Supporting Factual Statements with Evidence from the Web

Chee Wee Leong
Computer Science and Engineering
University of North Texas
Denton, TX, 76207
cheeweeleong@my.unt.edu

Silviu Cucerzan
Microsoft Research
Microsoft Way
Redmond, WA, 98052
silviu@microsoft.com

ABSTRACT

Fact verification has become an important task due to the increased popularity of blogs, discussion groups, and social sites, as well as of encyclopedic collections that aggregate content from many contributors. We investigate the task of automatically retrieving supporting evidence from the Web for factual statements. Using Wikipedia as a starting point, we derive a large corpus of statements paired with supporting Web documents, which we employ further as training and test data under the assumption that the contributed references to Wikipedia represent some of the most relevant Web documents for supporting the corresponding statements. Given a factual statement, the proposed system first transforms it into a set of semantic terms by using machine learning techniques. It then employs a quasi-random strategy for selecting subsets of the semantic terms according to topical likelihood. These semantic terms are used to construct queries for retrieving Web documents via a Web search API. Finally, the retrieved documents are aggregated and re-ranked by employing additional measures of their suitability to support the factual statement. To gauge the quality of the retrieved evidence, we conduct a user study through Amazon Mechanical Turk, which shows that our system is capable of retrieving supporting Web documents comparable to those chosen by Wikipedia contributors.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation, Measurement

Keywords

Fact verification, supporting evidence, Wikipedia, Web search, Web references, semantic term extraction.

1. INTRODUCTION

Web search engines have become the *de facto* standard for retrieving relevant data for informational needs that can be expressed in short queries, such as "Barack Obama major" or "Obama biography". However, enabling search engines to provide evidence for a complex factual statement, such as "In 1981, Obama transferred to Columbia University in New York City, where he majored in political science with a specialty in international relations" is largely an unexplored research problem. Retrieving evidence for such statements typically requires that users formulate short queries that contain words likely to occur all together on relevant Web pages. Furthermore, the snippets returned by Web search engines focus on the words of those short queries, making it hard for a user to determine if the retrieved Web pages support the factual statement without further navigation to the actual Web page content.

Is what one reads also what one believes to be true? The ability to verify factual information quickly is crucial in many business scenarios (e.g., politics, media, stock market, retail), as well as for the day-to-day needs of individual users. Fact verification has become particularly important due to the increased popularity of blogs, discussion groups, and social sites, as well as of encyclopedic collections that aggregate content provided by many contributors, such as Wikipedia and IMDB. While such collections generally provide accurate information, each factual statement may require additional verification/support due to the nature of the open contribution process. Currently, Wikipedia requires contributors to provide reliable sources for the edited content whenever possible, particularly for factual statements of controversial nature. This requirement presents both new annotated data opportunities and data annotation tool needs. On one hand, it resulted in numerous Web page citations being added to Wikipedia, which provides research opportunities for investigating a large collection of factual data annotated with references to supporting Web evidence. On the other hand, due to the size of the task, the majority of facts stated in Wikipedia still lack proper references; thus, building a tool that helps contributors and editors easily retrieve and/or improve such references can have a positive impact for Wikipedia and the Web community.

Our current objective is to investigate the retrieval of Web evidence to support any general factual statement. Since our investigation starts from the Wikipedia collection, this also provides a concrete method that can be employed by Wikipedia contributors to create relevant citations to Web sources. While assertions made in any particular statement may be questionable, we do not address here the task of determining their validity. The focus of the paper is the retrieval of the best supporting Web evidence for a factual statement as given. Future work will address models for retrieving also contradicting evidence to provide counterclaims to assertions of input factual statements.

We use Wikipedia as a starting point to derive a large collection of factual statements with supporting Web evidence, which we em-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

ploy further as training and test data while we study the retrieval of supporting evidence for any given factual statement. We use this collection to investigate the lexical matching between paired Web documents and factual statements and propose a system for rewriting factual statements into Web queries. We then investigate how to combine and re-rank the results of these queries to bring to the top the best supporting documents for the input factual statements.

2. RELATED WORK

To our knowledge, no previous work has targeted the effectiveness of information retrieval over the Web using complete factual statements. However, the investigated task has substantial common grounds with several other tasks previously tackled by the academic community, such as textual entailment, question answering, summarization, query rewriting and long queries, as well as topic-based retrieval and passage retrieval (TREC HARD 2003-2005[1]).

There exists a large body of work on processing long Web queries (some of which could be factual statements), which employs methods similar to those described here. Kumaran and Allan [16] present various methods for selecting a covering set of terms from a long query. They measure the information overlap between terms using their co-occurrence in a corpus, but this measure was found to perform rather poorly. In later work [17], they test interactive systems in which the user can choose from a few highly-ranked query reductions. They show that reducing a query to its (predicted) best performing sub-query can boost performance. Bendersky and Croft [1] address the task of identifying key concepts in long natural language queries from the TREC corpus. They identify the noun phrases in queries as a collection of concepts of interest and use a classifier to identify key concepts among those. They show that document ranking can be improved by boosting the rank of retrieved documents containing the identified key concepts. Bendersky and Croft [2] also propose a partitioning of the long queries from a major Web search engine into groups and analyze their distribution. Huang and Efthimiadis [12] describe a taxonomy of query reformulation operations of syntactic nature (word reordering, stemming, punctuation removal, etc.) and a rule-based classifier to detect each type of reformulation. Jones et al. [13] use Web search session data to learn term substitutions in order to generate query rewrites. However, their post-hoc annotation effort shows that the proposed system struggles to generate rewrites for queries with low frequencies, which is typical for long queries.

The task we are investigating can also be seen as related to question answering (TREC QA 1999-2007[2]). However, there are important differences between the two. On one hand, we work under the assumption that all facts of interest are contained in the input statement; thus, finding the supporting evidence is comparable with the step of mapping a question and a candidate answer retrieved from the Web to a document in the TREC collection by Web-based QA systems such as [4]. On the other hand, most factual statements that we employ are substantially more varied than the factoids targeted by TREC QA, in terms of length, domain, and syntactic and lexical complexity. Since Wikipedia is collaboratively edited by very many Web contributors, these factual statements encompass a large number of stylistic choices. Also, because they occur in running text and are judged as requiring verification for their claims by content contributors, we believe that they more representative of real-world applications. While there have been several efforts to extract structured information from the Web, such as [10], these have however focused on extracting very specific, typically binary relationships (e.g., "is a", "headquarters of", and "invented by") rather than complex and diverse factual statements. Hence, the methods used there are not directly portable to the task we are investigating.

[1]http://ciir.cs.umass.edu/research/hard/

[2]http://trec.nist.go/data/qamain.html

In late 1988, Obama entered Harvard Law School. He was selected as an editor of the Harvard Law Review at the end of his first year,[34] and president of the journal in his second year.[30][35] During his summers, he returned to Chicago, where he worked as an associate at the law firms of Sidley Austin in 1989 and Hopkins & Sutter in 1990.[36] After graduating with a J.D. magna cum laude from Harvard[37] in 1991, he returned to Chicago.[34] Obama's election as the first black president of the Harvard Law Review gained national media attention[30][35] and led to a publishing contract and advance for a book about race relations,[38] which evolved into a personal memoir. The manuscript was published in mid-1995 as Dreams from My Father.[38]

Figure 1: Example of paragraph with external references from the Wikipedia page for "Barack Obama". Some sentences do not have pointers to external resources (such as the first sentence), while others contain multiple such references, which could occur inside the sentence (such as "[37]") or at the end of the sentence (such as "[36]").

