Allen, R.B., Japzon, A., Achananuparp, P., and Lee, K-J., A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. HCI International Conference, 2007.

A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers

Robert B. Allen, Andrea Japzon, Palakorn Achananuparp, and Ki Jung Lee
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
rba@drexel.edu, acj26@drexel.edu, pa442@drexel.edu, and leekijung@gmail.com

Abstract

Large quantities of historical newspapers are being digitized and OCRd. We describe a framework for processing the OCRd text to identify articles and extract metadata for them. We describe the article schema and provide examples of features that facilitate automatic indexing of them. For this processing, we employ lexical semantics, structural models, and models of community events. Furthermore, we describe visualization and summarization techniques that can be used to present the extracted events.

1 Introduction

Worldwide, there are many projects to digitize historical newspapers. One of the largest of these is the National Digital Newspaper Program (NDNP) of the U.S. National Endowment for the Humanities (NEH) and the Library of Congress (LC) that proposes to digitize several million pages of newspapers which were microfilmed as part of the United States Newspaper Program (USNP). We describe a framework for indexing and providing access to that material. We employ an iterative process of developing modular models and improving the corpora based on the materials in the NDNP collections (Figure 1). We have received extensive files of the *Washington Times*[1] (1900-1902) from NEH/LC and we are using them to test our framework and tools.

We are focusing at first on improving access to the content and then on categorizing articles in the collection. Typically, we will first apply knowledge in which we have high confidence and then move to inference techniques that employ probabilistic inference. Text processing will help to categorize and index the articles and will be used to develop rich semantic representations. Moreover, the metadata and indexing derived from text processing can provide input for improved user interfaces. Both improved interfaces and representations of the historical events reported in the newspapers will organize, and facilitate access to this large news archive.

2 OCR Processing

2.1 Processing the OCR's Text

The words in the newspaper images have been extracted with OCR. The quality of the OCR is variable. In some places, the quality is high because the original text and the NDNP microfilms are also high quality. However, in other places, where the print is faded or smudged, the quality is marginal. The OCR'd text is delivered in METS ALTO XML

[1] No connection to the contemporary newspaper with the same name.

(http://www.loc.gov/ndnp/techspecs.html). LC itself has been digitizing the *Washington Times* and we obtained the XML files for it. We process the METS ALTO OCR files to extract the article metadata. A gradual convergence with inferences across several levels is needed so the processing needs to be modular and adaptive. Figure1 shows the extracted and lightly-processed text along with page coordinates and fonts. Figure 2 shows a longer passage.

```
HEAD(ID11): HP: 1176.0 OS: 1806.0 VP: 4056.0 WC:    1.0 HD:    1.0 ST: ID3 T: 20  CO:  Numerous
HEAD(ID11): HP: 1902.0 OS: 2256.0 VP: 4047.0 WC:    1.0 HD:    1.0 ST: ID11 T: 21  CO: Other
HEAD(ID11): HEAD(ID11): HEAD(ID11):
begin TEXTLINE
HEAD(ID11): HP: 1176.0 OS: 1308.0 VP: 4242.0 WC:    1.0 HD:    1.0 ST: ID11 T: 22  CO: of
HEAD(ID11): HP: 1389.0 OS: 1602.0 VP: 4236.0 WC:    1.0 HD:    1.0 ST: ID11 T: 23  CO: the
HEAD(ID11): HP: 2004.0 OS: 2274.0 VP: 4233.0 WC:    1.0 HD:    1.0 ST: ID3 T: 24  CO:  hers
HEAD(ID11): HP: 2367.0 OS: 2676.0 VP: 4221.0 WC:    1.0 HD:    1.0 ST: ID11 T: 25  CO: Slain
##(accept current)
HEAD(ID11): HP: 2811.0 OS: 3387.0 VP: 4209.0 WC:    1.0 HD:    1.0 ST: ID11 T: 26  CO: Captured
HEAD(ID11): ##(skip current)
begin TEXTLINE
HEAD(ID11): HP: 1173.0 OS: 1311.0 VP: 4434.0 WC:    1.0 HD:    1.0 ST: ID11 T: 27  CO: or
HEAD(ID11): HP: 1383.0 OS: 1860.0 VP: 4413.0 WC:    1.0 HD:    1.0 ST: ID11 T: 28  CO: Injured
```

Figure 1: Example of extracted OCR for the METS ALTO. The text on the right is news about the Boer War the word "Boers" is incorrectly identified as "hers".

