

# Evidence Finding using a Collection of Books

Marc-Allen Cartright, Henry Feild, and James Allan

Center for Intelligent Information Retrieval  
Dept. of Computer Science  
Univ. of Massachusetts  
Amherst, MA 01003  
{irmarc,hfeild,allan}@cs.umass.edu

## ABSTRACT

This paper introduces the task of *Evidence Finding*, a novel information retrieval task that uses books—a traditionally more trust-worthy source of information—to help provide evidence to support a statement. What makes this evidence-finding task different from other tasks, such as the related INEX *Prove It* task, is that both the statement for which evidence is sought *and its context* are given to the search system. A practical application of this system is to provide supporting or refuting evidence from books for a statement made within a Wikipedia article, using the entire article as contextual support for query generation. We provide details of this task as well as an analysis of a number of retrieval methods that address this task.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*search process*

## General Terms

Algorithms, Design, Experimentation

## Keywords

Evidence finding, statement support, supporting evidence, book retrieval

## 1. INTRODUCTION

Fact checking is the task of verifying whether an assertion is true or at least backed by evidence. We are interested in the latter, finding evidence that supports (or denies) an assertion in a document. For example, an assertion that the American Declaration of Independence was signed in 1776 is supported by a trustworthy source (like a page in a book) saying as much, and is refuted by a source that states that it was signed in a different year.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*BooksOnline'11*, October 24, 2011, Glasgow, Scotland, UK.  
Copyright 2011 ACM 978-1-4503-0961-5/11/10 ...\$10.00.

This task, which we call Evidence Finding (EF), is useful in any setting where text can benefit from citations—e.g., writing reports, scientific papers, or even opinion pieces. In this paper we describe early work toward applying this task to find support for assertions in Wikipedia articles.

EF is a similar task to fact-checking, however we restrict the setting as follows:

1. Returned evidence is sought only from a fixed collection — here, a set of published books.
2. The factual assertion is accompanied by the surrounding context.

An example of using the EF task in conjunction with Wikipedia is depicted in Figure 1, which shows an assertion highlighted in a small, dotted red box. Its immediate context is the paragraph in which it resides, highlighted with a slightly larger dashed blue box. A slightly larger context is the section in which the assertion is made, highlighted with a solid green box. The whole article could also be used as context, though this is not depicted in the figure. The article and section title are also readily available context for the assertion.

Although this paper provides an abbreviated investigation of EF, we show 1) a technique for candidate fact generation that we plan to further develop to create future EF test collections, and that 2) initial results indicating that the use of additional query structure and fact context provide useful information to improve query results.

## 2. RELATED WORK

EF is related to several other information retrieval tasks. In this section, we briefly describe these other tasks and the key distinctions between them and EF.

Likely the most related area of study is restatement retrieval [2, 9, 11], which tries to find sentences or passages that restate a given piece of text. EF differs in two ways. First, the task is not necessarily to find a restatement of the assertion. For example, sources that refute the fact will not necessarily restate the assertion. Second, rather than confining relevant information to a sentence or short passage, relevant documents within EF can be up to a page.

EF and question answering (QA) tasks have related goals, but the tasks are distinct. The aim of QA [5] and fact retrieval [4] is not to retrieve relevant documents from a collection, but rather to provide an explicit answer to the query, which is a question. QA involves establishing a list of answer candidates and picking the best one(s) based on evidence.

Franklin made the case for repeal, explaining the colonies had spent heavily in manpower, money, and blood in defense of the empire in a series of wars against the French and Indians, and that further taxes to pay for those wars were unjust and might bring about a rebellion. Parliament agreed and repealed the tax, but in the *Declaratory Act* of March 1766 insisted that parliament retained full power to make laws for the colonies "in all cases whatsoever".<sup>[19]</sup>

**1767–1773: Townshend Acts and the Tea Act**  
*Main articles: Townshend Acts and Tea Act*  
*Further information: Massachusetts Circular Letter, Boston Massacre, and Boston Tea Party*