Another similar line of work is textual entailment, which was first proposed by Dagan et al. [9] to account for the variability of semantic expression in languages. Directional entailment relationship can be achieved from different text variants to the same target meaning, and applications such as question answering, multi-document summarization and information retrieval often require this kind of inference to be more effective. Kozareva and Montoy [15] formulate the entailment problem as a classification task, where features extracted from a pair of texts are used with machine learning techniques to build a classifier for positive/negative entailment relationship. Other investigated approaches deal with semantic inferences using logic and ontology-based reasoning [25], as well as tree-edit-distance algorithms over syntactic representations of the text [14].

3. DATA COLLECTION

As stated in the Introduction, we employ Wikipedia to extract a labeled data set for training and evaluating our models. Wikipedia can be seen as a large repository of topic-driven factual statements contributed by Web volunteers. To ensure the quality of the collection, Wikipedia asks contributors nowadays to support factual statements by providing citations to external sources. Specifically, the Wikipedia manual recommends that all quotations and any challengeable claim be attributed in the form of an inline citation. In particular, exceptional claims should benefit of the support of *trustworthy evidence*, which the Wikipedia manual defines as information provided by third-party sources with a good reputation for fact-checking. Since determining the nature of claims or sources is subjective, Wikipedia contributors are asked to use common sense in deciding when and what type of external evidence is needed to support a claim. For example, statements about celebrities, such as "Lohan was voluntarily fitted with a SCRAM bracelet to monitor her sobriety" are more susceptible to being inaccurate and they would benefit of additional support. Likewise, blogs, tabloids, and fan pages are more likely to provide inaccurate information, and typically are not regarded as reliable external sources.

Figure 1 shows a text excerpt from the Wikipedia page for Barack Obama, in which most sentences contain citations to external resources. Some of these citations occur in the middle of a sentence (e.g., "[37]"), and are assumed to provide support for the claims

<*subject*> Martin Wendt </*subject*>
<*statement*> As of 2009, his total live tournament winnings exceed $900,000. </*statement*>
<*ref*> http://pokerdb.thehendonmob.com/player.php?n=29302 </*ref*>

<*subject*> Gamma-Aminobutyric acid </*subject*>
<*statement*> In general, GABA does not cross the blood-brain barrier, although certain areas of the brain which have no effective blood brain barrier, such as the periventricular nucleus, can be reached by drugs such as systematically injected GABA. </*statement*>
<*ref*> http://physrev.physiology.org/content/79/2/511.full </*ref*>

<*subject*> Jason Beghe </*subject*>
<*statement*> Kennedy and Beghe often hung out together outside the Metropolitan Museum of Art and in Central Park, and were monitored by Kennedy's Secret Service detail. </*statement*>
<*ref*> http://www.cnn.com/US/9907/22/jfk.growing.up.in.nyc/ </*ref*>

<*subject*> Los Angeles </*subject*>
<*statement*> Los Angeles enjoys a Mediterranean climate, with an average of 35 days with measurable precipitation annually. </*statement*>
<*ref*> http://www.weatherbase.com/weather/weather.php3?s=159227 </*ref*>

<*subject*> Murder </*subject*>
<*statement*> More than 500,000 people have been killed by firearms in Brazil between 1979 and 2003. </*statement*>
<*ref*> http://news.bbc.co.uk/2/hi/americas/4628813.stm </*ref*>

<*subject*> Armenian Genocide </*subject*>
<*statement*> Despite his previous public recognition and support of Genocide bills, as well as the election campaign promises to formally recognize the Armenian Genocide, the U.S. President, Barack Obama, although repeating that his views on the issue have not changed, has thus far abstained from using the term 'genocide'. </*statement*>
<*ref*> http://www.politifact.com/truth-o-meter/promises/obameter/... </*ref*>

Figure 2: Examples of data points extracted from the Wikipedia collection. The *statement* is a sentence or text fragment ending with a citation, the *subject* is the title of the Wikipedia page from which the statement is extracted, and the *ref* field is the URL of the citation in Wikipedia.

made up to that point in the sentence.[3] They will be referred to as *in-sentence citations* henceforth. Other citations, which occur at the end (e.g., "[36]") and will be referred to as *end-of-sentence citations*, may provide support for the entire preceding sentence, the paragraph, or, when the sentence contains in-sentence citations too, the end fragment (i.e., from the previous citation on). However, this distinction is not easy to make automatically. Because our goal is to investigate the retrieval of supporting evidence for well-formed factual statements, we decided to employ in our extracted data the whole preceding sentence corresponding to such citations. When multiple citations are used in support of a text fragment (e.g., "[30]" and "[35]" in Figure 1), we construct multiple data points with the same subject and statement, but different references.

Figure 2 shows several examples of data points that we extracted from Wikipedia. Each data point consists of a statement, a subject, and a reference, which would be explained as follows.

The *statement* is a Wikipedia text fragment for which a citation to an external Web source is provided. While our goal has been to extract full sentences as statements, automatic sentence segmentation is not 100% accurate, and thus, some data points in the extracted set may involve multiple sentences or sentence fragments.

The *subject*, which is the title of the Wikipedia page from which the statement was extracted, is employed to provide a context for the statement, mainly because of the extensive use of pronominal references to the subjects of the Wikipedia pages. We decided to preserve the original text of the statement and add the subject field in order to avoid deriving semantically incorrect sentences through anaphora resolution methods. As shown in Table 1, the subject is present as an exact string in the statement only one third of the time (35.8%). In an additional 6.6% of statements, all the words that compose the subject are present in the statement, but not together as a phrase.

Subject	As exact string	Component words only
in Statement	35.8%	6.6%
in Reference	60.1%	11.4%

Table 1: The percentages of time the subject of a data point in our training collection is present in the corresponding statement and in the referenced Web document.

The *reference* is the URL of the Web source used as citation for the extracted statement. Not all Wikipedia citations are referencing online sources or contain the URL of the source; only those for which we could identify a well-formed URL were extracted as data points in our collection.

Table 1 shows that 60.1% of the Web documents we retrieved from the reference URLs contain the subject as a phrase; consequently, a Web search query that contains the subject in between double-quotes would retrieve the Web page in the filtered set on average 6 out of 10 times. An additional 11.4% contain all the words that compose the subject, which means that querying for the subject would retrieve the referenced Web page in the filtered set 71.5% of the time (60.1 + 11.4).

As shown in the examples from Figure 2, we obtain factual statements for a diversity of subjects, from people and places to historical events, artifacts, and common concepts.

