The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-Century Newspaper collection

Gregory Crane
Perseus Project
Eaton Hall
Tufts University
(617)627-383
gregory.crane@tufts.edu

Alison Jones
Perseus Project
Eaton Hall
Tufts University
(617)627-3830
alison.jones@tufts.edu

ABSTRACT

This paper evaluates automatic extraction of ten named entity classes from a 19th century newspaper, the Civil War years of the *Richmond Times Dispatch*, digitized with IMLS support by the University of Richmond. This paper analyzes success with ten categories of entities prominent in these newspapers and the particular problems that these classes of named entities raise. Personal and place names are familiar but some more important categories (such as ship names and military units) illustrate some of the challenges that named entity identification confronts as it evolves into a fundamental tool not only for automatic metadata generation but also for searching and browsing as well. We conclude by suggesting the kinds of knowledge sources that digital libraries need to assemble as part of their machine readable reference collections to support named entity identification as a core service.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval-*digital libraries*

General Terms

Design, Performance, Experimentation

Keywords

Historical newspapers, Digital Libraries, Named Entity Recognition

1. INTRODUCTION

As Roy Rosenzweig has noted, in terms of the historical record, we are moving from the problem of scarcity to one of abundance, as more material becomes digital [28]. Michael Lesk has also suggested that scholars should support mass digitization projects as increasing the quantitative amount of information on the Internet will also lead to qualitative improvements [21]. Evolving work on various language technologies, such as machine translation, automatic summarization, cross language information retrieval, and question answering, to name only a few, has immense potential for helping scholars wade through these vast oceans of historical materials. An alphabet soup of evaluation forums (TREC, ACE, CLEF, DUC) publish results in this field and progress has been rapid, but these results have only begun to find their way into working systems accessible to wider audiences and into the digital library arena [9, 24]. More significantly, most research has focused upon contemporary materials such as usenet archives, article abstracts, job listings and current news reports. The materials analyzed tend to be structurally homogeneous and to cover only the recent past. The gazetteers, encyclopedias and heuristics in use thus do not directly address the needs of historical materials, which describe people, places, and other entities that often do not appear in modern knowledge sources.

This paper documents the results of work with one fundamental technology – named entity identification – applied to a set of historical newspapers digitized with IMLS support by the University of Richmond and builds upon previous work [30, 31, 32]. The results are significant for three reasons.

First, named entity recognition addresses documented needs of many users, who are anxious to search for particular people and places. Studies of information-seeking behavior by humanists frequently emphasize the importance of being able to search for particular people, places, and other proper nouns. Over a decade ago, the Getty Online Searching project found that humanities scholars typically searched for named individuals, geographical terms and chronological terms [7]. A more recent examination of the information-seeking habits of humanities scholars using digital libraries found that all those surveyed frequently used proper names in their searching, though this strategy often met with only limited success [10]. Several research studies that focused exclusively on historians also cited their preference to search by proper names, particularly the names of individuals and geographic locations [11, 15]. Similarly, a study on the research habits of genealogists illustrated that they wanted to be able to search by name, distinguish between different individuals with the same name, and be able to link current place names to historical ones in their searches [14]. This process of named entity identification in historical documents thus has the potential to serve a diverse group of audiences from humanities scholars to genealogists to the everyday user.

Second, named entity recognition not only supports direct user functions but also constitutes a fundamental step in ontology creation, a knowledge source for the emerging

generation of information retrieval systems and an important component in many emerging technologies [1, 8].

Third, named entity extraction is a notoriously domain dependent field. Chemists, for example, need to find chemical compounds, geneticists need to locate specific genes and relationships between them, and literary scholars need to identify source citations; each of these scenarios raises different issues and requires different strategies. Nevertheless, we need general systems that can select the best strategy with which to analyze a specific set of documents for a particular set of purposes. We thus need not only evaluation of the performances of different systems on various collections but also analyses of what entities matter in diverse contexts and what major sources of errors restrict performance.

This paper evaluates results of named entity extraction from the Civil War years of the *Richmond Times Dispatch*, a 24 million-word corpus. This paper (1) describes not only familiar categories (e.g., personal and place names) but also categories relevant to the period (e.g., ship names, Civil War military units), (2) evaluates the success for each of these categories, (3) illustrates major problems that we confront in extracting categories from 19th- century newspapers, and (4) suggests the impact that the use of more contemporary knowledge sources can have upon recall and precision.

2. BACKGROUND

Our current work with newspapers carries forward more than twenty years of work on digital libraries that exploits the automatic identification and analysis of full text to provide advanced searching, browsing and visualization. The Perseus Project has previously developed a general system to extract dates and places from all texts in the Perseus Digital Library [30, 31, 32]. That work demonstrated that American English posed greater challenges than our Greco-Roman or European collections: Americans were far more likely to reuse the same names such as Springfield and Washington over and over for different places and much less likely to mark place names with semantic classifiers.

We are currently working with Civil War issues of the *Richmond Times Dispatch*. This newspaper reflects the high end of data entry and markup, with double keying maximizing transcription accuracy, and careful markup of individual articles. This newspaper allows us to approach a variety of problems, the most significant of which is how to automatically analyze the content of articles that are both short and diverse in content. In these newspapers, an article on a local election may follow an allusive description of Garibaldi's activities in Italy. Since the source images are often hard to read, even the carefully entered texts contain more errors than in materials entered from clearer print. The newspapers contain thousands of personal, place, and organizational names, including detailed addresses (in a range of abbreviated formats), as well as advertisements and other commercial data.

