly. There are small columns in articles as the blue lines shown in Figure 1. To deal with this, we propose the following method to determine the vertically split column. Firstly, we trim the column with the x coordinate (horizontal direction) value. We then compare the minimum and maximum y coordinate (vertical direction) values with the newspaper coordinate value. If the difference is above a predefined threshold, we determine it as a small column and do not use it for trimming.

2.2. OCR

OCR is generally performed following the procedures of character delimiter recognition, size normalization, feature extraction, and classification. Google open-sources the OCR method Tesseract [5], which achieves 98.4% and 97.4% on newspaper articles in character and word level, respectively. However, after comparing the OCR accuracy of Google Drive[1]with Tesseract, we find that Google Drive works better. Therefore, we use the OCR function of Google Drive for extracting text from the article images.

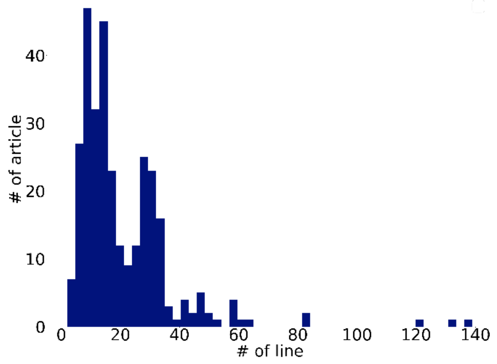


Figure 2 Line number distribution of the golden article data.

2.3. Filtering

We filter the OCRed articles that are not our target with a query phrase, leaving the target articles to be extracted. In order to allow the error of character recognition by OCR, we define similarities in character level. We use the Python *difflib* module SequenceMatcher[2] for calculating similarities. In SequenceMatcher, the similarities between a character string pair is defined as:

$$Similarity = (2.0 * M)/T$$

[1] https://www.google.com/intl/ja_ALL/drive/

[2] https://docs.python.jp/3/library/difflib.html

where M is the number of matched characters and T is the sum of character numbers in the character string pair.

We get word n-grams from the articles according to the number of words in the query character string. We then calculate the similarity between the n-gram and query character string, and take the articles with the highest similarity above a threshold as the target article. The threshold is tuned on a development set, which shows the highest F1-score.

3. Experiments

3.1 Data

We used the newspaper image data crawled from Trove[1], and the targeted articles were the ones containing the key phrase "public meeting". Trove is an online library database service maintained by the Australian government. We searched the key phrase "public meeting" on Trove to get the newspaper IDs. Next we obtained the newspaper pdf data through the API provided by Trove with the newspaper IDs. As OpenCV cannot handle pdf files, we converted the newspaper pdf data to PNG with ImageMagick[2].

In our experiments, we manually extracted 307 articles about "public meeting" spanning from 1838 to 1954, and split them into 149 and 158 articles for development and testing, respectively. Figure [2] shows the line number distribution of the golden data used for our evaluation.

Table 1 Article extraction evaluation results.

Method	Precision	Recall	F1-score
Baseline	**76.1**	56.3	64.7
Proposed	59.4	51.9	55.4
Baseline+Proposed	56.4	**97.5**	**71.5**

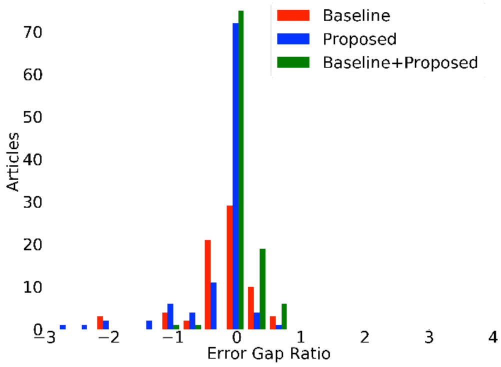


Figure 3 Line level evaluation results (beginning line).