The data set employed in our experiments was extracted from the June 20, 2011 version of Wikipedia, which is about 32GB and contains almost 3.7 million topic pages. We employed the wikifier described in [8] to extract all instances of entity/concept mentions in the text while preserving the original references in the text (commonly referred to as *Wikipedia interlinks*). Then, we segmented the text into sentences and extracted those sentences that contain in-sentence and/or end-of-sentence citations. We retain only those citations that contain a URL of a Web resource. This process resulted in a data set of more than 600,000 data points. The experiments described further in this study employed three distinct random subsets, of 123,318 data points for training, 4,988 data points for development, and 1,179 data points for the final test. Only the data points for which the cited URL was reachable at the time we performed the experiments were kept into these sets, and the reported sizes reflect this additional filtering step. [4]

To gain insight into the citation portfolio of Wikipedia, we first analyzed an additional random set of 100 data points, which cover various domains and topics. Of the 100 hyperlinked Web sources, only 87 were reachable at the time we performed the experiments. The type of cited documents ranged from personal blogs to governmental Web sites, and to Web pages of mainstream news agency. As expected, the overwhelming majority (97%, or 84 of 87 the reachable sources) were in English. 76 of these documents were single-page, while 11 of them were multiple-page documents. 97% of the references were text/html-based documents, 3% were videos, and none were Web images by themselves.

Surprisingly, we noticed in our pilot study that many document sources do not appear to obey the trustworthy, third-party requirement of Wikipedia, even though they do provide support for the factual statements for which they are cited. Furthermore, the non-negligible number of inactive/broken hyperlinks means that readers

[3]Depending on syntactic dependency rules/direction and citation conventions, this assumption may not hold for some languages other than English or other large document collections.

[4]Data employed in our evaluation experiments is available at http://research.microsoft.com/en-us/people/silviu/data/CIKM12.

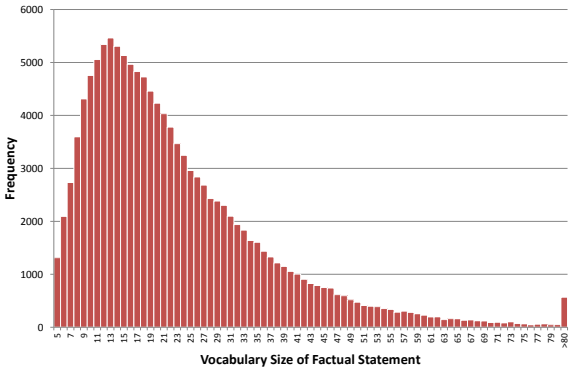

Figure 3: Histogram of the vocabulary size (distinct, lemmatized words) for the statements in our training set. The median vocabulary sizes for *in-sentence* and *end-of-sentence* factual statements are 11 and 13, respectively.

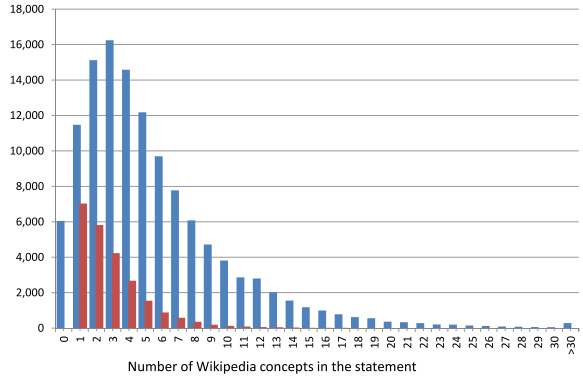

Figure 4: Histogram for the number of Wikipedia entities in the statements in our training set (light blue bars), in conjunction with the number of times all the entities from statements are also present in the referenced documents (dark red bars).

may be unable to verify the validity of a substantial number of factual statements through the citations provided in Wikipedia. Nevertheless, Wikipedia appears to be overall an extremely valuable resource for both training and testing a factual support system.

4. FRAMEWORK

Generating support evidence from the Web for a factual statement can be seen as a special type of search task, in which the information conveyed in the input statement is used to query the Web and retrieve documents that entail the input statement.

We start our investigation by making a strong assumption about the quality of the citations extracted in our data set from Wikipedia, namely that they refer to the most relevant documents available on the Web that support the corresponding statements. Obviously, while this is a desideratum for Wikipedia, such an assumption is likely to be unrealistic in practice (as already hinted by our pilot study). However, under this assumption, we can readily formulate an initial goal for our study: we would like to learn a retrieval model that, for as many statements in our data set as possible, retrieves the corresponding references as the top-ranked documents from the Web.

The naive method of using the statement directly as a query to a typical Web search engine is not an effective solution for this task. Table 2 summarizes the empirical results obtained by submitting as queries to the Bing API a subset of 10,000 factual statements picked at random from our training set. Only a very small percentage of the supporting documents cited in Wikipedia can be retrieved in the top position, while the vast majority of them do not get retrieved in the top 50 results. This shows that the cited Web documents are likely not to contain the Wikipedia statement verbatim. Instead, they may contain a paraphrased form, may contain the supporting data in tabular format, or entail the statement without explicitly summarizing the statement's content. Explicitly discarding the stopwords and functional words in the comprehensive list of Lewis et al. [20] before querying Bing Search proves beneficial for retrieval performance, as indicated by the improvement from 1.36% to 3.48% for retrieving the Web reference document at the top position.

Unfortunately, this lexical mismatch problem is not easily solvable through straightforward synonymy-based methods. As shown in Figure 4, the factual statements in our collection are typically long, and employing every possible combination of synonyms for all words in the original statement to generate alternate queries is not a practical approach. Moreover, many of the queries derived from the original statement by employing every word or a synonym of the word end up containing unlikely lexical combinations and do

Method	Top 1	Not Retrieved
Verbatim	1.36	95.67
Discard stopwords	3.48	90.08
Oracle: overlapping words	8.48	71.67

Table 2: The percentage of Wikipedia references that are retrieved by Bing on the top position for the corresponding statement when using as query (a) the verbatim form of the statement; (b) the statement from which stopwords were removed; and (c) only the words that are shared between the statement and the referenced document (the latter is seen as an oracle).

not retrieve any search result or produce only irrelevant results due to the combined effects of imperfect synonymy and homonymy.

Table 2 also shows the performance of an oracle system, which employs a query comprising all words that are common to the statement and the referenced Web document. Without having access to the various data streams the commercial engine employs for documents, such as the collection of anchor text (of links pointing to Web pages), this oracle constructs the longest possible query with words from the input statement that can retrieve the corresponding referenced document (in the filtered set). By using this oracle's queries, the Bing search engine retrieves the referenced document at the top position about 8.5% of the time. Surprisingly, the referenced documents are not retrieved in the top 50 even by the oracle queries more than 70% of the time.[5]

In the remainder of this section, we discuss a system that transforms a factual statement into a set of terms likely to occur in a supporting document. We propose a methodology of training this system on the dataset extracted from Wikipedia. False positive errors in term prediction (i.e. incorporating a term which does not appear in the supporting Web document into the predicted set) result in queries that would not retrieve the supporting document. Consequently, we investigate several term removal strategies, in which a number of terms are optionally dropped from the predicted set before performing Web searches, as well as the combination of the results obtained from these searches. We then employ a re-ranker based on text similarity measures with the original factual statement to determine the most relevant Web document(s) for the input statement. Figure 5 gives an overview of the proposed system.

[5]The Bing API provides a maximum of 50 Web retrieve documents for a single query request. We chose to use the allocated search traffic to perform queries on more data points rather than to run multiple searches on fewer data points.