3. RELATED WORK

The work presented in this paper is related to two different streams of research: trends in newspaper digitization and named entity recognition in historical text collections

3.1 Newspaper Digitization Projects

The onset of the National Digital Newspaper Program and a number of highly successful individual historical digital newspaper projects such as the Brooklyn Daily Eagle and Utah Digital Historical Newspapers has spurred research into the issues surrounding the digitization of older newspapers. Research has included case studies of projects [5, 12, 27, 33, 35], the challenges of newspapers digitization [6, 13, 16, 20], the development of metadata standards [3, 23], and the potential linguistic research uses of newspapers [22, 25]. Little of this work has focused upon how the technologies of information extraction and named entity extraction might be applied. One innovative application of language technologies was [26] where the authors examined how an incremental hierarchical clustering algorithm could be used to support topic detection and tracking in a newspaper digital library. [2,4] have reported on how automatically generated timelines and event gazetteers might be used to provide better browsing access to collections of historical newspapers. The most directly related work is that of [18] who developed a system that supports simultaneous access to a digitized collection of historical maps and newspapers. Their system, called HistoryMap, dynamically associates historical maps with Maori and other place names found in the New Zealand Digital Library Niupepa collection, and allows users to search the collection by place name.

3.2 Named Entity Recognition & Historical Collections

Named entity recognition systems and their potential use in historical documents is a topic that has received surprisingly little attention, although interest in this area is quickly growing. Two recently announced projects of note are the University of Sheffield's plan to explore data mining technologies in digitized eighteenth century materials[1] and the NORA project,[2] which is examining how text mining can be used in currently existing digital libraries of historical materials.

The problem of named entity recognition, particularly the problem of personal and place name disambiguation, in historical text collections has received some attention. Recent work reported by [17] discusses the possibilities of creating an electronic index to a corpus of 17th-century Dutch resolutions that are encoded in XML with all instances of people, institutions, ship names and geographic names tagged by hand. The authors found that they needed to create a separate relational database to store encoded information about institutions and people and that one of the most significant challenges was integrating searching between the two systems. Of particular issue was the problem of co-reference resolution; as they explained: "it may require quite a number of different string queries to find all locations in our texts where the person 'Louis, King of France' is mentioned in our corpus."

Similar work has been conducted by the creators of the Old Bailey Proceedings. As [29] reports, they were attempting to create a historical source that would allow historians to

[1] http://www.hrionline.ac.uk/armadillo/sources.html

[2] http://nora.lis.uiuc.edu/description.php

explore these records from the perspective of an individual Londoner. In this project, tagging consisted of both semantic classification and identification of entities. While some of the XML markup was done manually, personal names were automatically tagged with GATE (Generalized Architecture for Text Engineering).[3] They found that through the manual creation of rules for GATE, they were able to automatically tag about 80-90% of personal names successfully. While place name tagging was performed manually they still faced the issue of having one place match multiple places on a map. Since they found it too expensive to manually check and correct all names, they decided to let users specify the amount of uncertainty they were willing to tolerate while searching, either searching for all "precise matches" or "possible matches." For future work they plan to make increasing use of automated markup, by taking advantage of the structured nature of XML records and using the authority lists they had already compiled.

The open source system GATE has also been used by a number of different organizations to support named entity recognition and information extraction from historical collections. The most recent release of Greenstone has included GATE in order to support more sophisticated text mining of digital collections [34]. The Ephemeral Cities project is using GATE to extract personal names from a variety of local history materials including newspapers. [19]. As these projects indicate, named entity recognition has increasingly become identified as an important service that historical digital libraries can provide.

4. AGGREGATE FIGURES FOR MAJOR ENTITY TYPES IN THE *RICHMOND TIMES DISPATCH*

Our system draws upon three strategies for named entity analysis. First, it scans for distinctive multi-word expressions (e.g., "the Chicago Tribune," "the Boston and Maine Railroad"). Second, it applies rule sets to detect various classes of entity (e.g., "Lieutenant NAME1" _ NAME1 designates a person, "the schooner NAME2" _ NAME2 designates a ship name). We also compare patterns such as "NAME3, US_STATE" to identify "Boston, Mass." as not only a place but the particular place designated by the Getty Thesaurus of Geographic Names as "tgn,7013445." Third, it analyzes results from the first two stages to generate statistical models based on overall frequencies and on trigrams (the target word plus the two preceding words) to determine whether a given "Washington" designates a person or a place. If the "Washington" is a place, then we look at the geographic context and overall frequency of various "Washingtons" to choose which "Washington" is meant. If the "Washington" is a person, then we scan for nearby references to "Washingtons" with forenames (e.g., "George Washington," "Booker T. Washington"). If there are no nearby references, then we link a given "Washington" to the most commonly named "Washington" in the overall document (if there is one).

At this stage, we have chosen to defer analysis of one major class of named entity. Elliptical phrases such as "the Ohio" can, for example, refer to the Ohio River or to a ship. For now, we examine phrases of the form "the NAME" and classify them in two ways. If the phrase appears to be a self-standing noun phrase (e.g., "we saw the Ohio yesterday"), we use the TEI tag RS (reference string): "we saw the <rs>Ohio</rs>." If the name seems to be part of a larger phrase, then we use the TEI tag NAME: "we saw the <name>Ohio</name> representatives." With the exception of a few frequent entities (e.g., "the <rs>Mississippi</rs>_"the<placeNametype="river"key= "tgn,7022231">Mississippi</placeName>"), we do not, at present, classify these further, although we have worked on this problem and will address it subsequently. The *Richmond Times Dispatch* currently includes 184,511 entities with an untyped RS tag. The top five terms with such unmarked RS tags are "the South" (12,854), "the Government" (9,782), "the North" (8,584), "the State" (8,325), and "the President" (7,121). Two of these (North/South) are unambiguous and could be assigned standard identifiers. Three (Government/State/President) are ambiguous but could probably be assigned with high accuracy if we tracked the context (e.g., President _ Jefferson Davis unless there has been a recent mention of Abraham Lincoln). Unresolved RS tags have the greatest impact upon rivers (e.g., "the Pamunkey"), newspapers (e.g., "the Herald"), and ships (e.g., "the Monitor"). For further information on the system please consult our related technical report, which is currently being revised.[4]

Table One provides counts for the ten categories of named entities evaluated for this paper. Some elements were not included. The *Richmond Times Dispatch* includes more than 170,000 annotated quantities that include currency (e.g., "12$", "forty cents"), measures ("30 hogsheads," "15 barrels"), areas ("12 by 15 feet") and other units. These quantities are, however, on their own often not very useful but derive their significance when associated with particular objects (e.g., "15 barrels of rum," "a room of 12 by 15 feet"). Extracting propositional data is a crucial task but this study concentrates on the initial stage of named entity identification. Likewise time expressions ("10:30 AM," "in the afternoon") were not included in this analysis.