> The petitioners say that living [I]n the [] near the mill is a burden and they assert [Chat] the smoke is not only a menace to the [h]ealth of the neighborhood but that clothing bedding etc are injured by the smoke and that the [I]nteriors their houses are practical[] ruined [] [I]t.

Figure 2: A longer, relatively clean, fragment of processed OCR of text from the Washington Times (1902). Errors are indicated with brackets.

2.2 Article Segmentation

The basic unit of newspapers is the article. The METS ALTO standard used by awardees includes segmentation by textblock, which generally means by paragraph rather than by article.. Article segmentation for some historical newspapers is done manually. Therefore, we are developing an article segmentation utility to provide this function. For that tool, we will use a pixel analysis of the original TIF files as well as zones identified by word-coordinates from the OCR files and we will use knowledge about the layout for each newspaper (such as the layout of the masthead, see next section). Gatos et al. (1999) have reported considerable success segmenting articles in one newspaper but we believe this is a difficult task for most historical newspapers. Beyond the clues from the image, we will apply semantic-based segmentation (e.g., Hearst, 1994). We have developed an interface for human identification and tagging of articles (Figure 3). We have tagged a test set of articles which can serve as the ground-truth for tuning the automated article-tagging process.

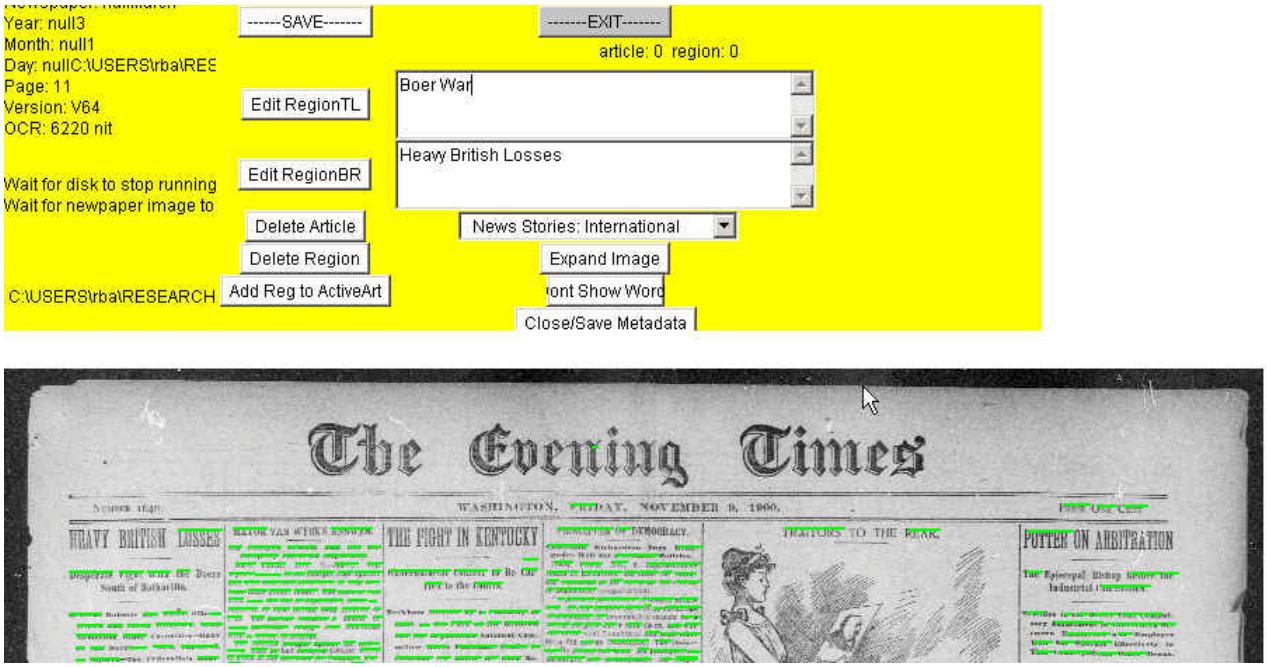

Figure 3: Interface for extracting ground truth about articles from newspapers. The upper section is a control panel. The user can mark out articles on the newspaper image. In addition, the word positions identified by the OCR can be presented. Note that the article and word boundaries are not shown, but well reproduced, in this image.

2.3 Newspaper Structure

Some basic information about the structure is helpful for effectively tagging and processing the articles. Indeed, this is used by some OCR extraction companies. Both the content and layout of individual news articles are highly structured. There is often redundancy within a single newspaper (across time) and across different newspapers. We can use the structure and redundancy to improve the accuracy of the text processing. We are developing hierarchical descriptions from sections, regions, and pages.

The newspaper structure goes beyond the definitions of ALTO structure. This would include the types of material we can expect to find on specific pages and in specific sections. The newspaper model will characterize the structure and style of the material typically found in the newspapers. Some newspapers have continuation of articles. Some have advertisements on the front page. Initially, it will be tuned for the *Washington Times*. The parameters include continuation, column width. Moving to higher level content such as where advertisements are likely to appear and the appearance of features are expected.

2.4 Content and Community Models