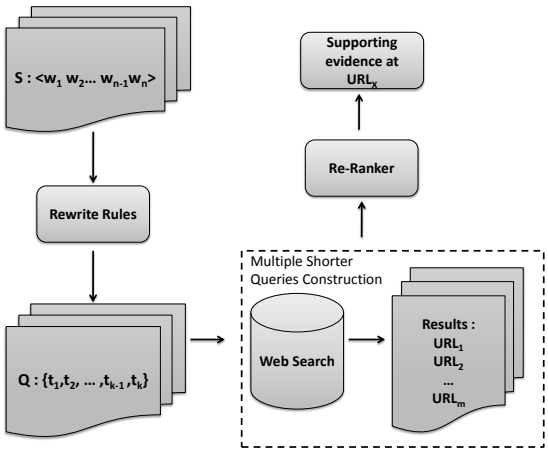

Figure 5: Overview of the proposed system: an input statement is transformed into a sequence of terms from which rewrite rules and *Drop-k* processes select subsets to be used for querying the Web; the search results are then aggregated to obtain a ranked list of supporting documents.

4.1 Factual Statement Rewriting

While the meaning of a sentence (or text fragment) is encoded in the meaning of its component words, not all words are equally important for understanding the information the sentence conveys. Additionally, as shown by Mihalcea and Leong [22], humans do not need to scan through the possible meanings of individual words in order to understand a sentence. Rather, they could identify important keywords in the sentence, and contextualize the concepts expressed by these keywords using their background knowledge. Similarly, we attempt to identify in each statement a subset of terms that capture its important semantic parts and use them subsequently as building blocks for constructing Web queries to retrieve documents relevant for that statement.

Let $S =< w_1 ... w_n >$ denote a factual statement comprising n words. We first perform entity/concept recognition, followed by stopword removal and word lemmatization. Given that our particular corpus of statements is extracted from Wikipedia, we preserve the interlinks created by Wikipedia editors, and identify other interlinked entities in the remainder of the text by using the entity linker described in [8]. For the sentence S, we then extract the set of *semantic terms* $T_S = \{t_1, ..., t_k\}$, defined as the set of all linked/identified entities and lemmatized word types after removing stopwords and function words. To facilitate this step, we employed the stopword list provided by Lewis et al. [20] and a lemmatizer derived from a large English thesaurus using simple regular-expression-based inflection rules.

Let U denote the URL pointing to the referenced Web document corresponding to the training statement S. For each term t in the statement S, we compute its number of occurrences in this target document. The set of semantic terms that occur at least once will be called the overlapping set O, $O \subseteq T_S$. They will also be referred further as the *overlapping terms*. Under the assumption that the Web search engine employed uses for retrieval/filtering only the content of the indexed documents and performs regular keyword searches (corresponding to the *AND* operator of a boolean search engine), O can be viewed as the maximal set of terms from S that can be used to retrieve the target document. Using additional terms from the statement (i.e., those not in the document) should result in a recall error. In general, the more terms from this set are employed, the smaller the filtered set of documents obtained will be.

	% citations retrieved in the Top 1	Top 10	Top 50
In-sentence			
No stopwords	2.04	7.14	12.24
Rewrite	5.10	13.27	20.41
Oracle (overlapping terms)	6.12	16.33	28.57
End-of-sentence			
No stopwords	3.61	7.77	9.71
Rewrite	3.89	10.92	16.47
Oracle (overlapping terms)	8.70	21.37	28.31

Table 3: Query rewrite performance for retrieving the Wikipedia referenced document as one of the top Web results.

	α	γ	P	R	F
In-sentence	0.680	11	0.808	0.616	0.761
End-of-sentence	0.610	13	0.791	0.644	0.756

Table 4: Performance of predicting overlapping terms in the constructed query.

While this does not guarantee that the position of the target document in the obtained ranked list improves with each new addition of terms from O, we note a strong correlation between the number of overlapping terms employed in the query and the document position in the ranked list of Web search results. In particular, we verified this hypothesis by running an experiment on a sample subset of 1000 data points in which we employed the queries comprising all overlapping terms, as well as those in which we dropped from the overlapping set one and two random terms. The positive correlation coefficients between query length and rank (inverse) of the target document are 0.956 and 0.949, respectively.

The main underlying step of our learning algorithm is to transform a statement S into a query Q by selecting the semantic terms most likely to retrieve the target U. More formally, we seek a function $g : T_S \rightarrow \{0,1\}^{|T_S|}$ that assigns 1 to terms that are likely to be in the target document and 0 to the other terms. This task is analogous to the problem of keyword extraction (e.g., [7]). We experimented with a number of syntactic, encyclopedic, and heuristic features for the terms in T_S, of which the most effective are reported in Table 5. These features are used with an implementation of a boosted decision trees classifier similar to that described by Burges [6]. To train this learner, we assigned the value 1 to all terms in O and 0 to all the terms in $T_S \setminus O$ for all data points in the training set. In our experiments on the development set, this learner outperformed all other machine learning methods that we experimented with (logistic regression, averaged perceptron, and support vector machines).

The learner outputs a probability $p(g(t) = 1)$ for each semantic term $t \in T_S$. We construct a query Q associated with the input statement S by sorting the terms from T_S in the reverse order of the output probabilities $p(g(t) = 1)$, and then by adding to Q terms t from the sorted list as long as $p(g(t) = 1) \geq \alpha$ and $|Q \cup t| \leq \gamma$. α and γ are model parameters dependent on the type of factual statement (inline or end-of-line). We optimized the values of α by using $F_\beta = (1 + \beta^2)\frac{PR}{\beta^2 P + R}$, a weighted harmonic mean of precision and recall, for which we compute *precision* as the percentage of query terms that belong to O (i.e., correctly classified as occurring in the referenced document), and *recall* (or *sensitivity*) as the percentage of semantic terms from O included in the query Q. We set $\beta = 0.5$ to favor precision over recall, since the inclusion of non-overlapping terms is more likely to compromise the retrieval of the supporting document. Note that setting β in this manner also avoids constructing highly precise queries, which tend to retrieve

Features	Description
Offset(t, S)	the backward distance (measured in semantic terms) from the end of the statement (the citation marker in Wikipedia)
TermCount(S)	the number of unique semantic terms
Keyphraseness(t)	$\frac{N_{sd}(t)}{N_s(t)}$, where $N_{sd}(t)$ is the # of times t is an overlapping term, $N_s(t)$ is # of statements in which t occurs in training
WikitextFreq(t)	the term's frequency in the associated Wikipedia article (or, generally, the article where the term is found)
POS(t, S)	the part-of-speech tag (Penn Treebank) of the term in the statement
TFIDF(t, S)	The product of the term's frequency in factual statement and its IDF derived from two corpora, BNC and NYT
Multiword(t)	binary value indicating whether semantic term is a multi-word compound

Table 5: Features used in the proposed learning algorithm for factual statement rewriting

supporting Web documents narrowly restricted to contain just a few selected semantic terms present in the factual statement. We set the value of γ as the average number of overlapping terms between the paired statements and referenced documents from the development set for each type of factual statement investigated.

While conceptually we follow a similar approach to [18, 26] for query subset construction from verbose queries, we account for the fact that our dataset contains a diverse set of well-formed sentences or text fragments of various lengths, which encompass a very large number of Wikipedia topics, as well as syntactic constructions and vocabulary sets significantly different from those typical for long Web search queries. Previously employed models such as conditional random fields (CRF) focus on dependency clues between query words to construct query subsets, which are appropriate for long Web queries that contain concatenations of terms, in order to narrow down the users' search intent. In our work, we hypothesize that good retrieval performance can be obtained through an accurate semantic correspondence between the factual statements and the target Wikipedia references, one that is expressed by a set of methodologically crafted semantic terms as close to O as possible.