The table provides figures both for the *Richmond Times Dispatch* and for the 19th-Century American Collection created at Perseus. We provide gross frequencies but normalize these, showing the frequency of each category per 10,000 words: thus, we identified in the Richmond newspaper 542,271 possible personal names or about 231 personal names in every 10,000 words of text. The final column provides the percentage by which the normalized frequency changes when we move from the newspapers to the overall collection. The overall density of named entities was, to our surprise, slightly higher in the general collection – perhaps a reflection of the fact that the general collection contains a number of encyclopedias, gazetteers and other name-intensive documents.

The top three categories are the same in each collection: personal names appear about 250 times per 10,000 words in each collection, place names 180 times. Dates are the third most common entity in both collections, but dates are almost twice as frequent in the general collection as they are in the newspapers. Again, the disparity may reflect the presence of reference works, but even this factor may not be enough to

[3] http://gate.ac.uk/

[4] Crane, Gregory and Alison Jones. The Perseus American Collection 1.0. (2005). http://www.perseus.tufts.edu/~gcrane/americancoll.12.2005.pdf

account for the gap. Preliminary analysis suggests that these newspapers assume that readers will infer the date from the newspaper itself and only specific dates that differ significantly are fully specified. While the gross difference is intriguing, the fact that dates are the third most common entity in both corpora is probably more important. People, locations, and time are the three most important classes of entity in each case.

Table 1: Aggregate Named Entities

Richmond Times Dispatch	Entity Type	Freq.	19th-Century American Collection	Freq.	Percentage change
23,486,847	total words	-	57,045,605	-	
542,271	personal names	230.88	1,538,829	269.75	16.84%
401,128	places	170.79	1,081,903	189.66	11.05%
156,009	dates	66.42	686,580	120.36	81.19%
106,773	products	45.46	348,703	2.32	-94.89%
106,357	organizations	45.28	158,177	61.13	34.99%
40,728	streets	17.34	36,506	3.26	-81.17%
36,731	newspapers	15.64	31,432	5.51	-64.77%
23,979	ships	10.21	18,624	6.40	-37.32%
20,366	regiments	8.67	13,244	27.73	219.77%
6,627	railroads	2.82	8,804	1.54	-45.30%
1,440,969	**total**	**613.52**	**3,922,802**	**687.66**	**12.08%**

The remaining seven categories begin to illustrate more sharply the differences between the general 19th-century collection (based on books) and the newspaper collection. The newspapers contain twenty times as many references to possible products than do the books. The disparity is, in fact, much greater: many advertisements are repeated verbatim from one issue to another. Because of limited resources, the Richmond team chose to sample advertisements, with the goal being to provide at least one instance of each distinct advertisement that ran more than once. Extracting products from advertisements represents the single most distinctive challenge posed by newspapers. The system which we built to address this problem has also seen beneficial results in the19th-Century American Collection.

While no other category varies by more than an order of magnitude, streets, newspapers, ships, and railroads are significantly more common in the newspaper corpus. Streets are particularly important because they provide the foundation for addresses and general locational data. In the newspaper corpus, space-saving abbreviations (e.g., "near Main and Smith" for "near Main and Smith Streets") are much more common and demanded specialized rules and/or statistical models. The same methods without modification produced comparable results for newspapers, ships, and railroads in both corpora.

Military units and organizations other than newspapers, regiments, and railroads were significantly more common in the general collection. A strong bias towards Civil War histories accounts for the prominence of regiments mentioned in the monograph collection. The number of such references spurred us to develop a special purpose driver that could recognize "54th Massachusetts," "Massachusetts Fifty-Fourth Regiment," and "Mass. 54th" as descriptions of the same entity. As a result, we were able to retrieve more military units from the newspapers, since the prominence of regimental units in the monographic collection led us to create more sophisticated patterns than the lesser frequency in the newspaper collection might have justified.

4.1 Evaluation: Precision of Identification

For each of the available 1,355 issues of the *Richmond Times Dispatch*, we calculated the top 20 named entities in each of the ten selected categories, providing a maximum of 200 unique entities to describe each newspaper issue as a whole. Since the newspaper issues are relatively short (about 17,300 words each), many do not mention 20 distinct newspapers or military units and thus generate profiles with fewer than 200 entities. In effect, we provide automatic metadata to illustrate the "aboutness" of each article. Our focus in this study was thus upon our success with the more frequently cited entities rather than low frequency entities. The automatically generated data shows encouraging results: "ammunition," for example, becomes a prominent commodity in late June 1861, just before serious hostilities begin. Robert E. Lee first appears as a significant figure in April 1861. Gettysburg appears at the start of July 1863. The most frequent dates in each article almost always correspond with the date on which the articles were published – indeed, scanning the probable publication dates, we noticed a number of gaps where we did not have issues for particular days.

We are also interested in understanding how much impact limited correction can have upon overall performance. The discussion below points out a number of instances where the removal of a few obvious and very common errors can dramatically improve accuracy. In subsequent work, we plan to study the secondary effects of such high impact/low cost corrections on subsequent analysis. Suppose the geographic identifier accidentally shifts from the US to the UK and 35 instances of Cambridge are mistakenly placed in England rather than in Massachusetts. Fixing that one systematic error, in turn, provides 35 references that will associate particular contexts with the US. This can help the automated system to correctly identify three Brightons as being in the US as well. Checking a few high frequency results could thus have a multiplier effect within subsequent automatic analyses. Table

Two illustrates the overall results of our evaluation, on which we will focus the remainder of this paper.

