Clustering-Based Article Identification in Historical Newspapers

Martin Riedl and **Daniela Betz** and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{riedlmn,pado}@ims.uni-stuttgart.de,mail@danielabetz.de

Abstract

This article focuses on the problem of identifying articles and recovering their text from within and across newspaper pages when OCR just delivers one text file per page. We frame the task as a segmentation plus clustering step. Our results on a sample of 1912 New York Tribune magazine shows that performing the clustering based on similarities computed with word embeddings outperforms a similarity measure based on character n-grams and words. Furthermore, the automatic segmentation based on the text results in low scores, due to the low quality of some OCRed documents.

1 Introduction

Historical newspapers are among the "most important" and "most often used" sources for many historians (Tibbo, 2003): Since the rise of regional and local newspaper culture in the late 18th and early 19th centuries, newspapers provide a window into national and global events and debates as well as into local everyday life (Slauter, 2015).

Traditionally, historical newspapers were stored on microfilms in local archives. Access was manual, required travel and authorization, and was often complicated by poor film quality (Duff et al., 2004). Digital availability of newspapers has scaled up the accessibility of historical newspapers tremendously and enabled large-scale analysis of phenomena like text re-use (Smith et al., 2015) or ethnic stereotyping (Garg et al., 2018).

Digital access to the full range of information in a newspaper is challenging, though. It requires (a), scanning of newspaper pages or microfilms into digital image files; (b), optical character recognition (OCR) to transfer images into text streams; and (c), identification of articles in the text stream.[1] Few historical newspapers have gone through all steps. For example, the vast Chronicling America archive of historical newspapers at the Library of Congress[2] only underwent steps (a) and (b), providing text files at the level of newspaper pages, without manual OCR post-correction (see Figure 1).

Due to the multi-column format of almost all newspapers, each text file contain multiple articles. In addition, many articles span several pages: they are split across text files. This is an obvious obstacle to any analysis requiring complete articles. It becomes particularly pressing for articles that span multiple issues (typically days or weeks). Notable among them are *serial stories* or *serial novels*, serialization being among the most important publication strategies for literary works in the 19th and 20th centuries (Lund, 1993).

In this paper, we investigate the task of *article identification across newspaper pages*, corresponding to step (c) above. We use only textual information from OCR as input, modelling the task as a sequence of a segmentation and a clustering step. Whereas most previous work solely uses image data for similar tasks, here, we examine the performance of an approach that uses textual information only. We introduce and provide a new annotated dataset sampled from the 1912 New York Tribune magazine. We find that clustering segments works relatively well for individual issues and becomes substantially more difficult across issues. Segment similarity based on word embeddings outperforms character n-grams similarities for most cases. The major challenge of the task is mainly the inferior scan quality which results in poor OCR text output.

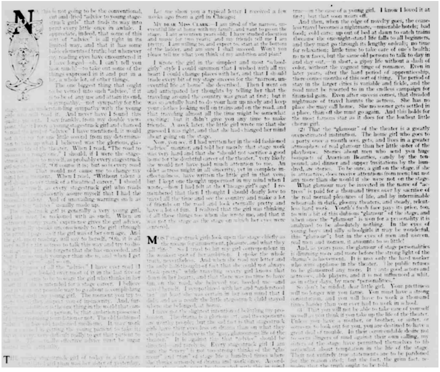
2 Related Work

The task tackled in this paper can be split into two sub-tasks: the detection of the different articles and the clustering of parts of the same article.

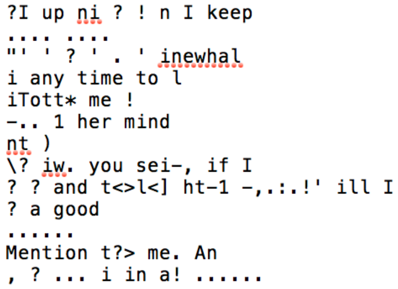
[1] In this paper, we ignore the issue of metadata extraction.

[2] https://chroniclingamerica.loc.gov

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 12–17
Minneapolis, MN, USA, June 7, 2019. ©2019 Association for Computational Linguistics



(a) newspaper page



```
?I up ni ? ! n I keep
.... ....
"' ' ? ' . ' inewhal
i any time to l
iTott* me !
-.. 1 her mind
nt )
\? iw. you sei-, if I
? ? and t<>l<] ht-1 -,.:.!' ill I
? a good
......
Mention t?> me. An
, ? ... i in a! ......
```

(b) text sample

Figure 1: Historical newspaper page with OCR output

Most previous work performs the segmentation of newspaper pages directly at the image level (Hebert et al., 2014; Meier et al., 2017). This strategy avoids having to deal with spelling errors arising from OCR. However, these methods are not applicable when only textual output is available.