In general, we observed in our experiments that the retrieved set of Web results is not independent of the term order in the query, most likely because Web search engines employ proximity features in their ranking algorithms. To help replicating the experiments reported in our work, it is necessary to note that, while the terms in the query set are selected in the decreasing order of the probability computed by our classifier, the actual queries we submit to the Web search engine preserve the order of terms from the original statement (to avoid additional noise from term reordering).

4.2 Employing Multiple Shorter Queries

4.2.1 Drop-k Search

As with any learning algorithm, the predictions made by our classifier are imperfect regardless of the type of features or the number of training instances. The average precision and recall obtained by our term classifier / query constructor are shown in Table 4. Since precision and recall are inversely related, attempting to improve one is generally done at the expense of the other. Arguably, in our process of predicting overlapping semantic terms, precision is preferred over recall to avoid the use of irrelevant terms for document retrieval, but targeting very high precision leads to constructing very short queries, which lead to rather general searches in which the documents retrieved, while relevant to the subject, do not provide accurate support for the target factual statement. Therefore, we attempt to improve the overall precision at a fixed recall rate by running multiple Web searches, which translates to an increased computational effort. Consider a scenario in which our classifier has a precision of $P = 0.8$, which means that 20% of query terms selected by our system are false positives. In other words, for an hypothetical average query Q of length ten, $|Q| = 10$, two of its terms do not occur in the target document. Without knowing which are those terms, we could run multiple searches in which we remove any combination of two terms from the query, which requires running $\binom{10}{8}$ = 45 Web searches in order to submit to the search engine the query that contains only overlapping terms.

We investigate a novel query modification paradigm, in which we perform dropping k semantic terms from the query generated using the rewriting rules, where the k terms are selected according to a utility scoring function in a *quasi-random* manner (i.e., randomly selected according to topical likelihood.) By performing multiple *Drop-k* searches in parallel, we are able to increase the chance of querying the Web using a subset of terms that are all correctly predicted as overlapping, while also allowing for a more "diversified" search process, in which dead-end scenarios caused by incorporating a non-overlapping term is prevented. The proposed parallelized Drop-k search is shown in Algorithm 1.

Algorithm 1 Drop-k Search

Input : Original Query Terms $Q = \{t_i | i = 1..N\}$
Input : Query Term Utility $L = \{l_{t_i} | i = 1..N\}$
Output : Set of Modified Q = $\{Q'_i | Q'_i \subset Q, i = 1..N\}$
Output : Set of N Search Results = $\{R'_i | i = 1..N\}$

Determining k
1: Compute average precision P of all Q_j, such that $\|Q_j\| = N$, using development dataset
2: Number of terms to drop, $k = \lceil N(1-P) \rceil$

Tagging t_i with Probabilities
1: Compute $S = \sum_{i=1}^{N} l_{t_i}$
2: Compute $Pr = \{p_i | p = 1..N\}$, where $p_i = l_{t_i}/S$

Constructing N Query Sets
1: **for** i = 1 to N **do**
2: $Q'_i = \{\}$
3: **for** 1 to k **do**
4: Randomly pick a term t from Q according to Pr, such that $t \in Q$ but $t \notin Q'_i$
5: $Q'_i = Q - \{t\}$
6: **end for**
7: **end for**

Performing N Parallel Searches
1: **for** i = 1 to N in parallel **do**
2: $R'_i = Search(Q'_i)$
3: **end for**

4.2.2 Topic Modeling

The investigated algorithm requires the use of a utility function to discriminate against query terms likely to be non-overlapping. Specifically, it computes the inverse of a measure of plausibility of a term to be included in the query. As a result, our Drop-k strategy is *quasi-random* in nature, where the least plausible terms are likely to be dropped. Generally, while such a utility function could favor a specific class of terms e.g. numerical tokens (e.g., "1.2%" and "$12mil."), we focus on selecting topic-specific terms (e.g., "election day") in our work. We employ the established Pachinko

	% citations retrieved in the		
	Top1	**Top10**	**Top50**
In-sentence			
Rewrite	5.10	13.27	20.41
+ Drop-k uniform	5.78	12.93	19.05
+ Drop-k PAM	6.12	14.29	21.43
+ Drop-k uniform parallel	8.84	18.71	26.19
+ Drop-k PAM parallel	9.52	20.75	29.93
Oracle (overlapping terms)	6.12	16.33	28.57
End-of-sentence			
Rewrite	3.89	10.92	16.47
+ Drop-k uniform	4.59	11.75	16.99
+ Drop-k PAM	4.72	11.29	17.05
+ Drop-k uniform parallel	8.45	18.78	25.84
+ Drop-k PAM parallel	9.96	21.40	28.96
Oracle (overlapping terms)	8.70	21.37	28.31

Table 6: Query rewrite results for various Drop-k models.

allocation model (PAM) [21] to model the topics in a text, where keywords forming the dominant topic are assumed as our set of annotation keywords. Compared with previous topic modeling approaches, such as Latent Dirichlet allocation (LDA) or its improved variant Correlated Topic Model [3], PAM captures correlations between all the topic pairs using a directed acyclic graph (DAG). It also supports fine-grained topic modeling, and it has state-of-the-art performance on the tasks of document classification and topical coherence of keyword sets. Given a factual statement, we use the PAM model to infer a list of *supertopics* and *subtopics* together with words weighted according to the likelihood that they belong to each of these topics. Each term is assigned a score given by:

$$Score(t_i) = \sum_{j \in supertopics} p_s^j p_j(t_i) + \sum_{k \in subtopics} p_s^k p_k(t_i)$$

where p_s^j is the inferred probability that topic j is exhibited in factual statement S, while $p_j(t_i)$ is the probability of generating term t_i given the topic j. Additionally, the utility for each semantic term t_i in S is computed as

$$l_{t_i} = \frac{1}{Score(t_i)} \sum_{t_j \in S} Score(t_j)$$

In essence, supertopics provide a coarse-grain topical gist of the statement, while subtopics further enhance the description with fine-grain topical variation. We employ 50 supertopics and 100 subtopics as operating parameters, since these values were found to provide good results in previous work on topic modeling [21]. The model includes several other parameters whose values must be determined empirically on the development set, such as randomization seed, Dirichlet prior for sampling multinomial distribution for subtopics, and number of samples to be drawn from the topic distributions.

While we cannot guarantee which semantic terms for a given factual statement overlap with its supporting Web document, the Drop-k algorithm is intended specifically for increasing recall for this document by including topical words hypothesized to be shared by both texts. As we shall explain later, by virtue of this pool of shared topical keywords, it is also hoped that other candidate supporting Web documents can be retrieved and promoted later in the re-ranking phase. The query results with the Drop-k algorithm implementation are shown in Table 6. As baselines, we include the *uniform* versions of Drop-k, in which the discarded terms are selected uniformly at random rather than based on the induced topic models. We conducted three independent experiments for each Drop-k implementation. We report the average and standard deviations of these. Note that for models that use multiple parallel queries, the results of those queries are aggregated, and the final rank of each unique Web page is based on its highest rank in the retrieved set of results, with ties broken by the average rank.