Table 2: Manual Evaluation of Entity Results

Entity Type	Total Entities Manually Evaluated	Total Correctly Tagged Entities	Total Incorrectly Tagged Entities	Tagging Accuracy	Major Source of Error
Personal Names	2246	1720	526	76.58%	Incorrect Name matching
Places	2592	2525	67	97.42%	Tagging incorrect place, tagging ethnic groups
Dates	1338	1291	47	96.49%	Tagging partial dates
Products	488	281	207	57.58%	Tagging of products/commodities when used as common nouns
Organizations	823	676	147	82.14%	Tagging of general nouns as organizations (offices, army, etc.)
Streets	1210	1202	8	99.34%	Tagging towns such as Brighton, Holyoke as streets
Newspapers	401	267	134	66.58%	Tagging of proper nouns as newspaper titles
Ships	226	211	15	93.36%	Tagging verbs as ship names
Regiments	400	367	33	91.75%	Tagging document sections as military sections
Railroads	126	112	14	88.89%	Tagging of partial railroad names
Total Entities	9850	8652	1198	85.01%	

4.2 Personal Names

We evaluated the number of times we were able to associate a surname with the correct forename given within the document. Our accuracy level for the tagging of this entity category is currently just over 75%. One major issue is that many references to individuals included a surname only, an abbreviated name, or referred to an individual that cannot be found within currently existing authority lists. Newspaper articles only rarely specified an individual's name in its entirety, particularly when discussing well-known individuals such as General Lee or President Lincoln.

Our system currently tags all floating surnames such as "Lee" or "Butler" and then attempts to match that name with the most likely possible full name. It searches the surrounding text and if it finds a fully specified name such as "David Butler," the system subsequently assigns this name to all other Butler references. If it cannot find a fully specified name with the surname it will tag it with the string "Butler, nomatch." The results of this system can be quite problematic within newspapers, where articles are short and widely disparate topics can be discussed within mere sentences of each other. Since the names of lesser known individuals, such as a criminal appearing in court, were more likely to be fully specified in the newspaper text, our system often assigns the wrong individual's name such as "David Butler" to references that are actually to "General Benjamin Butler." Similarly just one reference to a "Granny Scott" in a short anecdote caused ten incorrect attributions of a General Scott. We thus need a context dependent model for default names in order to be able to weigh the significance of a reference to a random "John Lee" against the significance of "Robert E. Lee" in a given context.

A variety of other errors were also encountered in the tagging of personal names. While we recognize patterns such as "Brown & Smith Co." as an organization composed of two surnames, we only tag "Brown & Smith" as two surnames, accounting for about 18% of all errors. This most frequently occurred with company names. The small pharmaceutical store "Goddin & Apperson" placed a large number of advertisements in our newspaper collection, leading to a large number of incorrect surname tags for both these names.

Another difficulty is that many common personal surnames and forenames can also be place names or proper nouns, leading to our two other major sources of error. Surnames such as Banks, Black, Cash, Church, Day, Price, Rice, White, Winter, and Young and forenames such as Bill occur quite frequently in everyday language. The most problematic name was Price. Almost all occurrences of this term that were marked as persons turned out to refer to the daily price of the newspaper or the price of a commodity. The term "White" would tag for a color and other times its use would lead to the tagging of expressions such as "White Persons" as a personal name. Similarly, "New National Banks," and "Virginia Banks" are often documented as individuals in our collection, despite the fact that all references typically are to bank buildings.

Occasionally certain grammatical patterns led to the tagging of phrases or expressions as personal names. Whenever two capitalized proper nouns occurred in a sentence, they were typically tagged as a proper name. This same error could also happen whenever a string of nouns appeared separated by commas. Newspapers made frequent use of all capital letters as well as capitalization of proper nouns, depending on the context. This makes the system's use of normally common grammatical or syntactical patterns problematic for newspapers. A number of phrases that were incorrectly tagged by the system as proper names in its first pass include "Medical Students," "Good Harness," and "Bill Head."

Role names, particularly those of royalty, also led to some interesting tagging results. A number of newspaper articles

frequently reference European royalty, so the tagging of role names using preexisting TEI tags such as Lady, Baron, and Lord has been attempted. The role name "Lady" proved most problematic and often the generic use of the word lady, or ladies, such as the expression "Ladies of Richmond" would lead to the tagging of a personal name such as Lady Richmond. In one case where an article detailed that a young lady had been murdered, her identity was tagged as "Lady Shot" since the headline for the article read "YOUNG LADY SHOT" in all capital letters. Another challenging role name was "Miss" where expressions such as "Miss May" or "Miss Semon" had "Miss" tagged as a forename rather than as a role name. Finally, tagging of the rolename General also occasionally leads to problems such as the tagging of the "Quartermaster General" as a general with the unfortunate last name of quartermaster.

4.3 Place Names

We found that some of our most accurate tagging results were with place names, with system performance currently at 97.42%. The types of errors that we found for this type were quite varied. The most significant error was when an entity was correctly classified as a place but the wrong place was subsequently identified such as the labeling of all references to "Manchester, Virginia" as "Manchester, New Hampshire". Another common source of error was the tagging of capitalized nouns as places. Several references to Providence (as in Divine Providence) were tagged as the city in Rhode Island, whereas a reference to the House (as in House of Delegates) was identified as House, North Carolina.

The latter error reflects the problem of using a very large gazetteer such as the Getty Thesaurus of Geographic Names (TGN): big lists accumulate unexpected place names (such as House) that often did not exist in the 19th century. Likewise reliance on the TGN causes a tendency to tag ethnic groups as nations, such as the tagging of several mentions of Europeans as the continent of Europe (European being listed as a synonym for Europe in the TGN).