In 1767, the Parliament passed the *Townshend Acts*, which placed a tax on a number of essential goods including paper, glass, and tea. Angered at the tax increases, colonists organized a boycott of British goods. In Boston on March 5, 1770, a large mob gathered around a group of British soldiers. The mob grew more and more threatening, throwing snowballs, rocks and debris at the soldiers. One soldier was clubbed and fell. All but one of the soldiers fired into the crowd. 11 people were hit; three civilians were killed at the scene of the shooting, and two died after the incident. The event quickly came to be called the *Boston Massacre*. Although the soldiers were tried and acquitted (defended by John Adams), the widespread descriptions soon became propaganda to turn colonial sentiment against the British. This in turn began a downward spiral in the relationship between Britain and the Province of Massachusetts.<sup>[31]</sup>

In June 1772, in what became known as the *Gaspée Affair*, a British warship that had been vigorously enforcing unpopular trade regulations was burned by American patriots including John Brown. Soon afterward, Governor Thomas Hutchinson of Massachusetts reported that he and the royal judges would be paid directly from London, thus bypassing the colonial legislature.

On December 16, 1773, a group of men, led by Samuel Adams and dressed to evoke American Indians, boarded the ships of the government-favored *British East India Company* and dumped an estimated £10,000 worth of tea from its holds (approximately £636,000 in 2008) into the harbor. This event became known as the *Boston Tea Party* and remains a significant part of American patriotic lore.<sup>[33]</sup>

**1774–1775: Quebec Act and the Intolerable Acts**  
*Main articles: Quebec Act and Intolerable Acts*

The *Quebec Act* of 1774 extended Quebec's boundaries to the *Ohio River*, shutting out the claims of the 13 colonies. By then, however, the Americans had little regard for new laws from London; they were drilling militia and organizing for war.<sup>[34]</sup>

The British government responded by passing several Acts which came to be known as the *Intolerable Acts*, which further darkened colonial opinion towards the British. They consisted of four laws enacted by the British parliament.<sup>[35]</sup> The first was the *Massachusetts Government Act*, which altered the Massachusetts charter and restricted town meetings. The second Act, the *Administration of Justice Act*, ordered that all British soldiers to be tried were to be arraigned in Britain, not in the colonies. The third Act was the *Boston Port Act*, which closed the port of Boston until the British had been compensated for the tea lost in the *Boston Tea Party* (the British never




Figure 1: An example of an assertion within the Wikipedia article “American Revolution”, along with its surrounding context.

In contrast, the primary goal of EF is to provide documents that can then be used as supporting or refuting evidence for a given statement. EF could be used to find that evidence for an answer candidate within a QA system, but the task of candidate generation and evidence rating is specific to QA.

We view the EF task as an extension to the INEX “Prove It” task<sup>1</sup>. In “Prove It”, the task items involve factual assertions made in isolation, which lack surrounding context. In EF, we assume that the factual assertion is a complete statement in need of confirmation or refutation, however we also assume that the assertion occurs within a larger context that is available to the query processing system. This setting makes the EF task distinct from other fact-related tasks.

### 3. METHODOLOGY

We view the experimental setup as follows: *Given an assertion and its context, the find pages in a book that either support or refute the assertion.* We use this section to discuss the data used in our study, the methods used to generate factual assertions, and the techniques used to generate queries based on the assertions and the available context.

#### 3.1 Data and Software

We use the collection from the INEX Book Track, which consists of approximately 50,000 scanned books with section markup provided by Microsoft as part of the track. We use the Galago<sup>2</sup> retrieval system which provides us with a flexible platform for trying various automatic retrieval tasks. Galago makes use of a query language similar to Indri and INQUERY, where such operations as proximity and field op-

<sup>1</sup><http://www.inex.otago.ac.nz/tracks/books/books.asp>

<sup>2</sup><http://www.galagosearch.org/>

erators can be used to express certain levels of structure in a given query. The INEX collection comes with metadata describing author, title, year, and so on. However, the metadata is inaccurate or missing for a large number of the books and cannot be relied on. Because that is a common problem for large collections such as this, we have chosen to explore retrieval techniques that do not use that metadata. We do, however, use the metadata fields to approximate the publication dates of the books in the collection in order to find Wikipedia articles appropriate for the collection.