Newspaper content is very different from other types of material; typically, it is full of named-entities, that is the names of Organizations, People, Places, Roles, Dates, and Holidays. Moreover, there is a distinctive discourse or style (e.g., Liddy et al., 1993). Further, the collection of historical newspapers is distinctive because we have the complete history of news events rather than processing them as they are streamed. Indeed, in many cases (at least for national and international

stories), we will be able to compare the way that they are presented across newspapers. All of this allows us to develop better models for processing and presenting the material..

The content characteristics of a particular community such as whether it is urban or rural and its governmental structure will be described. Events in the community's history will also be described. Some events are related to specific time periods. This includes both predictable events (e.g., elections, harvests, snow storms) as well as relatively unique events (e.g., a World's Fair).

The community constraints will reflect the traditions and distinctive activities of individual communities. For instance, we will develop models for two specific localities --- a rural model for the files from Utah and an urban model based on files for Washington, DC. The model will be elaborated as newspapers from other communities are added. The model will also include events scripts; in some cases, these will be generic (e.g., what types of events are likely to be reported following a flood) and in other cases, the event instances will be elaborated in detail (e.g., a specific flood).

2.5 Extracting and Using Named Entities

In this context, Entity-lists form the lexicon employed in the community constraint model. There are many sources of ontologies and of tools for developing ontologies (e.g., Bontcheva et al., 2002; Smith, 2002), but none of these is entirely appropriate for the application here in terms of the time-period covered. However, other types of material such as directories, gazetteers (e.g., Hill, 2000) and databases such as those available at Ancestory.com would be useful and we may explore research agreements to use them.

There is much activity on named entities. For processing historical U.S. newspaper articles, it will be essential to have systematic information about the relevant time periods. As an initial step, Allen and Achananuparp (submitted) have developed descriptions of prominent U.S. politicians with governmental structure both past and present. This extends the *Federal Register's US Government Manual* and should have independent value. However, Crane and Allison (2006) have recently described the difficulties in named-entity extraction. Furthermore, the named-entities in the articles need to be matched to named-entities in entity lists and lists expanded.

For processing the articles, we expect to use Gate, a standard named-entity extraction tool (Cunningham et al., 2002). While it is unclear how OCR degradation as seen in Figure 1 affects POS tagging, we expect to be able to repair the OCR sufficiently so that POS tagging performance will not be significantly degraded.