Personal names were also occasionally tagged as place names, such as the tagging of Sergeant Beverly as Beverly, West Virginia. Many of these errors are difficult to correct automatically. The capitalization of certain nouns such as Providence in text can make it difficult to distinguish when a word is being used as a place name.

The identification of accurate place names in newspapers thus presents a number of unique problems. The TGN does not include local place names such as "Broad Street Hotel," "Schad's Hall," or "Shockoe Slip." Thus part of our work entails identifying those local place names that appear prominently in the newspapers and adding them to our place name authority lists. We have spent a significant amount of time analyzing the problems of identifying this particular type of entity, which has perhaps led to our relatively high accuracy rate.

4.4 Dates

Currently the system tags a variety of dates including all numeric and textual references to days, months, and years. The tagging for this entity type performs quite well with general accuracy levels at about 96.4%. The system currently tags all types of dates including both full and partial expressions, such as the "nineteenth instance of May," "June 5th," and "September 4th 1862." The most common way of expressing dates in these newspapers was referring to individual days as instances, such as "the sale occurred on Monday last, the twentieth inst." The second most common way was to simply refer to a month and day, such as "December 1st." Authors of newspaper articles most likely felt little need to specify fuller dates that included years due to the fact that newspapers typically focused on recent events. Such partial specification can be problematic for our system, due to the many instances in the newspaper where a reference is made to "next Saturday" or the "ninth instance" without any nearby references to the actual month or year.

The largest source of error was when the system failed to tag a complete date reference, or where it would tag "May 2nd 1865" as "May 2nd." Due to unexpected pre-existing markup in the XML text, it would not connect 1865 to the Month/Day. The error rate also reflects simple bugs: we discovered problems with the rule set that led to the tagging of places such as "fourth street" as days of the week or of people's ages such as "in the twenty ninth year of his age" as days of the month, due to the these expression's use of ordinal numbers. Nonetheless, we have focused on tagging this type of entity quite broadly, which has led to largely satisfactory results.

4.5 Products

The automated tagging of products or commodities proved to be very difficult, and our system experienced its lowest level of accuracy with this type of entity at 57.58%. In part this reflects the need of a more comprehensive list of potential products. The current list consists of 1,900 commodities and was derived from an initial survey of the collection. Not only were there commodities in the 19th-century that are no longer sold or consumed today such as "burning fluid," but also the same commodity may go by a number of names or be represented several ways in the text such as "windowglass," "window-glass," or "window pane glass." The identification of 19th-century commodities that no longer exist often involved the use of the Oxford English Dictionary Online to determine what a particular word meant and if it was being used as a commodity. One popular sales term for advertisements in the 19th-century was "furnishing goods," yet the system routinely tagged this as a generic commodity labeled "goods" until we added the full commodity name to our authority list.

Refining the results of this system has at times proved both entertaining and quite challenging. Our system still struggles with commodities that contain multiple terms in their names such as "anthracite coal" and "cider vinegar." These items are frequently missed because the system tags the first as "coal" and the second as "vinegar." This partial tagging of commodity names accounted for about 2% of total errors.

The most significant source of tagging error for our system was that many items that were sold or consumed as commodities and products also occurred frequently as general nouns in everyday language. The examples are numerous and include items such as apparel, bacon, butter, cotton, dress, flour, gold, jewelry, salt, and tobacco. The mistagging of these nouns as products when they were not actually items for sale accounted for about 25% of all errors. Another major source of error was the tagging of government organizations as products, particularly bureaus, houses, and cabinets. Several frequent items for sale included the "Bureau of Conscription," the "Cabinet of the President," and the "House of Representatives." This error should prove relatively easy to fix by restricting the context in which these items tag as commodities.

All other sources of error in the tagging of commodities proved to be relatively minor. The system would occasionally tag personal names as products (e.g. Porter, Wheat, Wood) as well as place names (e.g. "brick house," "State room," "forks in the road") and company names (e.g. tagging "coal" in the "Catawba Coal and Iron Company" as a product).

4.6 Organization Names

We are currently recognizing a very broad range of organization types such as armies, banks, assemblies, banks, commissions, companies, districts, offices and parties, to name only a few, with varying ranges of success. Organization tagging begins by drawing on gazetteers of organizations mined from the corpus and reviewed by editors, then scans for noun phrases with headwords such as "assembly" or "association." Overall 82.14% of the entities tagged are indeed organizations.

Lexical ambiguity is a common source of error. While "commissions," "districts," and "offices" are often actual organizations, these words also occur quite frequently as nouns in everyday language. The largest source of error for this entity category is the tagging of general nouns or expressions as organizations. While we are quite successful at tagging fully specified references to organizations such as the "Young Men's Christian Association," and the "Virginia Legislature," we have less success at avoiding the tagging of expressions that are too general such as "central committees" and "general assemblies." This issue caused over half of all the errors we found.

A number of examples can help to illustrate this persistent problem. Expositions were very difficult to tag due to extensive use of expressions such as "Exposition of Divine Truth" tagging as the "Divine Exposition." While the term "commission" occasionally tagged for an actual military or political commission, these phrases typically involved a commission for a specific person, such as "Willey's commission in the navy." Tagging for "districts" also proved to be problematic. While most of the terms tagged were districts, the districts referred to were places such as "Wheeling District" rather than formal organizations. This organization type also had the habit of tagging actual individuals such as a "district commissioner" or "district attorney." Attempts have also been made at tagging "offices" in order to tag expressions such as "Office of the President." The most frequent use of this term in newspapers, however, was to offices that were being referred to in directions in an advertisement or as a place name, such as the location of a meeting. Examples includes "apply at the "Manager's office" and "across the street from "Cloptin's Office" tagging as organizations.