#### 3.2 Generating New Facts

We selected verified statements from Wikipedia pages to serve as assertion candidates for the EF task. We first selected several topics (i.e. articles) from Wikipedia that overlapped the estimated publication dates contained in the INEX collection. Table 1 shows a brief summary of the selected topics. We then automatically extracted sentences that contained references at the bottom of the wikipedia page, using these sentences as our assertions. Figure 2a shows an example factual assertion extracted from a Wikipedia article, and 2b shows the corresponding reference for that assertion. Notice the hyperlinked terms/phrases in Figure 2a. We extract these terms, and note that they have special status in the wiki page text, as they were used in the description of the attached hyperlink. In the remainder of the paper, we refer to these terms as wikiwords.

Once we have an assertion (e.g., the text within the dotted red box in Fig. 1), we extract several additional components from the Wikipedia article, including:

- the wikiwords
- the text of the enclosing paragraph (e.g., the text within the dashed blue box in Fig. 1)

(a)

The Eighteenth Amendment was the result of decades of effort by [temperance movements](#) and at the time was generally considered a [progressive amendment](#).<sup>[1]</sup>

(b)

## References

[edit]

- <sup>^</sup> Hamm, Richard F. (1995). *Shaping the Eighteenth Amendment: temperance reform, legal culture, and the polity, 1880-1920*. UNC Press Books. p. 228. ISBN 9780807844939. OCLC 246711905 [g](#).

**Figure 2:** (a) One of the assertions automatically extracted from the Wikipedia article on the American Revolution and (b) a supporting reference given by the author of the assertion.

Topic
American revolution
Battle of Gettysburg
William Shakespeare
Incandescent light bulb (History of the light bulb)

**Table 1: Selected topics from Wikipedia. Each topic corresponds to a distinct page in Wikipedia bearing the topic name as the title.**

- the text of the enclosing section (e.g., the text within the solid green box in Fig. 1)
- the enclosing section title
- the title of the wikipedia article

Each of these additional components is made available to the query processing system. We assume that “in the wild”, this information would accompany the factual assertion as input to the system.

For this pilot study, we extracted 12 facts from 4 Wikipedia pages. These can be seen in Table 4.

### 3.3 Query Generation

A primary focus of this investigation is to gain insight into what kinds of query structure are important for EF. Towards this end, we try several different manual and automatic query generation techniques, which we describe here.

#### 3.3.1 Manual Runs

We manually performed anaphora resolution to ensure that each of the candidate facts could stand alone as assertions without requiring references to additional sentences or pages to form a complete assertion.<sup>3</sup>

Each of the authors manually generated a Galago query based on each anaphora-resolved fact. The query-writers were not able to edit their queries after they were submitted to the system, therefore this simulates the popular “ad-hoc” query style used in the TREC Web Track<sup>4</sup>. These manual runs provide some provisional evidence of the ability of three trained IR system users to generate queries using a highly

<sup>3</sup>We note that anaphora resolution falls under the broader approach of coreference resolution, which could be used to fully associate all references in the entire Wikipedia article. We chose not to use full coreference resolution due to the expense and we feel that anaphora resolution of *only* the factual assertion will provide most of the benefit of coreference resolution at a fraction of the cost.

<sup>4</sup><http://plg.uwaterloo.ca/~trecweb/2011.html>

expressive query language. In the Results section, we label the manual runs as M1, M2, and M3.

The manually generated queries served as a model for developing automatic query generation methods. Table 3 shows three example manual queries. We found that the manual queries tended to use a mix of phrases, unordered windows, synonyms, and keywords external to the fact itself, although no weighting was used. In addition to keyword expansion we also encountered cases of “concept expansion”—stating a particular entity in multiple ways. Due to the difficulty in automating concept expansion, we leave implementation of this technique to future work.

#### 3.3.2 Automatic Runs

We use the multinomial Language Model [10] as our baseline, which we label as BOW in results. Although several techniques are known to outperform this model, it provides a good comparison point for the expected performance of a retrieval system without the use of any additional structure.

Using the manual queries as a guide, we know the important aspects of query generation includes the use of phrases, anaphora resolution, and terms related to the fact, but not included within the fact. We strive to capture these aspects in the automatic query techniques in the following manner.

To make use of phrase information, we use the Sequential Dependence Model of Metzler and Croft [8], which incorporates both term proximity and term dependence. We refer to this run as SDM.