A different line of research addresses the detection of segments in texts. Often, contemporary newspaper texts, Wikipedia articles or novels are artificially merged (e.g. Choi, 2000; Galley et al., 2003). Most of these methods are based on similarities between adjacent sentences or segments. The similarities are mostly computed using words (Hearst, 1997; Choi, 2000) or dense vector representations like topic models (Bestgen, 2006; Riedl and Biemann, 2012) or embeddings (Alemi and Ginsparg, 2015).

Another related task is genre classification, in particular for newspaper texts. Lorang et al. (2015) present a classifier for detecting poetic content, which is however based again on images and incorporates image preprocessing techniques. Lonij and Harbers (2016) build a general genre classifier for text spans, but only for historical Dutch newspapers. A general limitation of this approach is that the articles which we want to separate may not differ in gender: this is often true (e.g., editorial content in the middle with advertisements on the side) but not always (e.g., multi-column pages such as title pages).

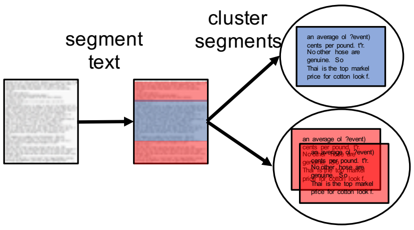

segment text
cluster segments
an average ol ?event) cents per pound. t'r. No other hose are genuine. So Thai is the top markel price for cotton look f.
an average ol ?event) cents per pound. t'r. No other hose are genuine. So Thai is the top markel price for cotton look f.
an average ol ?event) cents per pound. t'r. No other hose are genuine. So Thai is the top markel price for cotton look f.

Figure 2: Overview of the method for detecting and merging serial stories

At the textual level, article identification is related to author identification (Stamatatos, 2009) and style breach detection (Tschuggnall et al., 2017), which group texts by author. However, these settings typically do not attempt grouping at the story level and use predefined lists of authors. Also, noisy texts are generally not considered.

3 Method

Recall that in this article we have the goal of turning a collection of (textual) newspaper pages into a collection of (textual) articles.

We follow the intuition that articles should be recoverable through *coherence* at multiple levels. Not only are articles *semantically* coherent in terms of vocabulary and names by virtue of typically covering one topic, but they are also *stylistically* coherent since they are typically written by one author. We operationalize this intuition by recovering articles through *semantic clustering* of text segments.

The most straightforward type of text segment provided by historical newspapers is the individual line. However, multi-column layouts lead to very short lines which are too information poor for reliable clustering. Therefore, we adopt a two-step procedure as shown in Figure 2: We first subdivide the pages into *segments* (stretches of text that presumably belong to the same article). Then, we cluster segments within and across pages to assign all segments of the same article in one cluster.

Text Segmentation. TextTiling (Hearst, 1997) is based on the intuition that chunks that are semantically coherent use a similar vocabulary. First the document is segmented into sentences and tokens. In the next step the lexical similarity between two neighboring blocks of $b = 10$ sentences is computed. TextTiling computes lexical similarities of pairs of adjacent blocks around the i-th gap, s_i, as the cosine similarity between the lexical distributions of both blocks. Plotting these scores, TextTiling assumes that minima within this line indicate also segmentation boundaries. In order to find segmentation boundaries, a depth score, $D_i = (s_{i-1} - s_i) + (s_{i+1} - s_i))$, is computed and local minima are selected.

Segment Clustering. Subsequently, we cluster the segments into articles. In this study, we focus on semantic similarity among segments and do not take positional information into account. We use a simple but powerful clustering method, spectral clustering (Ng et al., 2002). Spectral clustering applies k-means not to the original similarity matrix, but to a dimensionality-reduced version, increasing expressiveness and robustness of the method. Thus, we first build the matrix by computing similarity scores between all segments. Based on this matrix, we then perform the spectral clustering.

Two measures of pairwise segment similarity appear particularly appropriate for OCRed, and thus noisy, texts. The traditional one is the similarity of words or character n-gram distributions, using the Jaccard coefficient.

We hypothesize, that due to OCR errors, character n-grams might work better than using complete words. Thus, we compute the Jaccard coefficient on words as well as on character n-grams (n=2–8). A more recent approach is using the cosine similarity between 200 dimensional embeddings defined as centroids of their fastText word embeddings (Bojanowski et al., 2017). Using fastText we benefit from the functionality that embeddings can be generated from out-of-vocabulary words.