We seek to answer the following questions: For improved document retrieval, can we modify our queries to better capture the statement's topical inclination by sampling semantic terms with respect to a skewed distribution induced by a topic model such as PAM? Also, is it possible to leverage parallel execution of multiple queries to fetch a more diversified set of documents in order to increase the chance of retrieving supporting Web references?

We note that Drop-k with PAM topic modeling provide consistent improvements over the implementation using exclusively rewrite rules for both in-sentence and end-of-sentence citations. For the parallel implementations, there are significant differences between rewrite rules only and rewrite rules + Drop-k with PAM, the latter providing a performance increase of 5.10% to 9.52% ($p < 0.01$) for Top 1, in-sentence citations, and of 3.89% to 9.96% ($p < 0.01$) for Top 1, end-of-sentence citations. Overall, the increase in referenced documents retrieved in the top 10 and the top 50 for both citation types is also significant ($p < 0.01$) for the parallel implementations of rewrite rules + Drop-k with PAM. Although employing queries constructed by dropping terms uniformly at random results in significant retrieval improvements over the basic rewrite rules too, there is a large standard deviation between their independently executed experiments. Specifically, when evaluated on the non-parallel versions, the average standard deviation of Drop-k (uniform) is 2.03% while that of Drop-k (PAM) is 0% for the In-Sentence citation type, and 0.31% and 0.02% respectively for the End-of-Sentence citation type. We hence conclude that dropping terms uniformly, while increasing the chance of retrieving the correct supporting referenced document overall due to fetching a diversified set of documents, is not as effective as the Drop-k strategy that performs quasi-random dropping of terms based on topically-induced distributional term information.

Surprisingly, when compared to querying the Web using the overlapping semantic terms, the Rewrite + Drop-k PAM-parallel implementation consistently scores better document retrieval performance at all rank levels and for both citation types. Recall that a parallel implementation submits N set of modified queries constructed from a single factual statement and aggregates the resulting sets of retrieved documents, where N corresponds to the number of original semantic terms prior to dropping. To confirm that our parallelized runs are not gaining advantage due to more queries being sent for each statement, we conducted an experiment in which we fetch as many unique Web pages per overlapping query as the total number of unique Web pages retrieved by the Rewrite + Drop-k PAM-parallel. This average *normalized retrieval rate* is 29.01% for Rewrite + Drop-k (PAM-parallel) and 18.15% for using overlapping query retrieval, indicating that sending multiple queries constructed in parallel by randomly dropping semantic terms sampled from a topic model such as PAM results in fact in a more diversified set of retrieved documents, which leads consequently to improved document retrieval performance.

4.3 Re-ranking the Retrieved Results

Until now, we have considered the ranking of Web documents retrieved by Bing search API to suffice for our task on hand without taking into account other evidence. To boost the ranking of those Web documents most similar to the input sentence, we investigate a re-ranking paradigm that employs several similarity measures between the input sentence and the retrieved documents. For each unique Web document d in the combined set of N ranked list of candidates,we compute the best rank of d in any of the lists as

Re-ranking heuristics	Rewrite	+Drop-k PAM	+Drop-k PAM parallel
In-sentence			
+ Cosine	0.00	0.00	+2.04
+ LSA	-2.04	-4.08	-1.02
+ PageRank	**+1.02**	+1.02	+2.04
+ Cosine + LSA	0.00	-3.06	+1.02
+ Cosine + PageRank	**+1.02**	**+2.04**	**+4.08**
+ LSA + PageRank	-1.02	-3.06	0.00
+ all	-1.02	-2.04	+1.02
End-of-sentence			
+ Cosine	+0.46	+0.46	+2.77
+ LSA	+0.19	-0.83	+0.65
+ PageRank	-0.09	-0.09	+0.65
+ Cosine + LSA	+0.37	-0.46	+1.29
+ Cosine + PageRank	**+0.93**	**+0.74**	**+2.96**
+ LSA + PageRank	+0.28	-0.83	+0.65
+ all	+0.37	-0.37	+1.48

Table 7: Performance gain/loss in absolute percentage points for retrieving Wikipedia references at the Top 1 position by adding various heuristics to the re-ranking formula.

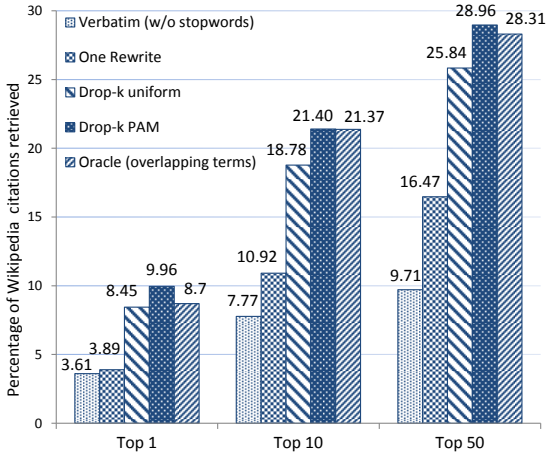

Figure 6: Bar chart summary for the percentage of Wikipedia citations retrieved in Top *n* results for all models proposed.

$minRank(d)$. We then compute the score of the Web document d as:

$$Score(d) = \frac{1}{minRank(d)} \cdot \prod_j h_j(d)$$

and we re-rank all retrieved documents by sorting them in the decreasing order of their scores. $h_j(\cdot)$ are heuristic measures of similarity between Web documents and the input statement. We use the *cosine similarity* metric [23], which quantifies the degree of lexical overlap between the factual statement and the Web reference normalized by their lengths. To measure their similarity on a *semantic level*, we also employ Latent Semantic Analysis (LSA) [19]. LSA quantifies the relatedness between the statement and each of the retrieved Web document using second-order co-occurrence statistics of semantic terms between the two.

Intuitively, by employing both these heuristics simultaneously, Web documents which are semantically- and lexically- similar to the input factual statement are promoted to the top of the ranking, possibly allowing us to discover supporting Web references that qualify as alternatives to the reference provided in Wikipedia.

As a third heuristic, we employ the PageRank [5] algorithm for creating a re-ranked order among references with discrimination for trustworthiness and reliability. PageRank capitalizes on the hyperlink structure of the Web to allow the implementation of a "recommendation system", one in which a Web page having a higher PageRank score possess more incoming recommendations than those with lower PageRank scores. We believe that the higher the PageRank score a Web page possess, the more authoritative and trustworthy it is with respect to its information content, which is somewhat consistent with findings from previous work that demonstrated the effectiveness of using PageRank as a proxy for detecting spam websites [11]. In their work, they concluded that although the Web hyperlink structure is subjected to manipulation by spammers, top-ranked websites by PageRank are mostly legitimate and trustworthy. Given two documents with the same $minRank$, promoting the document with a higher PageRank score also aligns with Wikipedia's policy on selecting reliable and trustworthy sources as citations.

We utilize lightweight implementations of cosine similarity[6], LSA[7] and PageRank[8] to avoid significant overheads in the re-ranking phase. All component heuristics scores are normalized on a 0-1 scale before integrating them into the re-ranking formula.