Attempts at tagging political parties also frequently led to false tags such as tagging expressions such as "old party feeling" as the "Old Party" or a "large party of women" as the "Women's Large Party." Similar problems were encountered when attempting to tag unions. Many of the items the system tagged as a union were place names such as "Union Hill." It also tagged any general use of the word as a noun such as in "numerous union troops," "Strong union speeches," and "prayers for the union."

One of the most frequently found organizations within historical newspapers are small businesses owned by local individuals. Almost half the text of these newspapers consisted of advertisements, purchased by individuals such as lawyers, druggists, and storeowners. Many of these company names consist of a list of surnames such as "Harris, Spicer & Harris." These strings of text currently tend to tag as a list of individual surnames. The variety of naming patterns for small businesses in newspapers was extensive, and the same company name could include abbreviations one time while including the full name another, such as for the "Chas. T. Wortham & Co." which also contained listings as "Charles T. Wortham & Co." While the use of certain patterns such as "& Co" or more specifically the XML string, "& Co" may help to identify text strings as organizations, not all companies contain this expression.

While some more major companies such as the "Powhatan Steamboat Company" and the "Richmond Trunk Factory" are relatively easy to identify, the rules for tagging these names, such as "tag all expressions that occur with the term 'company' and 'factory,'" can have unforeseen problems. The use of terms that are too common or generic as patterns in identifying organization names can lead to the tagging of sentences where factory or company are simply used as nouns.

Another difficulty in identifying company names is that both businesses and military organizations can contain the key term "company" such as "Company A." This is an example of where an authority list generated from a 19th century city directory could be invaluable in providing the system with a list of local business names to tag. A preliminary list of company names has been created while checking the results of automated tagging. Further examination has illustrated that a small number of companies accounted for a significant number of advertisements, which may make creation of a manual list feasible.

The identification of banks can also be particularly challenging, as riverbanks have on more than one occasion been tagged as financial institutions. Examples include the frequently tagged "Bank of the Mississippi" and "Bank of the Rio Grande" when all references were to the rivers. Another issue that came up was the tagging of expressions such as "shares city bank," and "shares continental bank" as full bank names. Financial sections of the newspaper typically listed all share prices in this manner, leading to a number of incorrect tags.

4.7 Street Names and Addresses

One of the most difficult entity types that we had trouble initially tagging were street names, particularly identifying the individual streets listed as intersections in the newspaper text, such as "Marshall and Clay" or "7th and Franklin." In our preliminary work these terms often tagged as standalone surnames rather than street names. The addition of various heuristics to our system that made use not only of a "gazetteer" of street names but also took into account the context (the words which appeared in the text around those street names) has improved our results considerably. For example, we found that if a term such as "Cary" or "Marshall" appeared before a word such as "corner" or "between" it was much more likely to be a street name than a surname. Tagging accuracy for this entity is at over 99%. We currently tag all street names such as "Broad Street" and "Cary Street" as well as individual street addresses such as "148 Main." The most common source of error for this entity type is the tagging of place names as streets, such as when the system tagged the town of Brighton, Massachusetts, as "Brighton Street." Patterns that typically work well can also lead to occasional

errors. While phrases that end in "place" are often street names, we also found a number of street name tags for "Vacant Place."

4.8 Newspapers

The tagging of newspaper titles has relied greatly on the use of historical authority lists. Identification of newspaper names utilizes lists of names drawn from reference works such as the 19th-century Rowell's *Directory of American Newspapers*. Since a large number of the stories in 19th -century newspapers are quoted or borrowed wholesale from other newspapers, there are many titles to identify. The preliminary results for newspaper tagging show a rather low level of accuracy, about 67%. A closer look at the source of error, however, reveals a common problem. Several newspaper names, particularly "Southern Confederacy," "Saturday Morning," "Stars and Stripes," and "Star Spangled Banner," while actual newspaper titles, were also common phrases that were used quite frequently in our newspaper collection. After removing these errors from our results, the tagging accuracy rises to 85.30%. The one mistag "Southern Confederacy" as a newspaper led to over half of all errors we found. A related issue was that many generic expressions referring to the press as a whole such as "Richmond papers" and "Republican Journal" also tagged as individual newspapers. A number of newspaper titles from the 19th-century were general expressions such as "American Union," "Tax Payer," and "Southern Churchmen." One advantage of this kind of problem is that it is easily fixed by updating our authority lists. Once all the errors that involved general expressions such as these were removed from our evaluation, only six actual tagging errors were found, which led to an accuracy rate of over 97%.

Manual evaluation also led to the realization that our system was also failing to tag a number of different newspaper titles. Many newspaper titles involve a place name and then a general noun such as the "Portsmouth Transcript," our system currently tags the place names in these titles as actual places and misses the newspaper title. Newspaper titles in the text can also be displayed in a number of different ways such as the ""Hartford (ct.) courant" and the "Brownville (Texas) Flag." This causes our system to miss the newspaper reference due to the introduction of the state abbreviation.

4.9 Ships

While this type of entity tags at a high level of accuracy at over 93%, our system currently emphasizes recall over precision due to the difficulty of identifying ship names. Since ship names were often place names (e.g., "Brooklyn," "Cumberland," "Monticello"), personal names (e.g., "Louisa," "General Page," "Mary Pierce") or other nouns (e.g., "Warrior," "Ruby," "Leviathan"), the identification of patterns and terms that introduce ship names has been very important. Certain contextual clues are helpful, such as that newspaper articles entitled "SAILED" and "ARRIVED" or "Marine Intelligence" tended to indicate a section that will list recent departures and arrivals of ships. A list of ship "term types" has been created such as "Canal boat" and "packet schooner" to help the system identify ship names. In newspapers, ship names typically appeared after a listing of the ship's type, often in all capital letters.