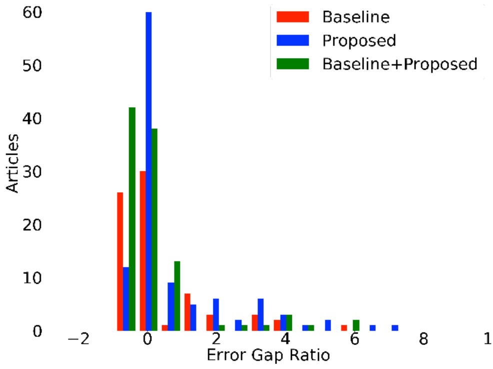


Figure 4 Line level evaluation results (ending line).

3.2 Comparison

We compared the following methods in our experiments:

- Baseline: We compared a baseline method, that is based on text features to identify articles directly from the OCRed text provided by the Trove website. The baseline method extracts features from the beginning and ending sentences of articles for article identification. The features are as follows:
  - Beginning sentence: Take 2 sentences before the sentence that contains "public meeting".
  - Ending sentence: We first apply name entity recognition using the Stanford parser[3]. Then we take the sentence containing LOCATION, DATE, PERSON tags, but the following sentence that does not contain these tags as the ending sentence.
- Proposed: This is our proposed method presented in Section 2.
- Baseline+Proposed: Use our proposed method for articles which the baseline model fail to extract because the end sentence corresponding to the start sentence is not found.

3.3 Parameter Tuning

To tune the thresholds for the rule line and small column identification described in Section 2.1, we used the newspaper data spanned in one month and determined the thresholds empirically. For the threshold used for filtering as described in Section 2.3, we tuned it on the development data and chose

[1] https://trove.nla.gov.au

[2] https://www.imagemagick.org/

[3] https://nlp.stanford.edu/software/lex-parser.shtml

the one achieving the highest F1-score. We tuned the threshold from 0 to 1 with an increment of 0.05, and it turned that 0.8 was the best and thus we used the threshold of 0.8 for filtering.

3.4 Evaluation Methods

In our experiments, we conducted article level evaluation to evaluate if the articles are successfully extracted. In addition, we also conducted line level evaluation to evaluate the accuracy for article extraction. These two evaluation methods are described as follows:

Article level evaluation method

We calculated the similarity following Equation 1 between the sentences containing the keyword "public meeting" in the extracted article and golden article, respectively. If the similarity is higher than a threshold then the extraction is evaluated as success, otherwise it is failure. The threshold was empirically determined to be 0.6. Then we calculated the precision, recall, and F1-score for the baseline and proposed method.

Line level evaluation method

We compared the beginning and ending lines of the extracted articles to the golden articles to investigate the difference. Then we calculated the ratio of excess and deficiency lines between the extracted and golden articles.

3.5 Results

Article level evaluation

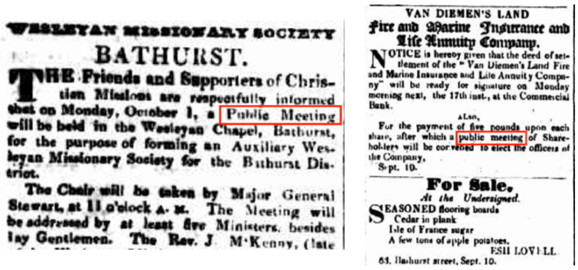
Table 1 shows the results for article level evaluation.

We can see that Baseline has a higher F1-score than Proposed. The reason for this is that Baseline uses the feature of a sentence that includes ``public meeting" when getting the first lines of the article, and thus the sentence used for article level evaluation is extracted. However, there are still some failures in the extraction of Baseline. This is because firstly there can be multiple public meeting articles in a newspaper image, secondly there are OCR errors about the keyword ``public meeting."

After combing our proposed method with the baseline method, the article extraction results are improved.

Line level evaluation

Figures 3 and 4 show the line level evaluation results for the beginning and ending lines, respectively. The horizontal axis represents the gap ratio of the number of the excess and deficiency lines against the entire number of lines in an article. The vertical axis represents the number of articles. We can see that on both the beginning and ending lines, Proposed extracted significantly more articles without excess and deficiency than Baseline. In addition, for the case of articles without excess and deficiency in both the beginning and ending lines, Baseline only successfully extracted 5 (3.1%) articles but Proposed extracted 19 (18.0%) articles. Therefore, we can say that the proposed method that uses visual features to identify the article split, is more effective for extraction articles with specific topics.