To find out which combinations of heuristics are most effective in our re-ranking step, we carried out a study in which each heuristic is added individually and in combination with other heuristics. The absolute gains or losses in ranking the Wikipedia referenced documents at the top position by comparison with the basic models are shown in Table 7. Regardless of the combination of heuristics, performance gain or loss are mostly marginal when re-ranking only the set of results for one query (the basic Rewrite method and Drop-k PAM). The benefits of re-ranking become apparent for Drop-k PAM parallel, in which we re-rank a much larger set of Web results corresponding to the N shorter queries employed. The combination of cosine similarity and PageRank provides significant improvements for both end-of-sentence citations (+2.96) and in-sentence citations (+4.08). In general, we note larger variations for in-sentence citations, for which the statements are generally shorter and contain less terms to be used for search, hence may result in less specific queries. Re-ranking their search results by employing measures of similarity with the input statement can therefore have more impact on retrieval performance. Interestingly, of all the heuristics applied individually, LSA turns out to be the worst, decreasing performance significantly in some cases. Unlike cosine similarity, LSA measures texts similarity indirectly and can have the side-effect of promoting documents with little overlap with the input factual statement. Such documents may not be favored by the Wikipedia contributors that follow the Wikipedia guidelines of providing citations to documents that "directly support" the statement. For larger sets of Web results retrieved (e.g., Drop-k PAM parallel), employing the PageRank heuristic individually or in combination with cosine similarity improves the system performance substantially, confirming our hypothesis that it can be used as an effective gauge for selecting trustworthy Web sources that align with those selected by Wikipedia contributors.

Figure 6 provides a summary of the performance of all models investigated prior to the re-ranking phase. The model that employs rewrite rules with a quasi-randomized parallel Drop-k strategy consistently matches or outperforms the Oracle system across all retrieval scenarios. This demonstrates a lot of promise for a system which does not share the Oracle's advantage of being informed which words from the sentence appear in the target Web document.

[6]http://search.cpan.org/~tpederse/Text-Similarity-0.08/lib/Text/Similarity/Overlaps.pm

[7]http://code.google.com/p/semanticvectors

[8]http://search.cpan.org/~ykar/WWW-Google-PageRank-0.17/lib/WWW/Google/PageRank.pm

In each assessment task, you are presented with a subject, such as *Company X*, a statement about the subject, such as "*Company Y was acquired by Company X for $13.5 million in 2007.*", and three Web pages, which may provide evidence about the statement. For each Web page, you are asked to judge:

(A) EVIDENCE QUALITY: Whether you could conclude that the statement is true based on the content of the page.

- **COMPLETE (2):** The text of the Web page supports the statement. In other words, after reading the text of the Web page, a person would conclude that the statement is correct.
- **PARTIAL (1):** The Web page supports the main fact (such as that *Company X* bought *Company Y*), but some details of the statement cannot be inferred from the Web page. For example: the Web page states that *Company X* acquired *Company Y* for *less than $15 million*, but the exact amount in the statement (*$13.8 million*) cannot be concluded from the information provided by the Web page.
- **POOR (0):** The page does not provide evidence about the main fact in the statement (for example, it states that *Company X* made acquisitions in *2007*, but does not mention the acquisition of *Company Y*).

(B) EASE OF EFFORT: How effortless it was to find the supporting information in the Web page?

- **EASY (2):** All of the supporting information for the statement is in one segment of the Web page and it does not take much effort to find (less than one minute)
- **MEDIUM (1):** The supporting information can be found and put together with reasonably little effort (less than five minutes).
- **HARD (0):** A person has to carefully read the content of the entire Web page to make a decision whether or not the Web page supports the statement.

(C) TRUSTWORTHINESS: Is the Web page trustworthy? In other words, do you believe the Web page provides reliable information on the given subject?

- **RELIABLE (2):** You would trust the information provided by this Web site for the given subject.
- **ACCEPTABLE (1):** The Web page is well written and seems to provide credible information.
- **SUSPICIOUS (0):** You would not trust the information from this Web site without double checking.

Figure 7: The instructions provided to the Amazon Turk annotators in the comparative evaluation of the results obtained by the proposed methods and the original Wikipedia reference.

5. HUMAN EVALUATION

The results reported to this end have been based on the important assumption that each factual statement is best supported by the Web reference corresponding to the Wikipedia citation, which we attempted to retrieve via the proposed retrieval algorithms. However, in practice, there are multiple high-quality Web resources that support most statements, and choosing one over another as reference is a subjective undertaking.

We evaluate the overall quality of the Web pages retrieved by the proposed methods, both in terms of how well they support the input statement as well as their status being trustworthy and authoritative sources of evidence. For 100 factual statements randomly selected from the test set, we perform a comparative study of the supporting evidence provided by the original Wikipedia reference (U1), the Web page retrieved at the top position by using the Rewriting system from Section 4.1 (U2), and the Web page obtained as the top ranked result by using the Rewriting system in conjunction with Drop-k PAM parallel (U3). For both U2 and U3, we use additionally a re-ranking process that employs the best combination of heuristics found in Section 4.3, i.e., $Rank(d) = \frac{1}{minRank(d)} \cdot Cosine(d, S) \cdot PageRank(d)$.

We use the Amazon Mechanical Turk (AMT) for these annotations, which has been shown as an effective way to obtain accurate

Reference Type	Quality	Ease of Effort	Trustworthiness
Wikipedia	1.37	1.26	1.48
One Rewrite	1.15	1.18	1.36
Drop-k PAM	1.15	1.24	1.45

Inter-annotator agreement

45-48%
68-77%

44-48%
67-71%

45-51%
79-86%

Reference Type	Quality	Ease of Effort	Trustworthiness
Wikipedia	83.0%	79.3%	90.0%
One Rewrite	71.0%	75.3%	87.0%
Drop-k PAM	71.0%	74.7%	88.7%

Figure 8: (a) The top table shows average scores assigned by AMT annotators in a comparative evaluation of Wikipedia references and the top documents generated by two of the investigated systems. The maximum achievable score is 2. (b) The bottom table displays the percentage of time pages for each method get assigned the best rating after coalescing the top two ratings. (c) The percentage figures above the arrows represent inter-annotator agreement range for assigning the best rating by all three models, before and after coalescing, respectively.

annotations for human intelligence tasks [24] through non-expert contributions. Our decision to engage human subjects in such a study is also justified by the complexity of an accurate assessment of the proposed supporting evidence, which typically involves deep understanding of the texts and requires paraphrasing and entailment skills achievable only by humans at this point. To increase the validity of our findings, we only employ experienced Turkers with 97% approval ratings and who have successfully completed 50 tasks approved by requesters.

For each data point comprising a subject, a factual statement, and the corresponding three Web pages (U1, U2, and U3), we invited three independent annotators to provide judgements on three criteria: (a) evidence quality; (b) ease of effort; and (c) trustworthiness of the Web page/site, as defined in the instructions shown in Figure 7. The purpose of this setup is to allow each annotator to discriminatively evaluate the effectiveness of the three Web pages in providing support for a subject and a factual statement about the subject from our dataset. To ensure a fair evaluation, we randomize the order of the Web pages for each data point presented to each annotator. We set no time limit for each annotation task; annotators could take as long as they need to scrutinize the Web page for any information that helps them to provide an answer for each evaluation criterion. However, such information is restricted exclusively to the Web pages shown.