4 Dataset

To our knowledge, there is no standard dataset for article identification in historical newspapers.[3] Thus, we created such a dataset.

We selected the five March 1912 issues of the *New York tribune* Sunday magazine[4] for annotation since this dataset contains long articles, some but not all of which are serializations that extend over multiple issues. We annotated a total of 82 pages.

The annotation was performed by three annotators so that each page was annotated by two different annotators. We annotated each segment in the OCR output, marking it either as part of an article with a unique ID, or as an advertisement.

The high number of short advertisements, combined with the low OCR quality due to very small and artistic typesetting, led to high disagreement on the segmentation annotations. Since our focus is on articles, we merged all advertisement blocks. The resulting annotation achieves a Cohen's (Cohen, 1960) kappa score of κ = 0.85, (”almost perfect” agreement). Subsequently, we manually checked the disagreements and merged the annotations.[5]

In the following experiments, we consider either all pages of one issue (ByIssue setting), or all pages of all issues (AllIssues setting). The ByIssue dataset contains an average of 37 gold segments corresponding to 12.6 articles. The AllIssues dataset consists of 53 different articles split among 185 gold segments — i.e., we have an average of 3 to 4 segments per article.

5 Experimental Setup

Preprocessing. We remove all non-alphanumeric characters and transform similarities exponentially for clustering. The fastText embeddings are trained on all 1912 English-language newspapers available from Library of Congress.

Design. We conduct two experiments. In the first experiment, we use our gold standard (manually annotated) segment boundaries and perform only clustering. This setup reveals the performance of the clustering method. The second experiment adopts a more realistic setting and evaluates clustering performance when using automatically predicted segments obtained by TextTiling.

Evaluation. In the first experiment, only the clustering needs to be evaluated. For the evaluation, we rely on the B-cubed measure, an adaptation of the familiar IR precision/recall/F_1 measure to the clustering setup (Bagga and Baldwin, 1998). In the second experiment, we additionally evaluate automatic segmentation, for which we report precision and recall. Using this measure is motivated as when using automatic text segmentation as a preprocessing step, we prefer high recall, resulting in fine-grained

[3] The National Library of the Netherlands (https://www.kb.nl/en) gives access to Dutch newspaper and also provides a classifier to detect different genres. However, they do not detect articles crossing pages and avoid advertisements.

[4] This data is made available as PDF and text by the Library of Congress via Chronicling America: http://chroniclingamerica.loc.gov/

[5] The annotation and source code is published at: https://github.com/riedlma/cluster_identification.

Similarity		n	B-Cubed Prec.	 Rec.	 F1
Cosine fastText			0.6983	0.6316	0.6591
Jaccard	n-gram	2	0.5335	0.5349	0.5298
		3	0.5621	0.5343	0.5432
		4	0.6153	0.5595	0.5824
		5	0.6234	0.5507	0.5813
		6	0.6634	0.5698	0.6097
		7	0.6774	0.5712	0.6158
		8	0.6576	0.5510	0.5963
	word		0.6880	0.5905	0.6328

Table 1: Effect of similarity measure on clustering performance for a fixed number of clusters of 12 (ByIssue setting, gold standard segmentation)

Sim.	Cl.	B-Cubed Prec.	 Rec.	 F1
Jaccard word	10	0.6290	0.6063	0.6139
	11	0.6511	0.5870	0.6148
	12	0.6880	0.5905	0.6328
	13	0.7053	0.5749	0.6296
	14	0.7213	**0.5659**	0.6315
	15	**0.7427**	0.5565	**0.6330**
Jaccard 7-gram	10	0.6162	**0.5790**	0.5927
	11	0.6519	0.5737	0.6060
	12	0.6774	0.5712	0.6158
	13	0.6938	0.5626	0.6177
	14	0.7063	0.5543	**0.6185**
	15	**0.7096**	0.5424	0.6120
Cosine fastText	10	0.6161	0.6276	0.6176
	11	0.6523	0.6342	0.6387
	12	0.6983	0.6316	0.6591
	13	0.7270	**0.6371**	0.6757
	14	**0.7504**	0.6309	**0.6810**
	15	0.7485	0.6095	0.6671

Table 2: Experiment 1: Article identification with gold standard segments, ByIssue setting

segments. Due to the non-deterministic nature of the spectral clustering, we perform each clustering run 5 times and report averages.