WESLEYAN MISSIONARY SOCIETY

BATHURST.

THE Friends and Supporters of Christian Missions are respectfully informed that on Monday, October 1, a Public Meeting will be held in the Wesleyan Chapel, Bathurst, for the purpose of forming an Auxiliary Wesleyan Missionary Society for the Bathurst District.

The Chair will be taken by Major General Stewart, at 11 o'clock A. M. The Meeting will be addressed by at least five Ministers, besides lay Gentlemen. The Rev. J M'Kenny, (late

VAN DIEMEN'S LAND

Fire and Marine Insurance and Life Annuity Company.

NOTICE is hereby given that the deed of settlement of the "Van Diemen's Land Fire and Marine Insurance and Life Annuity Company" will be ready for signature on Monday morning next, the 17th inst., at the Commercial Bank.

ALSO,

For the payment of five pounds upon each share, after which a public meeting of Shareholders will be convened to elect the officers of the Company,

Sept. 10.

For Sale,

At the Undersigned.

SEASONED flooring boards
Cedar in plank
Isle of France sugar
A few tons of apple potatoes.

ESH LOVELL

63, Bathurst street, Sept. 10.

Figure 5 Article extraction examples

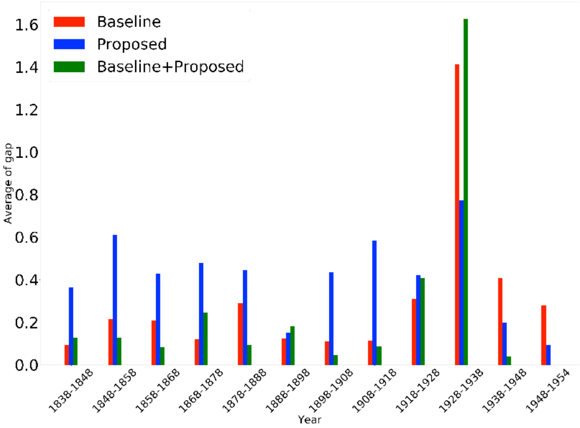


Figure 6 The average of the Line level evaluation results in each year.

3.3 Discussion

Figure 5 (left) shows an example of an article that failed to be extracted. The ``public meeting" area surrounded by the red rectangle in the image was incorrectly recognized as "Pohlle Meeling" by OCR. Therefore this article was not identified as a target article during filtering and thus failed to be extracted. This happens because the resolution of this newspaper article was lower than others, and thus OCR failed. Among all the test data, 8.1% of the articles were failed to be extracted due to OCR errors.

Figure 5 (right) shows an example of an article that was successfully extracted but with some excess. We can see that the trimmed newspaper image contains not only a target article but also an article that is not our target. The reason for this is that the ruled line has been cut off in the middle, which makes the identification of the article split fail, leading to the improper trimming. Among all the test data, 26.7% of the articles were successfully extracted but with some excess like this example.

There were also extracted articles with some deficiency compared to the golden articles. One reason for this is that the trimming algorithm treated some paragraph splits as article splits. There is 8.1% of this type of error among the test data. In addition, the proposed method cannot deal with an article spanning multiple columns, making this type of article being split into multiple articles. There is 1.9% of this type of error among the test data.

Figure 6 shows the average of the Line level evaluation results in each year (divided from 1838 to 1954 every 10 years). We can see that the average of gap ratio is large from 1928 to 1938 in all methods. This is because, as newspapers are new, the variation of advertising design increases (like Figure 7) and the accuracy of our trimming method decreases.

4 Content Delivery

Using the "Baseline+Proposed" method, we extracted 269,044 public meeting articles spanning from 1838 to 1954 from Trove. To provide exploration of the content, we extended the search engine Visual Cloud [6] so it may support a much larger number of documents to search. We extract name entities from each article using Stanford CoreNLP[1]. Each article represents then a document which is indexed by its list of named entities.