2.6 Article Metadata

Articles will be categorized by genre (e.g., editorials, reviews, advertisements) and by topic (international news, sports). This extends the earlier work of Allen and Schalow (1999). The International Press Telecommunications Council (http://www.iptc.org/). Both types of categories can be useful as metadata. Automatic genre categorization is based on the structural constraints. For instance, in some newspapers, the editorials appear on the center page. Figure 4 illustrates a possible descriptive metadata schema based on the Dublin Core. This is human generated and therefore costly in terms of time but thorough in description.

Element	Scheme	Example of content
DC Title		Tuan in Open Rebellion : Boxer Leader and General Tung-fuhsiang take the Field
DC.Subject	LCSH	China--History--Boxer Rebellion, 1899-1901
DC.Subject	NITF	Unrest, conflicts and war--civil unrest--rebellions
DC.Subject	Personal Names	Tuan, Tung-fuhsiang, Bazaure, Chwang, Yu Hsien
DC.Subject	Topic	International
DC.Type	NITF-newscode – genre	Actuality
DC.Coverage.Geographic Location.Historic		
DC.Coverage.Geographic Location.Current		Gansu Cheng - China; Kansu
DC.Coverage.Date		November 15, 1900
DC.Identifer.Serial	Standard No, LCCN	sn82-16310
DC.Identifer.Issue		1654
DC.Identifer.Page		01
DC.Identifer.Article		
DC.Relation	Use this to connect related articles	

Figure 4: Qualified Dublin Core descriptions proposed for articles.

After news and advertisements, notices are the next largest group of articles in the *Washington Times*. Figure 5 lists the article headings for a variety of notices. Notices are the most varied in terms of topic and structural presentation. Personal notices are distinguished from corporate or social entities through the font and formatting of the headings. Notices contain a variety of topics from death announcements to sales of coal.

Notices

Death	Personals
Special Notices	Birds, Dogs, Etc.
Commissioner's Notices	Recovery from Illness
Society announcements	Educational
Crime report	Foreign Mails
Collaterals Forfeited	Help Wanted

Figure 5: Subdivisions of Notices.

The primary indication of a unique article of the *Washington Times* is the shape of an article given by its borders and visual cues. Advertisements are typically the easiest to detect with dark solid lines for borders or borders that are laced asterisk styled characters (e.g., Figure 6). Further, advertisements are the only type of article that contains graphic images as part of an article. Political cartoons are the only other form of images in the *Washington Times*.

The first line in the heading of a news article is in all capital letters and the largest font size of all titles/sub-titles in the heading. If there is a sub-title to the article, then a thin line of about a half inch in length is drawn. After this line, the subtitle appears and it is bold faced but not in all capital letters. If there is an additional subtitle, then another thin line will appear before the next subtitle which will be smaller in font size than the first subtitle. Articles that are not from DC will begin with the location of city and sometimes the state, followed by the date and a dash for example, “BALTIMORE, Md., Nov. 15-- ”.

This consistent way of formatting news articles is helpful in discriminating articles that are not news. There are many medical cures advertised that try to pass for news articles but lack the

formatting particular to news articles.. Also, as the *Washington Times* fills the entire column, many very small articles are included in its pages. Typically, these are a one-to-three sentence excerpts from other newspapers; these excerpts are consistently formatted as well.

TUAN IN OPEN REBELLION

Boxer Leader and General Tungfuhsiang Take the Field.

Reported to Have a Large Army in the Province of Kansu—The Insurgent Prince and Chwang Deprived of All Rank and Turned Over to the Court by an Imperial Edict. Governor Yu Hsien Ordered Imprisoned. Pending His Banishment to the Farthest Borders of the Empire—Serious Disturbances Occur.

LONDON, Nov. 15.—Despatches from Shanghai say that, according to reports from reliable Chinese sources, Prince Tuan and General Tung-fuhsiang are in open re-

The Saks Stores

Pennsylvania Ave. and 7th St.

Extra Special Sale of Men's Fancy Vests

Entire Fancy Vest Stock Among Our Stores.

$1.00 for choice of $2.00 values.
$1.50 for choice of $2.75 values.
$1.90 for choice of $3.50 values.
$2.40 for choice of $5.00 values.
$3.45 for choice of $6.00 values.