To resolve anaphora, we assume that the subject of the anaphora will appear in the paragraph. We include SDM elements from the context paragraph in addition to the SDM of the fact itself, giving less weight to the paragraph elements. As mentioned earlier, we did not use automatic coreference resolution due to its complexity, though this would be an obvious step to try in the future. We call this method SDM+P.

To address the use of related terms, we use three different methods for incorporating terms from the surrounding context. We assume that terms and phrases located closer to the fact are more useful and that the article and section titles include important text. The first technique uses SDM in addition to ordered windows of bigrams extracted from wikiwords found within the assertion’s source paragraph. We call this SDM+PWW. We then created two additional generation techniques that build on the SDM+P method described earlier. The first includes SDM elements from the assertion’s source section, placing more of the weight on terms from the assertion’s paragraph and the most weight on terms from the assertion itself; this is referred to as SDM+S. The

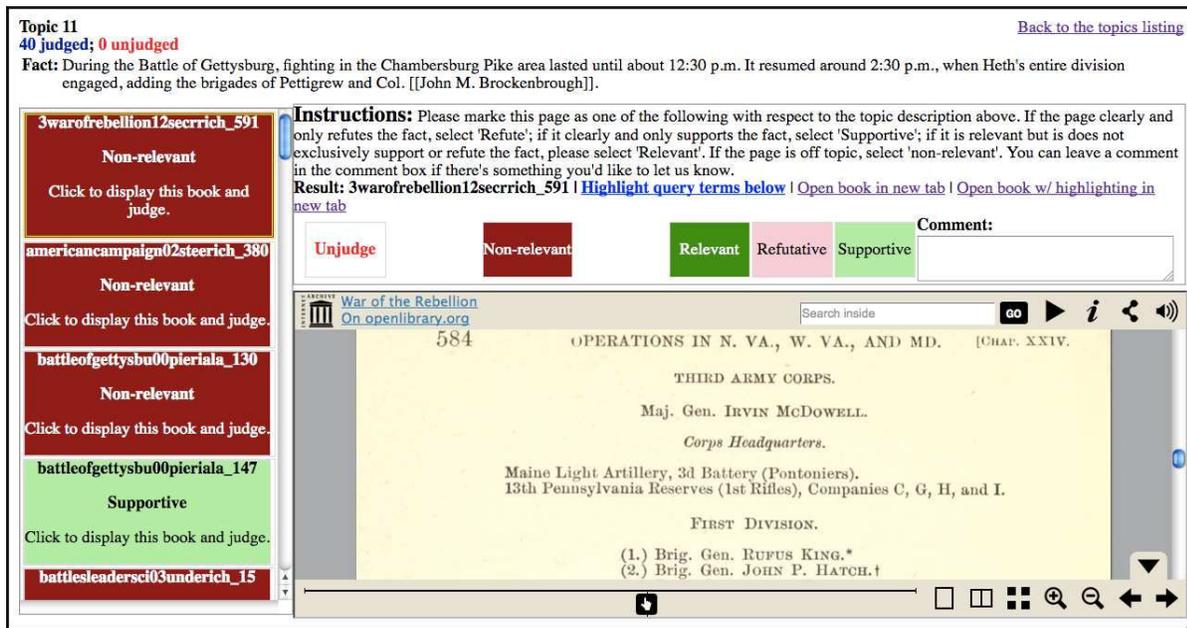


Figure 3: Interface used for judging documents.

final method uses SDM+S and adds bigrams taken from the section and article title. We call this SDM+T.

We did not attempt to automatically expand queries with synonyms, though synonyms were used in the manual queries. Future methods should explore using utilities such as WordNet<sup>5</sup> to introduce synonyms.

### 3.4 Mixed Runs

We applied two additional approaches that involve manual pre-processing to resolve anaphora followed by automatic processing using BOW and SDM to form the finalized query. These methods are labeled as BOW+AR and SDM+AR, respectively.