For each type of Web reference and each evaluation criterion, we compute an average score per annotator per data point using the range 2-1-0. The results are shown in Figure 8. The original Wikipedia reference (U1) has the highest averages for all three evaluation criteria, which is consistent with our initial assumption that the references provided by the Wikipedia contributors represent some of the best supporting sources of evidence, and could be regarded as an upper bound. For evidence quality, U2, which is based exclusively on searching the Web using our baseline rewriting system, ties with U3, the model enhanced with Drop-k (PAM Parallel). Both score an average of 1.15, which is lower than the score for U1 (1.32). Because U1 has a standard deviation of 0.76 for evidence quality, 0.78 for ease of effort, and 0.67 for trustworthiness, and also, the sample size is only 100, the averages for our

		(A) Evidence quality		(B) Ease of effort		(C) Trustworthiness	
		Best rating(%)	Inter-agreement(%)	Best rating(%)	Inter-agreement(%)	Best rating(%)	Inter-agreement(%)
U1:	Three-value range (**2** vs. 1 vs. 0)	54.00	45.67	47.33	47.67	57.67	50.33
Wikipedia	Coalesced (**2** vs. 1 ∪ 0)	54.00	54.67	47.33	58.00	57.67	58.00
	Coalesced (**2** ∪ **1** vs. 0)	83.00	76.67	79.33	70.67	90.00	86.00
U2:	Three-value range (**2** vs. 1 vs. 0)	44.00	45.67	42.67	44.00	48.67	45.67
One rewrite	Coalesced (**2** vs. 1 ∪ 0)	44.00	59.34	42.67	58.00	48.67	58.00
	Coalesced (**2** ∪ **1** vs. 0)	71.00	71.33	75.33	68.00	87.00	79.33
U3:	Three-value range (**2** vs. 1 vs. 0)	44.33	47.33	49.00	43.67	56.00	45.00
Drop-k PAM parallel	Coalesced (**2** vs. 1 ∪ 0)	44.33	60.67	49.00	57.33	56.00	52.67
	Coalesced (**2** ∪ **1** vs. 0)	71.00	68.00	74.67	66.67	88.67	82.00

Table 8: Percentage of time pages generated by the discussed methods were assigned the best rating, and annotator inter-agreement.

rewrite model (U2) and the Drop-K model (U3) situate well within margins of the spread, and could be seen as similar to the averages for U1. Additionally, U3 is indistinguishable from U1 for ease of effort (1.24 vs 1.26) and trustworthiness (1.45 vs 1.48).

To our knowledge, there is currently no way to ensure that AMT annotations are performed by a fixed set of annotators, each of whom completes the entire set of tasks. AMT promotes a "free market" of annotation transactions, in which each annotator decides whether to work on a task based on goodwill and monetary compensation offered for the task. As such, we are unable to obtain agreement statistics using standard measures such as Fleiss's Kappa. Instead, we measure the inter-annotator agreement by considering pairwise comparisons of the three annotations and dividing the number of matching annotations by the total number of annotations. Table 8 shows detailed evaluation results as percentages of time annotators assigned the maximum score to each type of document, as well as the inter-annotator agreement ratios. Overall, the annotators assign the best score for evidence quality to the Wikipedia reference a higher number of times than to the Web pages generated by our system. At the same time, we note that agreement ratios are relatively low, even when judging Wikipedia references. Because most disagreements are due to the assignment of scores 2 and 1 by different annotators, we hypothesize that the interpretation of the top two labels as defined in our guidelines leads to the highest inconsistency across annotators, due to the inherently subjective nature of such a labeling exercise. Therefore, Table 8 presents also coalescing statistics, for merged adjacent scoring labels (2 ∪ 1 vs 0, as well as 2 vs. 1 ∪ 0). In all of the scenarios, the inter-agreement ratios improve dramatically when merging the top two labels (2 ∪ 1 vs 0), which indicates that annotators tend to disagree about the strength of the evidence presented rather than whether or not a document supports a given statement. Overall, the obtained numbers indicate that the proposed system is capable of retrieving supporting Web documents comparable to those provided by the Wikipedia contributors.

6. CONCLUSIONS AND FUTURE WORK

We presented an investigation of the task of automatically retrieving supporting evidence from the Web for factual statements, in which we employed a large corpus of factual statements paired with supporting Web documents, as derived from the Wikipedia collection. Based on this corpus, we proposed a training paradigm for rewriting factual statements for querying the Web, which is competitive in performance with an oracle system that matches lexically the input statement with the paired Wikipedia reference. Furthermore, a user study based on Amazon Mechanical Turk showed that the supporting documents retrieved by the proposed system are comparable in quality to those selected by Wikipedia contributors.

We envision that the studied task can be further expanded to cover a more comprehensive scenario of fact verification, in which both supporting and contradicting evidence must be retrieved and organized/summarized for any input factual statement.

7. REFERENCES

[1] M. Bendersky and B. Croft. Discovering key concepts in verbose queries. In *Proc. of SIGIR*, pages 491–498, 2008.

[2] M. Bendersky and B. Croft. Analysis of long queries in a large scale search log. In *Proc. of WSCD*, pages 8–14, 2009.

[3] D. Blei and J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[4] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. In *Proc. of TREC*, pages 393–400, 2001.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117, 1998.

[6] C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. *Technical Report*, MSR-TR-2010-82, 2010.

[7] A. Csomai and R. Mihalcea. Linguistically motivated features for enhanced back-of-the-book indexing. In *Proc. of ACL*, 2008.

[8] S. Cucerzan. TAC entity linking by performing full-document entity extraction and disambiguation. In *Proc. of TAC*, 2011.

[9] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *Proc. PASCAL Workshop*, 2005.

[10] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. M. Popescu, T. Shaked, S. S. abd D. Weld, and A. Yates. Web-scale information extraction in KnowItAll. In *Proc. of WWW*, pages 100–110, 2004.

[11] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. of VLDB*, pages 576–587, 2004.

[12] J. Huang and E. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proc. of CIKM*, pages 77–86, 2009.

[13] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proc. of WWW*, pages 387–396, 2006.

[14] M. Kouleykov and B. Magnini. Tree edit distance for textual entailment. In *Proc. of RANLP*, 2005.

[15] Z. Kozareva and A. Montoyo. MLEnt: The machine learning entailment system of the University of Alicante. In *Proc. of PASCAL RTE Workshop*, 2006.

[16] G. Kumaran and J. Allan. A case for shorter queries, and helping user create them. In *Proc. of HLT*, pages 220–227, 2006.

[17] G. Kumaran and J. Allan. Effective and efficient user interaction for long queries. In *Proc. of SIGIR*, pages 11–18, 2008.

[18] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proc. of SIGIR*, pages 564–571, 2009.

[19] T. Landauer and S. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition. *Psychological Review*, 104(2):211–240, 1997.

[20] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[21] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proc. of ICML*, 2006.

[22] R. Mihalcea and C. Leong. Towards communicating simple sentences using pictorial representations. *Machine Translation*, 22(3):153–173, 2009.

[23] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::Similarity - measuring the relatedness of concepts. In *Proc. of AAAI*, 2004.

[24] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast - is it good? Evaluating non-expert annotations. In *Proc. of EMNLP*, 2008.

[25] M. Tatu and D. Moldovan. COGEX at RTE 3. In *Proc. of PASCAL RTE Workshop*, 2007.

[26] X. Xue, S. Huston, and B. Croft. Improving verbose queries using subset distribution. In *Proc. of CIKM*, pages 1509–1068, 2010.