Manual evaluation has also proved very important in this process. Checking of automated tagging led to the identification of "schr," an abbreviation for schooner, as an important "term type" to add into the list of types for ships. The most frequent error still encountered is the tagging of a verb that follows a ship type as a ship name, leading to voyages on the ships "DESTROYED," "STOPPED," and "MAROONED." This error accounted for over 40% of the mistags found in the current evaluation. Due to the frequent use of both personal and place names for ship names, there were also a number of instances where an individual or a location tagged as a ship, such as Columbia in the "District of Columbia" tagging as a ship name. Another significant error was the abbreviation "Brig Gen"(for Brigadier General) tagging as a ship, due to the fact that the term "Brig" often introduces a ship name. This is an example of when a good general pattern can lead to noise in the results.

4.10 Regiments

One of the most difficult types of entities to fully identify in newspapers are military organizations such as regiments. We have digitized a number of military reference works that include lists of military organizations that may assist in the disambiguation of these organizations in newspaper text. While our accuracy level for this category of entity is relatively high at 91.75%, during the manual evaluation we determined that our current system still misses a large number of military regiments. One major problem is the variations in how military names can appear. The same article will make references to the "1st Reg Volunteers," "1st Reg Va Volunteers," and "Virginia Volunteers," all of which upon closer examination are the same regiment. Currently, we are quite successful at tagging expressions such as the "Third MA Reg" and "Second NY Cav" with their full names of "Third Massachusetts Regiment" and "Second New York Cavalry." Regiments that we missed were often missed due to unusual spacing of state abbreviations, such as the "4th N C" and the "8th S C" not tagging due to the extra space.

The major source of error identified for this entity category was in attempting to identify military organizations with "section" in their name. Tags for the "14th section" typically referred to the 14th section of a legal document or a book. This accounted for over half of all errors. The only other major source of error was the partial tagging of a regiment name, such as "regiment 5" tagging for a reference to the "Fourth and Fifth Regiments." One small source of error was the tagging of expressions such as a "four horse team" and "six horse team" as the fourth and sixth cavalries. Part of our current task is to identify those sections of the newspaper that typically provide a listing of military units, in order to examine patterns of how military units are typically labeled in newspaper texts.

One particular challenge has been in deciding what level of specificity we want to support in the tagging of military organizations. Items such as "French fleet" and "French riflemen" are currently tagging but often lead to tagging of generic expressions rather than actually military organizations. Evaluation of several newspaper issues also led to the identification of a number of military terms that were not tagging and needed to be added to the list of military term types such as "battalion," "greys/grays," "cadets," and "troops," for we were missing expressions such as "Carolina grays." Evaluation also led to the identification of several negative phrases to be added to a list of stop-words including, "national guard hat," "such rifles," and "colt's rifles."

4.11 Railroads

Railroads proved to be relatively easy to tag due to the fact that railroad names, such as the "Orange and Alexandria Railroad" were frequently spelled out in the newspaper text. This type of entity currently tags with an accuracy rate of about 89%. The only major source of error identified for this category was the failure of the system to associate partial railroad names with their full entities. There were a number of tags for the "Danville Railroad" when the full entity that should have tagged was the "Richmond and Danville Railroad." A frequent problem was that some of the more common railroad names would be abbreviated in the text, such as the "Orange and Alexandria Railroad" being referred to as the "Orange and Alex R.R." or at times just the "Orange and Alex." We are currently trying to identify the different ways in which railroad names can be expressed. Automatic tagging, however, many never be able to resolve all errors. Occasionally entertaining expressions were mistagged, such as an article entitled "Shocking Railroad Accident" tagging as the "Shocking Railroad." Since these article titles followed the same grammatical patterns as actual railroad names, most of these mistags have had to be identified through manual evaluation.

5. CONCLUSIONS

The developers of digital libraries and domain specialists may or may not play an active role in developing named entity recognition systems. Only domain specialists working with digital libraries, however, can assemble the knowledge sources that we will need if we are to see sophisticated named entity systems emerge that can adapt themselves to the needs of particular communities working with a range of different collections. We need bigger authority lists and probably more refined rule sets, but the biggest progress will most likely come from using many rich knowledge sources as training data. The results with 19th-century English language materials presented here build on work with pre-modern Greco-Roman materials and lead us to the following conclusions:

• Simple authority lists of names are not always enough. We mined the names of Civil War era ships from Admiral David D. Porter's *Naval History of the Civil War* (1886) but the range of proper names reused as ship names was so great the extensive (if not comprehensive) list picked up too many errors. It was better to mine the text for patterns such as "the cruiser NAME" than to draw upon a large set of brief, ambiguous names.

• The longer the entity name, the more effective the list. This principle is hardly a surprise but worth stressing. We do not (yet) have a comprehensive list of railroads but have generated a list by mining the corpus for patterns ending in "Railroad/RR." We expect a comprehensive list of railroads to be particularly effective.

• Even long entity names need to be checked. We find ships with names such as "the City of Atlanta" and newspapers with names such as "the Stars and Stripes." We need to check longer keys for ambiguity.

• Shorter but more historically relevant lists may be better than longer modern ones. While we currently rely upon the TGN to identify new place names, we first consult a corpus driven gazetteer of 5,300 places – a process that provides better results than simply using the TGN. This gazetteer was derived from a Civil War atlas, a list of towns large enough to have newspapers in 1870, and additional places noted as we worked on the general 19th-century American collection. We also hope to replace the one million plus entry TGN with the 75,000 entry 1855 *Harper's Gazetteer of the world* to support more accurate tagging of historical place names that no longer exist.

• We need our knowledge sources to be more robust than simple lists. In the case of places, the 1855 *Harper's Gazetteer* lists 150 places named Washington. We need to use geographical location, population size, home state and county and other features to provide more precise matches. Likewise we need to be able to use foundation dates of organizations, birth dates of people, coordinates of locations, and dates of affiliations to further assist in disambiguation.