## 4. EVALUATION

We labeled the candidate pages using one of the four relevance classes as defined by the INEX “Prove It” task:

**Non-Relevant** The page does not contain any relevant content.

**Relevant** The page pertains to the assertion, but does not directly support or refute the assertion.

**Refutative** The page refutes the factual assertion.

**Supportive** The page confirms the factual assertion.

Note that for the purposes of this task, we can group the classes into “relevant” and “non-relevant”, since we are primarily concerned with finding a page that helps to confirm or refute the factual assertion. Using this reasoning, we can place all of the classes except “non-relevant” into the “relevant” class, resulting in a set of binary judgments. For the evaluations measures, we are most interested in mean reciprocal rank (MRR) and success at rank 5 (Success@5). We

<sup>5</sup><http://wordnet.princeton.edu/>

define Success@ $n$  to be 1 if at least one relevant document appears in the first  $n$  ranks, and 0 otherwise. A high value for either of these suggests the technique will pull pages with relevant evidence into the top-most ranks.

We pooled the top 10 results from each run on the 12 topics, and then gathered judgments on a majority sample of the resulting pool using computer science graduate students as assessors. In order to perform assessments, we made use of the Internet Archive<sup>6</sup> reader application to provide the original text to the assessor. A screenshot of the judgment system used is shown in Figure 3. Using this system, we were able to judge 468 book pages that were retrieved using the queries generated as described in Section 3.3. To generate the performance statistics below, we used the `ireval` program provided by Trevor Strohman.<sup>7</sup> We used the program considering the judgments as binary judgments, grouping the classes as described above.

## 5. RESULTS

As mentioned earlier, a sample of the top 10 results returned for each run were judged, and as such the numbers presented here from the \*@10 metrics should be considered a lower bound. We looked at the upper bound for these measures by assuming all non-judged documents are relevant; though the exact results change, the relationships between methods remains roughly the same.

### 5.1 Anaphora Resolution

The step of resolving anaphora in the queries provided considerable improvement over the original queries. Both the BOW+AR and SDM+AR runs outperformed their non-resolved counterparts. Note however that even the BOW+AR and SDM+AR runs did not perform as well as most of the manual runs (although they did outperform one

<sup>6</sup><http://www.archive.org>

<sup>7</sup><http://ciir.cs.umass.edu/~strohman/code/>

Method	MRR	Success@5	Success@10	P@10	MAP	NDCG
BOW	0.4824	0.7500	0.8611	0.4000	0.2658	0.4966
SDM	0.5037	0.8333	0.8889	0.4500	0.2937	0.5329
SDM+P	0.6242	0.8750	0.8889	0.4833	0.3453	0.6079
SDM+S	0.5828	0.8750	0.8889	0.4500	0.3084	0.5843
SDM+T	0.6245	0.8750	0.8889	0.4583	0.3415	0.6105
SDM+PWW	0.5176	0.8333	0.8889	0.4333	0.2994	0.5449
BOW+AR	0.6369	0.8333	0.9167	0.5583	0.3907	0.6538
SDM+AR	0.6833	0.9167	0.9167	0.5833	0.4219	0.6812
M1	0.5585	0.9167	0.9444	0.4250	0.2349	0.5072
M2	0.8361	1.0000	1.0000	0.5750	0.3904	0.6683
M3	0.8167	0.9583	1.0000	0.5500	0.3588	0.6339

Table 2: Results of various methods used for EF.

manual run). In addition, SDM+P, which used phrases from the surrounding paragraph in order to help automatically resolve anaphora, also performed better than SDM or BOW. This evidence suggests that anaphora resolution in the factual assertion itself is a potent source of information, but still more information may be incorporated to improve results even further.

## 5.2 Use of Additional Context

We can see that including terms, phrases, and wikiwords from the paragraph surrounding the fact helped in the automatic runs. SDM+P and SDM+PWW outperform both BOW and SDM. However, including terms and phrases from context further away from the assertion (i.e., in the section text and titles: SDM+S and SDM+T) yields diminishing returns over only using the paragraph context. This suggests that including the additional term may have caused topic drift. Note that using additional context does not perform as well as the manual runs.