6 Results

6.1 Experiment 1: Gold boundaries

First, we inspect the effect of computing similarity in different ways for the ByIssue setting for 12 clusters, the average number of articles per issue (cf. Section 4). The results in Table 1 show that among the Jaccard-based similarities, there is an interesting tendency for relatively long n-grams to work well, with the best results for n=7. Furthermore, in contrast to our intuition that the word level would suffer from OCR errors, we see better results for words than for n-grams. The overall best results are achieved by Cosine similarity on fastText embeddings which can be understood as an optimized combination of word and character n-gram information.

Next, we vary the number of clusters and retain the three best-performing similarity measures. (The analysis shown in Table 1 is robust across numbers of clusters). For the ByIssue setting (see Table 2), we consider between 10 and 15 clusters. We find that Precision generally increases with increased number of clusters, while Recall decreases, as could be expected. The maximum F1 score of just above 68% is obtained for cluster sizes of 14 (fastText-based and 7-gram similarities) and 15 (word-based similarity). This corresponds closely to, and is a bit higher than, the average number of gold clusters in that dataset (viz., 12.6). Embedding-based similarity outperforms trigram-based similarity by about 2.8 points F1.

In the AllIssues setting, we expect to see around 53 articles and thus explore performance between 50 and 55 clusters (see Table 3). The F1 scores are generally lower than for the ByIssue setting, but still substantial. We find similar tendencies as before (Precision increasing and Recall decreasing with the number of clusters). However, there is more variance than in the ByIssue setting, so the patterns are less clear. We achieve best performance for 7-gram-based similarity with 55 clusters, for the word-based similarity with 54 and for embedding-based similarity with 54 clusters. The best performing number of clusters is again close to, and a bit higher than, the true number of articles. Here, also the 7-gram Jaccard similarity performs better than using words and is essentially on par with the fastText embeddings. We interpret this finding as showing that long n-gram shared between segments (e.g. person names, place names, etc.) are a surprisingly good indicator of article identity, even in the face of noisy OCR output.

6.2 Experiment 2: Automatic boundaries

We first evaluate TextTiling, our automatic segmentation method (cf. Section 3) and find a low Precision (0.1168) but a comparatively high Recall (0.6602). This means that precise segmentation of the noisy, OCRed historical texts is challenging indeed: TextTiling over-segments the texts. This happens, for example, when parts of a page “look different” in a scan (e.g. due to folds) and OCR introduces systematically different errors. We still prefer over- to under-segmentation, since over-

Sim.	Cl.	B-Cubed Prec.	 Rec.	 F1
Jaccard word	50	0.5581	0.4313	0.4865
	51	0.5618	0.4340	0.4896
	52	0.5645	0.4467	0.4986
	53	**0.5705**	**0.4493**	**0.5026**
	54	0.5622	0.4435	0.4957
	55	0.5608	0.4503	0.4995
Jaccard 7-gram	50	0.5930	0.4753	0.5274
	51	0.5843	0.4668	0.5189
	52	0.6045	0.4968	0.5451
	53	0.6116	0.4796	0.5376
	54	0.6059	0.4773	0.5339
	55	**0.6214**	**0.5010**	**0.5546**
Cosine fastText	50	0.5917	**0.5085**	0.5466
	51	0.5878	0.4876	0.5328
	52	0.5876	0.4746	0.5251
	53	0.5798	0.4751	0.5221
	54	**0.6246**	0.4927	**0.5506**
	55	0.6064	0.4839	0.5381

Table 3: Experiment 1: Article identification with gold standard segments, AllIssues setting

	Sim.	Cl.	B-Cubed Prec.	 Rec.	 F1
BI	JC Word	15	0.4363	0.2125	0.2843
	JC 7-gram	14	0.4631	0.3313	0.3857
	Cos. fastText	14	**0.6168**	**0.3650**	**0.4563**
AI	JC Word	53	0.2442	0.0923	0.1339
	JC 7-gram	55	0.2726	0.1884	0.2228
	Cos. fastText	54	**0.4409**	**0.2105**	**0.2848**

Table 4: Experiment 2: Article identification with automatic segments (AI: AllIssues, BI: ByIssue)

segmented articles stand a chance of being recombined in the clustering step.

Table 4 shows the results for article identification on automatically segmented text (we report only results for the previously best numbers of clusters). As can be expected given the segmentation results, performance drops substantially compared to Experiment 1. What is notable is the difference between the ByIssue and the AllIssues settings: For ByIssue, performance drops moderately from 0.68 to 0.46 F1, while for AllIssues we see a huge decrease from 0.55 to 0.28 F1. Similarity behaves consistently: fastText performs best for both settings, while word-based similarity yields the lowest scores.