Saks & Company.

STOCKS ACTIVE AND FIRM

Little Change in the Features of the Market.

Pacific Railroad Shares Conspicuous in the Business—The Industrial Quarter Set Particularly Buoyant. Local Traction Group Higher. Irregular Price Changes in London.

NEW YORK, Nov. 15.—The Stock Market opened active and decidedly firm. At was the case yesterday, the preferred shares of the Pacific roads were conspicuous in the dealings, and Louisville, Pacific, preferred, and Union Pacific preferred, stocks rapidly ran up to unusually large values of each. The industrial quarter was not particularly active, but kept very strong, and a fairly steady tone was continued to the market.

The local traction group opened materially higher on reports that the franchise tax law would be modified at the approaching session of the Legislature.

London furnished irregular price changes for American stocks and the dealings here for foreign account were on both sides of the market.

Little or no change appeared in the financial features of the situation over night.

New York Stock Market.

Corrected daily by W. B. Hibbs & Co., members of the New York Stock Exchange, 1420 F

Figure 6: Three examples of newspaper structures that can facilitate automated processing. In the article on left, note the distinctive layout and fonts. In the advertisement, center, note the dark border. For the stock reports on the right, note the distinctive tabular structure.

There is usually a large number of clues in each article relevant to category assignment. We will develop predictors of such clues for each category. For topic categorization, two approaches are commonly used: (a) the distribution of words that appear in the text (e.g., Hovy & Lin, 2000); and (b) the presence of specific named-entities.

2.7 Refining the Constraints and Named-Entity Lists

We are exploring the use of back-propagation neural networks learning about and adapting to the constraints (e.g., Rumelhart & McClellan, 1986) but refining the constraints (models) and entities will necessarily be a joint human-machine activity. The members of the project staff will make revisions from their own knowledge and we will encourage librarians and community groups to enter local knowledge. Thus, we will provide interfaces for people to work with the constraints and named-entities. Furthermore, because this may be a hybrid human-machine update process, we will track the sources of all revisions (i.e., data provenance) in case the learning drifts and needs to be reset and restarted. Because the extent of community participation may vary greatly, the text processing tools we will also need the option of tools that can operate with minimal human intervention.

2.8 From Historical Facts to Histories

Once we will have extracted entities, we can use them in many ways. For instance, users of the system could browse the people mentioned in the news for a given year. These brief descriptions for individuals can link to specific news articles. We would also seek to provide summaries of major news topics and even to synthesize responses to open-ended queries about the news events in

a given period. These are increasingly ambitious text processing tasks and they depend increasingly on statistical inferences.

2.8.1 Clustering News Stories and Topic Threads

News articles often report an evolving news story so that later articles expand upon earlier ones. A person viewing any one of these stories, may want to know about the other ones. To identify similar articles, a cluster analysis can be conducted. There has been a great deal of work on topic detection and tracking (TDT, Allan, 2002; Fiscus & Doddington, 2002). In TDT, nascent news topics are detected and tracked as they are received. In this project, we are looking for topics in an historical collection rather in a news stream. Thus, we can examine the entire corpus when trying to identify a topic thread. In fact, we could identify the topic first and work backwards to find the first story.

2.8.2 Multi-Document Summarization of Articles

One benefit of identifying topics sets of articles as described in the previous section is that we can pick articles as summaries from that set. Some articles will provide more background and could even provide overviews. For instance, an article appearing at the end of a sequence of articles on a given topic might summarize the earlier articles. However, none of the articles may be an adequate summary. Thus, we might be able to analyze the contents in more detail and to summarize them. For example, someone interested in the history of Washington, DC might ask the question "Describe the results of the 1902 commission to redesign Washington" and a summary could be provided.