## 5.3 Differences Between Manual Runs

Although all of the manual runs performed reasonably well, we can see a large disparity in performance between the three runs, across all measures. This led us to analyze these queries and their results in greater detail, to better determine what may have caused such disparity. As an example, we choose the assertion:

This dumping of tea from British East India Company ships into the Boston harbor became known as the [[Boston Tea Party]] and remains a significant part of American patriotic lore.

resulting in the three queries being generated by three different authors, shown in Table 3. The double brackets around the phrase “Boston Tea party” indicates that a link exists to a Wikipedia page focused on the hyperlinked phrase. For those unfamiliar with the Galago query language, we now briefly describe each query component used. The *ordered window*, or `#od:n`, operator enforces the rule that the contained terms must occur in text the order they are specified under the operator, with no more than  $n$  spaces between each term. Therefore `#od:4( boston tea party )` enforces the rule that an “occurrence” includes the term *boston*, then after no more than 4 additional terms, the term *tea*, and then after no more than another 4 terms, the term *party*. The *unordered window* operator, indicated using `#uw:n`, is defined differently. We define a window of size  $n$ , and require that all terms under the operator may appear in any order, but

they must occur within that window for the window to be a valid “occurrence”. Therefore, `#uw:10( boston harbor )` enforces the rule that the terms *boston* and *harbor* must occur within 10 terms of each other, but they may occur in neither order. Note that pairings are performed over the occurrences and not the unique terms, therefore the string *boston harbor boston* will produce 2 occurrences under this instance of the operator.

We can now gain a better intuition behind the intents of the three users. Each user thought it important to look for the entire phrase “Boston Tea Party”, however User A thought it important to allow for non-phasal terms between the given terms, whereas Users B and C afforded much less latitude. User B also included the phrase “boston harbor”, but was willing to allow for a significant number of terms to occur between the two terms. User C also thought “harbor” was an important unigram. User A appears to approach the topic from one side, thinking that it would be important to reference the idea that the Boston Tea Party was part of American culture—specifically, as lore. In contrast, User B thought it more relevant to note the added association that the tea was British in origin, and it involved dumping. None of these ideas were expressed in an overly structured manner (all the additional terms were unigrams, and therefore assumed to be independent), but the intention can be easily inferred.

Looking at a brief summary of the results in Table 3, we see that User B was significantly more successful in retrieving relevant documents, particularly in the first 5 ranks of retrieval: User A retrieved only 1 relevant document, whereas User B retrieved 5. In order to understand the cause of such a great disparity, we analyze the top 5 documents of each run to see what User B apparently exploited that User A did not. We analyzed 3 sets of documents: the top 5 documents retrieved by User A’s run, the top 5 retrieved by User B’s run, and an 4 additional documents that were judged relevant that were returned in the top 1000 of User B’s run, but not User A’s run.

The additional unigram terms seemed to have made the most difference between the runs at the top 5 ranks. The non-relevant documents of User A included documents that simply mentioned the Boston Tea Party, but either in a listing of events or under a larger theme of American lore. Therefore including *american* and *lore* did little to help find relevant documents, and seemed to actually introduce noise. In contrast, the unigrams included by User B seemed to consistently count towards the score of the documents retrieved. Additionally, the “Boston Harbor” phrase helped in several

User	Run	Query	RR	P@5	AP
A	M1	#combine( #od4( boston tea party ) american lore )	0.25	0.20	0.05
B	M2	#combine( #uw:10( boston harbor ) #od:1( boston tea party ) tea british dump )	1.00	1.00	0.47
C	M3	#combine( #od:1(boston tea party) harbor #od:1(east india company) )	1.00	1.00	0.39

**Table 3: Three different queries generated for the same factual assertion, along with several retrieval statistics.**

instances. In the documents that only appeared for User B, the use of *british* and *dump* seemed particularly useful. Although none of the results are unexpected, it illustrates that good term selection is vital to good performance, as inclusion of the wrong terms can often cause topic drift in the query, and produce erroneous results. Both users omitted any form of the phrase “East India Trading Company”, however User C did not. We hypothesize that it may have introduced a certain number of non-relevant documents that discuss the East India Trading Company in other contexts.

## 6. DISCUSSION

Based on the initial experiments, we have already learned that techniques such as anaphora resolution, concept selection, and use of phrases greatly improve retrieval performance for the EF task. Manual generation of queries using these techniques provided good retrieval performance, and our efforts to automatically replicate these techniques recovered most of the gains from the manual generation method.