6.3 Discussion

Sim.	min. OCR quality ≥-1.0	 ≥0.0	 ≥0.5
Jaccard Word	0.6315	0.6491	0.7133
Jaccard 7-gram	0.6185	0.6628	0.7252
Cosine fastText	0.6810	0.7008	0.7629
# of pages	82	74	55

Table 5: Article identification on pages filtered by OCR quality (Exp. 1, ByIssue, B-Cubed F1, 14 clusters)

The results of our experiments show that processing historical newspaper is a challenging task, due to the high variance of the OCR quality. Sometimes, pages are hardly readable (cf. Figure 1); on other pages, the quality varies greatly among sections.

We further investigated the impact of OCR quality by annotating each page with an OCR quality indicator on a four-point Likert scale (-1: unusable, 0: bad, 1: medium, 2: good), averaging over two annotators. Then, we repeated the ByIssue setting of Exp. 1 with 14 clusters, including only pages with a quality at or above different thesholds.

Table 5 shows the results. Even though performance might be expected to decrease for filtered datasets since the fixed number of clusters becomes less appropriate, it mostly remains similar (0.0) and improves using a threshold of 0.5.[6] This shows that OCR is indeed a leading source of problems.

7 Conclusion

This paper has introduced a new dataset for the text segmentation and identification of articles in historical newspapers with OCR-induced noise. We have shown results for two tasks: a) article segmentation and b) article clustering. Overall, results are promising for clustering based on gold standard segmentation, but degrade significantly when segmentation is performed automatically. This indicates manual segmentation, which involves much less effort than OCR postcorrection, is a worthy target when some manual annotation resources are available. Arguably, segmentation can also be improved further by the inclusion of visual features (Meier et al., 2017), which appears a promising direction for future research.

Acknowledgments

This study was supported by the DFG grant Oceanic Exchanges (KO 5362/2-1) and the CRETA center funded by the German Ministry for Education and Research (BMBF).

[6] We cannot apply any higher threshold filtering, as for some issues the number of clusters is higher than the number of sentences, i.e. possible segment boundaries.

References

Alexander A. Alemi and Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of COLING*, pages 79–85, Montreal, Canada.

Yves Bestgen. 2006. Improving text segmentation using latent semantic analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, pages 26–33, Seattle, WA.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Wendy Duff, Barbara Craig, and Joan Cherry. 2004. Finding and using archival resources: A cross-canada survey of historians studying canadian history. *Archivaria*, 58(0):51–80.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*, pages 562–569, Sapporo, Japan.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

David Hebert, Thomas Palfray, Stephane Nicolas, Pierrick Tranouez, and Thierry Paquet. 2014. Automatic article extraction in old newspapers digitized collections. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pages 3–8, Madrid, Spain.

Juliette Lonij and Frank Harbers. 2016. Genre classifier. KB Lab: The Hague. `http://lab.kb.nl/tool/genre-classifier`.

Elizabeth Lorang, Leen-Kiat Soh, Maanas Varma Datla, and Spencer Kulwicki. 2015. Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Magazine*, 21(7/8).

Michael Lund. 1993. *America's Continuing Story: An Introduction to Serial Fiction, 1850–1900*. Wayne State University Press.

B. Meier, T. Stadelmann, J. Stampfli, M. Arnold, and M. Cieliebak. 2017. Fully convolutional neural networks for newspaper article segmentation. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, pages 414–419.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. 2002. On spectral clustering: Analysis and an algorithm. In *Proceedings of NIPS*, pages 849–856.

Martin Riedl and Chris Biemann. 2012. Text Segmentation with Topic Models. *Journal of Language Technology and Computational Linguistics*, 27(47-69):13–24.

Will Slauter. 2015. The rise of the newspaper. In Richard R. John and Jonathan Silberstein Loeb, editors, *Making News: The Political Economy of Journalism in Great Britain and the United States from the Glorious Revolution to the Internet*, pages 19–46. Oxford University Press.

David A. Smith, Ryan Cordell, and Abby Mullen. 2015. Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, 27(3):E1–E15.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3).

Helen Tibbo. 2003. Primarily History in America: How U.S. historians search for primary materials at the dawn of the digital age. *The American Archivist*, 66(1):9–50.

Michael Tschuggnall, Efstathios Stamatatos, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2017. Overview of the author identification task at PAN-2017: Style breach detection and author clustering. In *Working Notes of CLEF 2017*, Dublin, Ireland.