• We need heterogeneous as well as domain specific knowledge sources. A gazetteer may provide detailed information about places and a biographical encyclopedia about people, but separate resources may not help us determine whether a given Washington is a person or a place. A comprehensive encyclopedia that includes people, places, and organizations, can provide clues to help us determine which entries are more likely (e.g., entries with longer articles are probably more likely).

6. REFERENCES

[1] Alani, Harith, et.al. (2003). Automatic ontology-based knowledge extraction from web documents, *IEEE Intelligent Systems*, 18(1), January-February, 14-21.

[2] Allen, Robert B. (2005). A focus context browser for multiple timelines. *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 260-1.

[3] Allen, Robert B. and John Schalow. (1999). Metadata and data structures for the historical newspaper digital library. *Proceedings of the eighth international conference on information and knowledge management*, 147-53.

[4] Allen, Robert B. (2004). Query interface for an event gazetteer. *Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries*, 72-3.

[5] Arlitsch, Kenning and John Herbert. (2003). Digitalnewspapers.org: The digital newspapers program at the University of Utah. *Serials Librarian*, 47, (1/2), 2-6.

[6] Arlitsch, Kenning and John Herbert. (2004). Microfilm, paper, and OCR: Issues in newspaper digitization. *Microform and Imaging Review*, 33 (2), 58-67.

[7] Bates, M. J., D.N. Wilde, and S Siegfried. (1993). An analysis of search terminology used by humanities scholars: The Getty Online Search Project report number 1. *Library Quarterly*, 63(1), 1-39.

[8] Bontcheva, Kalina, et. al. (2004). Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering*, 10 (3/4): 349-373.

[9] Bontcheva, Kalina, et. al. (2002). 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.

[10] Buchanan, George, et. al. (2005). Information seeking by humanities scholars. *Proceedings of the European Conference on Digital Libraries*, 218–229.

[11] Choi, Youngok and Edie M Rasmussen. (2003). Searching for images: The analysis of users queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology*, 54(6): 498–511.

[12] Deegan, Marilyn, et. al. (2002). The British Library newspaper pilot. http://digitalcooperative.oclc.org/digitize/britishlibrarynewspaper.html.

[13] Deegan, Marilyn. et. al. (2002). Digitizing historic newspapers: progress and prospects. *RLG Diginews*, 6 (4). http://www.rlg.org/preserv/diginews/diginews6-4.html#feature2.

[14] Duff, Wendy and Catherine Johnson. (2003). Where is the list with all the names: information seeking behavior of genealogists. *American Archivist*, 66 (1), 79-95.

[15] Duff, Wendy M. and Catherine Johnson. (2002). Accidentally found on purpose: Information-seeking behavior of historians in archives. *Library Quarterly*, 72 (4), 472-499.

[16] Gilboe, Lynda James. (2005). The challenges of digitization: libraries are finding that newspaper projects are not for the faint of heart. *Serials Librarian*, 49 (1/2), 155-63.

[17] Hoekstra, Rik. (2005). Integrating structured and unstructured searching in historical sources. *Proceedings of the XVI international conference of the Association for History and Computing,* 149-55.

[18] Jones, Steve, et. al. (2004). Searching and browsing in a digital library of historical maps and newspapers. *Journal of Digital Information*, 6 (2), Article No. 324. http://jodi.tamu.edu/Articles/v06/i02/Jones1/.

[19] Kesse, Erich. (2004). Ephemeral Cities. *RLG Diginews*, 8 (6). Retrieved from http://www.rlg.org/en/page.php?Page_ID=20492#article0

[20] King, Edmund. (2005). Digitisation of newspapers at the British Library. *Serials Librarian*, 49 (1/2), 165-181.

[21] Lesk, Michael. (2005). The qualitative advantages of quantities of information: bigger is better. *Journal of Zheijang University Science,* 6A(11): 1169-78.

[22] Macqueen, Donald S. (2004). Developing methods for very-large-scale searches in Proquest Historical Newspapers collection and Infotrac The Times Digital Archive: The case of two million versus two millions. *Journal of English Linguistics*, 32 (2), 124-43.

[23] Murray, Ray L. (2005). Toward a metadata standard for digitized historical newspapers. *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 330-1.

[24] Oard, Douglas W. (2004). Language technologies for scalable digital libraries. Presented at the International Conference on Asian Digital Libraries, New Delhi, India.

[25] Popik, Barry. (2004). Digital historical newspapers: A review of the powerful new research tools. *Journal of English Linguistics,* 32 (2), 114-23.

[26] Porrata, Aurora Pons, et. al. (2003). Building a hierarchy of events and topics for newspaper digital libraries. *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14-16, 2003. Proceedings*, 588-596.

[27] Readings, Reg and Mark Holland. (2003). 'The Thunderer' on the web - The Times Digital Archive 1785-1985. *Library + Information Update*.

[28] Rosenzweig, Roy. (2003). Scarcity or abundance? Preserving the past in a digital era. *American Historical Review*, 108 (3), 735-762.

[29] Shoemaker, Robert. (2005). Digital London: Creating a searchable web of interlinked resources on eighteenth century London. *Program: Electronic Library and Information Systems*, 39 (4), 297-311.

[30] Smith, David and Gregory Crane. (2001). Disambiguating geographic names in a historical digital library. *ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, 127–136.

[31] Smith, David. (2002). Detecting and browsing events in unstructured text. *SIGIR02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 73–80.

[32] Smith, David. (2002). Detecting events with date and place information in unstructured text. *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint conference on Digital libraries*, 191–196.

[33] Terpstra, Judith A. K; et. al. (2005). The Tundra Times newspaper digitization project. *RLG Diginews*.

[34] Witten, Ian, et. al. (2004). Text mining in a digital library. *Int. J. On Digital Libraries*, 4(1):56–59.

[35] Zweig, Ronald W. (1998). Lessons from the Palestine Post project. *Literary and Linguistic Computing*, 13 (2), 94-7.