### 6.1 Evaluation Critique

Using the relevance classes provided, assessment of the retrieved pages from the INEX collection proved to be exceedingly difficult at times. Our technique for generating factual assertions occasionally generated facts that could not be directly supported by the available collection. As an example, consider the fact

During the Battle of Gettysburg, fighting in the Chambersburg Pike area lasted until about 12:30 p.m. It resumed around 2:30 p.m., when Heth’s entire division engaged, adding the brigades of Pettigrew and Col. John M. Brockenbrough.

which was generated from the “Battle of Gettysburg” topic. Almost all of the pages retrieved in pooling discussed the Battle of Gettysburg, and therefore were in some sense topical. However given the breadth of coverage on the topic (the reference for the above fact is a book analyzing the first day of the battle alone), it is reasonable to assert that simply discussing the Battle of Gettysburg is insufficient for relevancy. In this case, we could require that the topic of the candidate document must at least discuss the fighting focused in the Chambersburg Pike area.

We have come to the conclusion that extracting factual assertions from Wikipedia articles in this way allows assessors the opportunity to quickly generate a candidate list of facts that are suitable for experimentation. However in light of the difficulty we encountered during the judging period some manual supervision is required to generate the final list of topics. Additionally, judgments may ultimately need to incorporate the idea of factual “nuggets” similar to the TREC QA Track [5] in order to assess how much of the fact has been confirmed or refuted. The spread of a topic must also be taken into consideration—The Battle of Gettysburg is a very large topic, with multiple books discussing various

aspects of it. In other cases, the topic may not be as extensive, therefore finding *any* source discussing the topic may be beneficial.

### 6.2 Future Directions

Based on the preliminary results we have demonstrated here, we see two major threads of research that we intend to pursue. The first involves refinement and extensions of the EF task, while the second focuses on further work involving the use of books in relation to Evidence Finding.

#### 6.2.1 Evidence Finding Going Forward

Although we have gained some insight into what components of a query are useful for Evidence Finding, a considerable amount of work remains to better understand and utilize this process. The manually generated queries made use of additional structure, such as term equivalence classes (synonym structure) and additional query expansion simply using domain knowledge. Additionally, although we used term and phrase weighting techniques in the automatically generated query runs, we made no attempt to perform any training to optimize these weights. Previous research has shown that training these weights can provide substantial gains in retrieval performance [3, 7]; however we would like to expand these ideas to incorporate the additional context available in this EF task.

Near-term future steps also include generalizing our approach over a larger portion of Wikipedia. Wikipedia has become a ubiquitous source of information for millions of users and it is important to maintain the quality of the information in articles on the site. The work introduced here speaks directly towards addressing this issue and could be used to find sources for statements that have a “citation needed” footnote. Over time, we would like to generalize the task to other online sources where an assertion may be made, and supporting evidence is required. Such sites as political websites, blogs, and news aggregation sites could all benefit from EF. In a larger setting, we can imagine this task generalizing to corroborating statements in academic papers, or transcripts of speeches from political leaders. Therefore, we see the scope of this task extending outside of the online world, and having potential impact in multiple disciplines.

#### 6.2.2 Books in the Future

We earlier noted that for the most part, we ignored the metadata included with the books in the collection due to significant noise. We believe that if the quality of the metadata were higher, we could make use of the metadata fields to improve retrieval performance further [6]. Therefore, we see focus on better metadata curation as paramount to increasing the utility of online books in future search tasks.

In this instance we considered books to be an authority to verify Wikipedia statements, we must also consider that statements in books themselves may be in need of verification. Therefore, if we turned Evidence Finding inward to the books themselves, we find another task, *fact provenance*,

which would focus on finding the first time a factual assertion was made in published work. The ability to infer such information could serve as an important tool for literary and historical scholars. We note that this is in some sense similar to topic detection and tracking [1], though in the case of books, the material is not streaming and can be analyzed holistically.

In addition to the ideas mentioned above, we also invite contributions from the BooksOnline community for further ideas for evolving Evidence Finding and its interaction with book collections.

### 6.3 Limitations

As it stands now, Evidence Finding and our proposed techniques to address the task have several limitations, some of which we note here.

**Assertion composition**—facts or statements may be composed in several ways. Some may consist of one atomic fact that is easily verified by a single source. Others may be compound statements, requiring a different source to verify each constituent assertion. One possible method to tackle this problem is to automatically decompose assertions into atomic facts and retrieve supporting documents for each separately. Alternatively, the evaluation for EF could be modified to allow for multiple sources for each atomic fact.