Multi-document summarization combines evidence from several documents. The simplest method of summarization is extracting sentences; this is often accomplished by picking sentences with salient terms (Salton et al., 1994). However, such extractive summaries are often unsatisfactory, and it may be helpful to consider a wider range of features such as the rhetorical structure of articles ---that is, the discourse purpose of different sections (Radev, 2000). Techniques developed for dynamic collections (i.e., streaming news sources) may also be adapted. For instance, "NewsBlaster" is a news summarization tool (Hatzivassiloglou, et al., 2000; McKeown, et al., 2002). As with TDT, news stories are processed in NewsBlaster as they are received, and we could likewise adapt this technique to fixed news collections.

Even more broadly we may want to understand the historical context of news articles. Consider a user who has identified an ancestor from a rural county and found news stories mentioning that ancestor that referred to unfamiliar events in the county's history. An interface may be developed to present summaries of those events to provide context about that ancestor. Finally, we will explore interfaces for coordinating with other historical resources.

2.9 Interfaces and End-User Access

NDNP will greatly benefit from improved end-user access to its content. The current generation of interfaces for digitized historical newspapers depends largely on either browsing by date or searching the OCR'd text. Many other types of navigation and search are possible. A graphical view may be the most effective way to present large amounts of information. Specifically, timelines which display sets of temporal events would be the central organizing metaphor. This could be sets of articles (Swan & Allan, 2000) or it could be attributes of Entities such as events in

a person's life (Allen, 1995, 2005a, b). Figure 7 shows a novel timeline developed and applied to Brooklyn history. Significant events of Chinese immigration in Brooklyn are shown to the right of the center line. Broader historical context is shown to the left of the center line. Both (manually generated) text summaries of an article on Chinese immigration in Brooklyn and related news articles are shown in pop-up windows.

While simple timelines show sets of events, we'd also like to show dependencies among events. Events may be shown as causal relationships and the causal relationships may be woven into narratives. In previous work, we have explored graphical approaches showing causal relationships and narratives suitable for presentation in a Web browser. Specifically, a model of causation between events in scientific explanations was proposed by Allen, et al. (2005a). In Allen and Acheson (2000), narrative sequences of events were identified for a children's story. We will explore extending these techniques to large and complex descriptions of historical events as drawn from the newspapers.

Figure 7: A timeline browser for the *Brooklyn Eagle* (from Allen, 2005a). Focus events are shown right of center while sets of contextual events are shown on the left side.

3 Conclusion

We have outlined a framework for processing the digitized historical newspapers such as those from the NDNP. The first step is article segmentation based on image characteristics and the OCR'd text. Articles will be categorized based on their text, their position in the newspapers, and relationship to other articles. This text processing will facilitate access to the articles and eventually to the extraction of events.

4 Acknowledgements

This work is funded in part by NSF grant #0329111 for ``Interacting with Threaded Event Scenarios".

5 References

Allan, J., Introduction to topic detection and tracking, in *Topic Detection and Tracking: Event-Based Information Organization*, Allan, J. (ed), Kluwer, 2002, 1-16.

Allen, R.B., Timelines as information system interfaces, *Proceedings International Symposium on Digital Libraries*, 1995, 175-180.

Allen, R.B., Developing a query interface for an event gazetteer, *IEEE/ACM Joint Conference on Digital Libraries*, 2004, 72-73.

Allen, R.B. A focus-context timeline for historical newspapers, *IEEE/ACM Joint Conference on Digital Libraries*, 2005a, 260-261.

Allen, R.B., Using information visualization to support access of archival records. *Journal of Archival Organization*, 3, 2005b, 37-49.

Allen, R.B., and Acheson, J., Browsing structured multimedia stories, *ACM Digital Libraries Conference*, 2000, 11-18.

Allen, R.B., and Achananuparp, P., Providing civic information via a Web service, submitted. *International Conference on Information Systems.*

Allen, R.B., and Schalow, J., Metadata and data structures for the Historical Newspaper Digital Library Project, *ACM Conference on Information and Knowledge Management*,, 1999, 147-153.

Allen, R.B., Wu, Y.J., and Jun, L., Interactive Causal Schematics for Qualitative Scientific Explanations, *International Conference on Asian Digital Libraries,* 2005, (LNCS 3815/2005,411-415).