The Visual Cloud provides a full search engine, with an interactive timeline and tag cloud, as illustrated in Figure 6: upon a query, here "*meeting*", the results are placed on a timeline, a word cloud describes the semantic content of the search results, and individual access to each result is possible.

5 Conclusion

In this paper, we constructed a corpus of public meeting articles via image processing and OCR.

Experiments conducted on the newspaper data from Trove indicated that we can successfully extract 97.5% of the targeted articles and 18.0% of the extracted articles are without excess and deficiency.

THE EXAMINER, TUESDAY, JANUARY 1, 1929.

Holiday Sporting Fixtures

THE WEEKLY COURIER

The King of New Year Gifts!

SOUVENIR PLAYING CARDS

of the Queen of Tourist Resorts–TASMANIA!!

Every Card a Different Picture of Tasmania's Beauty Spots

Get a Packet for Yourself and Friends

Price - - - - 5/-

Obtainable at all Stationers

See the result of your Photography the Same Day

HATTON AND LAWS

ROBUR

Your Holiday Snaps

Spurlings

Do you use powder?

Figure 7 Newspaper example in 1928 (has an unusual design)

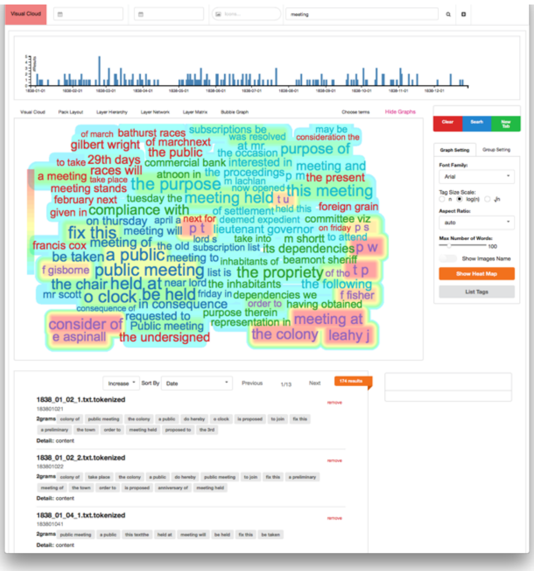


Figure 8 Search and visualization of the public meeting corpus.

We further enabled content delivery of the public meeting articles through search and visualization. In the future, we plan to apply and verify our methods to historical materials in other fields.

This work was supported by JSPS Kakenhi 19H01330.

[1] https://stanfordnlp.github.io/CoreNLP/index.html

References

[1] Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank, Computational Linguistics, Vol. 19, No. 2, pp. 313–330 (online), available from ⟨http://dl.acm.org/citation.cfm?id=972470.972475⟩ (1993)

[2] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation, Proc. of Machine Translation Summit, pp. 79–86 (online), available from ⟨http://mt-archive.info/MTS-2005-Koehn.pdf⟩ (2005).

[3] 大津, 展.: 判別および最小 2 乗規準に基づく自動しきい値選定法, 電子通信学会論文誌 D, Vol. 63, No. 4, pp. 349‐356 (online), available from ⟨https://ci.nii.ac.jp/naid/40002557720/⟩ (1980).

[4] Suzuki, S. and Abe, K.: Topological structural analysis of digitized binary images by border following, Computer Vision, Graphics, and Image Processing, Vol. 30, No. 1, pp. 32–46 (online), DOI: https://doi.org/10.1016/0734-189X(85)90016-7 (1985).

[5] Smith, R.: An Overview of the Tesseract OCR Engine, Proc. of International Conference on Document Analysis and Recognition, Vol. 2, pp. 629–633 (online), DOI: 10.1109/ICDAR.2007.4376991 (2007).

[6] Ren, H., Renoust, B., Viaud, M.-L., Melanc¸on, G. and Satoh, S.: Generating “visual clouds” from multiplex networks for tv news archive query visualization, 2018 International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, pp. 1–6 (2018).