**Anaphora resolution**—we used several runs that used manual anaphora resolution as an upper bound on performance. Human-equivalent resolution performance may not be achievable, even with state of the art coreference resolution systems. An in-depth exploration into anaphora resolution techniques is required to understand how realistic it is to achieve the performance of the manual resolution runs using fully automatic techniques.

**Evaluation**—it is unclear how to best evaluate the EF task. As discussed earlier, it may be appropriate to use the notion of “nuggets”, as in the TREC QA task. Additionally, as mentioned above, if a compound assertion is being analyzed, it may be beneficial to allow a document to support one atomic fact while refuting another, i.e., to consider each atomic fact independently.

## 7. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF CLUE IIS-0844226 and in part by NSF grant #IIS-0910884. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

## 8. REFERENCES

- [1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, et al. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, volume 1998, pages 194–218. Citeseer, 1998.
- [2] N. Balasubramanian and J. Allan. Syntactic query models for restatement retrieval. In *String Processing and Information Retrieval*, pages 143–155. Springer, 2009.
- [3] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 491–498, New York, NY, USA, 2008. ACM.
- [4] W. S. Cooper. Fact retrieval and deductive question-answering information retrieval systems. *J. ACM*, 11:117–137, April 1964.
- [5] H. Dang, D. Kelly, and J. Lin. Overview of the trec 2007 question answering track. In *Proc. of TREC*. NIST, 2007.
- [6] J. Kim, X. Xue, and W. B. Croft. *A Probabilistic Retrieval Model for Semistructured Data*. Lecture Notes in Computer Science. Springer, 2008.
- [7] M. Lease, J. Allan, and W. Croft. Regression rank: Learning to meet the opportunity of descriptive queries. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 90–101. Springer Berlin / Heidelberg, 2009.
- [8] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.
- [9] A. Mittelbach, L. Lehmann, C. Rensing, and R. Steinmetz. Automatic detection of local reuse. In M. Wolpers, P. Kirschner, M. Scheffel, S. Lindstaedt, and V. Dimitrova, editors, *Sustaining TEL: From Innovation to Learning and Practice*, volume 6383 of *Lecture Notes in Computer Science*, pages 229–244. Springer Berlin / Heidelberg, 2010.
- [10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.
- [11] J. Seo and W. B. Croft. Local text reuse detection. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 571–578, New York, NY, USA, 2008. ACM.

<b>Topic</b>	<b>Assertion</b>
American revolution	The peace treaty with Britain, known as the Treaty of Paris, gave the U.S. all land east of the Mississippi River and south of the Great Lakes, though not including Florida.
American revolution	This dumping of tea from British East India Company ships into the Boston harbor became known as the Boston Tea Party and remains a significant part of American patriotic lore.
American revolution	Many of the descendants of Black Loyalists of Nova Scotia still live in Sierra Leone, as well as other African countries.
American revolution	Calling themselves "Federalists," the nationalists convinced Congress to call the Philadelphia Convention in 1787.
American revolution	While the Battle of Bunker Hill was a British victory, it was at a great cost; about 1,000 British casualties from a garrison of about 6,000, as compared to 500 American casualties from a much larger force.
Battle of Gettysburg	During the Battle of Gettysburg, fighting in the Chambersburg Pike area lasted until about 12:30 p.m. It resumed around 2:30 p.m., when Heth's entire division engaged, adding the brigades of Pettigrew and Col. John M. Brockenbrough.
Battle of Gettysburg	The morning of June 27, Maj. Gen. Jubal Early departed for adjacent York County, Pennsylvania.
Battle of Gettysburg	The inconclusive Battle of Brandy Station, the largest predominantly cavalry engagement of the war, proved for the first time that the Union horse soldier was equal to his Southern counterpart.
William Shakespeare	Shakespeare's blank verse is often beautiful, but its sentences tend to start, pause, and finish at the (End-stopping) end of lines, with the risk of monotony.
William Shakespeare	In his will, Shakespeare left the bulk of his large estate to his elder daughter Susanna.
Incandescent light bulb	In 1841, Frederick de Moleyns of England was granted the first patent for an incandescent lamp, with a design using platinum wires contained within a vacuum bulb.
Incandescent light bulb	Light loss in incandescent lamps is due to filament evaporation and bulb blackening.

**Table 4: Assertions used for the study.**