Beitzel, S.M., Jensen, E.C., and Grossman, D.A., Retrieving OCR text: A Survey of Current Approaches. *Symposium on Document Image Understanding Technologies (SDUIT)*, Greenbelt, MD, 2003.

Bontcheva, K., Maynard, D., Cunningham, H., and Saggion, H. Using Human Language Technology for automatic annotation and indexing of digital library content. *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, 613-625. 2002.

Caplan, P., Barnett, B., Bishoff, L., Borgman, C., Hamma, K., and Lynch, C., *Report of the workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content*, 2003. http://www.imls.gov/pubs/pdf/digitalopp.pdf

Cohen, W.W., Infrastructure components for large-scale information extraction systems, Conference on Innovative Applications of Artificial Intelligence, 2003.

Cox, R.J., *Documenting Localities: A Practical Model for American Archivists and Manuscript Curators,* Scarecrow Press, 2001.

Gatos, B., Mantzaris, S.L., Chandrinos, K.V., Tsigris, A., and Perantonis, S,J, Integrated Algorithms for Newspaper Page Decomposition and Article Tracking, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, 559.

Crane, G., and Jones, A., The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-Century newspaper collection, *IEEE/ACM JCDL*, 2006.

Cunningham, H., Bontcheva, K., Tablan, V., Ursu, C., and Dimitrov, M., *Developing language processing components with Gate (user*'s guide), Technical report, University of Sheffield, U.K., (2002). http://www.gate.ac.uk/

Fiscus, J.G, and Doddington, G.R., Topic detection and tracking evaluation overview. in *Topic Detection and Tracking: Event-Based Information Organization*, Allan, J. (ed), Kluwer, 2002, 17-31.

Hearst, M., Multi-paragraph segmentation of expository text, Proceedings *Association for Computational Linguistics*, 1994, 9-14.

Hill, L., Core elements of digital gazetteers: Placenames, categories, and footprints, *Proceedings of the European Conference on Digital Libraries*, 2000, 280-290.

Hovy, E., and Lin, C-Y., Automated text summarization in SUMMARIST. In I. Mani and M. Maybury (eds.), *Advances in Automated Text Summarization*. MIT Press, 1999.

Liddy, E.D, McVearry, K.A., Paik, W., Yu, E., and McKenna, M., Development, implementation and testing of a discourse model for newspaper texts, *Human Language Technology*, 1993.

Mantzaris, S.L., Gates, B., Gouraros, N. and Tzavelis,P., Integrated Search Tools for' Newspaper Digital Libraries.

McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans. J.L., Nenkova, A., Sable, C., Schiffman, B., and Sigelman, S., Tracking and summarizing news on a daily basis with Columbia's Newsblaster, *Human Language Technology*, 2002.

Murray, R., Towards a metadata standard for digitized historical newspapers, *IEEE/ACM JCDL*, 2005, 330-331.

Petras, V., Larson, R.R., and Buckland, M., Time Period Directories: A Metadata Infrastructure for Placing Events in Temporal and Geographic Context. *IEEE/ACM JCDL*, 2006.

Radev, D.R., A common theory of information fusion from multiple text sources step one: Cross-document structure, *ACL SIGDIAL*, 2000, (http://www.sigdial.org/workshops/ workshop1/proceedings/radev.pdf).

Riloff, E. and Lehnert, W., Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*,. 12, 1994, 296-333.

Rumelhart, D.E., and McClellan, J., editors. *Parallel Distributed Processing*, 2 volumes. MIT Press, Cambridge, MA, 1986.

Salton, G., Allan, J., Buckley, C., and Singhal, A., Automatic analysis, theme generation, and summarization of machine-readable texts, *Science*, 264, 1994, 1421-1426.

Smith, D.A, Detecting events with date and place information in unstructured text. *ACM/IEEE Conference Digital libraries*, 191-196, 2002.

Swan, R., and Allan, J., Automatic generation of overview timelines, *ACM SIGIR*, 2000, 49-56.