Identification of Live News Events using Twitter *

Alan Jackoway
University of Maryland
jackoway@umd.edu

Hanan Samet
University of Maryland
hjs@cs.umd.edu

Jagan Sankaranarayanan
University of Maryland
jagan@cs.umd.edu

ABSTRACT

Twitter presents a source of information that cannot easily be obtained anywhere else. However, though many posts on Twitter reveal up-to-the-minute information about events in the world or interesting sentiments, far more posts are of no interest to the general audience. A method to determine which Twitter users are posting reliable information and which posts are interesting is presented. Using this information a search through a large, online news corpus is conducted to discover future events before they occur along with information about the location of the event. These events can be identified with a high degree of accuracy by verifying that an event found in one news article is found in other similar news articles, since any event interesting to a general audience will likely have more than one news story written about it. Twitter posts near the time of the event can then be identified as interesting if they match the event in terms of keywords or location. This method enables the discovery of interesting posts about current and future events and helps in the identification of reliable users.

Categories and Subject Descriptors

H.3 [**Research Paper**]: Systems and Services

General Terms

Algorithms

Keywords

tense identification, future tense, event detection, geotagging, Twitter

1. INTRODUCTION

Though Twitter presents a massive source of information on current events, it is an incredibly noisy medium, so automatically selecting which posts (i.e., Tweets) are reliable and interesting for a general audience can be very difficult. Many users post information that is only interesting to their individual group of followers, post spam, or post incorrect information. Furthermore, some users could be very reliable on one topic while being completely unreliable about another topic. Our goal is to help identify which users post reliable Tweets and which Tweets are interesting.

We accomplish this by identifying and predicting future events as well as where they will be happening. Events are discovered using a constantly updating corpus of news articles, called NewsStand and described fully in [28]. After news articles have been downloaded and grouped into clusters based on their story, events can be discovered as described in Section 5. Using information about the events, we extract relevant and reliable information from posts on Twitter. This is aided by making use of a large and constantly updated corpus of news articles. The prediction capability is achieved by using the future events to define new features (such as keywords and location) related to the events which can be subsequently extracted so that Twitter traffic (i.e., postings) about the events can be identified. Our work is also useful in identifying which Twitter postings are associated with real time events. Moreover, it aids us in identifying Twitter posters who are posting on real events and, of equal importance, which posters can be deemed reliable on a particular issue. We can also begin to determine a geographic region with which a user is familiar by finding out where the events that a user posts about are occurring. The more reliable posts about events in a certain area, the more likely future posts about that area are to be reliable. This is all in the spirit that there are millions of people posting on Twitter, and as we cannot follow all of them, we would like to have some measure of their reliability and credibility. This work is a temporal extension of our prior work [23] which attempted to determine the spatial locations associated with posts on Twitter despite the absence of location information in the Tweets.

The rest of this paper is organized as follows. Section 2 provides a brief introduction to Twitter, which is expanded with regard to news in Section 3, while Section 4 reviews the geotagging process. Section 5 describes our tense identification methods. Section 6 contains results of an experimental evaluation of these methods, while concluding remarks are drawn in Section 7.

2. TWITTER OVERVIEW

Twitter is a social networking website that has expanded greatly over the last few years. Twitter users post messages (termed *Tweets*), which can be seen by all users who choose to "follow" them. One user can become another user's "follower," in which case everything posted by the second user will be viewable by the first user (though there is a way to block users from reading posts if desired). Each Tweet is at most 140 characters, allowing Tweets to be posted via text message as well as using the Twitter website. As of April 2010, Twitter had 105 million registered users, with 300,000

*This work was supported in part by the National Science Foundation under Grants IIS-10-18475, IIS-09-48548, IIS-08-12377, CCF-08-30618, and IIS-07-13501.

LBSN '11, November 1, 2011. Chicago, IL, USA

more registering each day. These users were posting about 55 million Tweets per day [30]. Twitter's enormous popularity could be due to its simplicity—by limiting users to 140 character-long messages, it means that posting or reading a Tweet is guaranteed to not take very long. Additionally, Twitter is available on a number of platforms, including the website, SMS messaging, and a multitude of clients for both phones and computers. As a social network, Twitter's popularity increases because it encourages users to follow many people and get as many people as possible to follow and read their Tweets.

Because of the enormous number of Tweets posted per day on a huge variety of topics, Twitter can be an ideal source for finding out up-to-date information on events. However, there are a few major obstacles to extracting this information from Twitter. First, because of the character limit on posts, Tweets do not have as much information as a post on a blog or website, which can make it difficult to assess their value. Posts with links (which ordinarily would provide quite a bit of information because of the hyperlink) are made more difficult to analyze because they are almost always "shortened" by passing them through a forwarding service specifically designed to map real URLs to 10-20 character links. Another confounding factor in finding useful information on Twitter is the prevalence of spam messages. Many users posting spam attempt to make their messages look as legitimate and interesting as possible to catch unsuspecting eyes. Twitter has taken major steps to reduce spam with some success. In February 2009, spam messages accounted for about 5.5% of Tweets. In the summer of 2009, spam neared 10%, but as of February 2010, spam has been reduced to just over 1% of Twitter's posts [4].

The most pressing problem with finding interesting and reliable Tweets is the volume of Tweets that are not interesting to a general audience. Twitter does not have a specific topic or goal like some websites, but rather it encourages people to post about anything they like. As a result, many Tweets are brief thoughts of a user, often interesting only to people who know that user personally. Though this is not a problem for Twitter—as long as the people following a user are interested in what he or she tweets, Nevertheless, Twitter remains useful as a social network, although it can be quite troubling for those trying to extract Tweets that would be interesting to an audience other than a users' followers. This paper provides a way to identify which users are reliable by using the fact that we can verify their posting on particular events on account of having discovered the actual events independently by looking at news articles from NewsStand [28], a constantly updated corpus of online news, similar to Google News. By verifying users based on their comments on events that are known to be happening in the world, we can determine what users are knowledgeable about certain events or geographic regions.

3. BREAKING NEWS USING TWITTER

As we pointed out, Twitter is an electronic medium that allows a large user populace to communicate with each other simultaneously. We recently demonstrated a system called TwitterStand [23] that uses Tweets posted by users in Twitter to capture breaking news faster than conventional news aggregators (e.g., Google News, Yahoo! news). In fact, a constant criticism of conventional news aggregators is that they are slow in responding to breaking news [6]. When important news occurs, it takes quite a while for news aggregators to prominently display it. This shortcoming of news aggregators is not surprising as they themselves do not produce the content, but rather simply crawl and index news sources that produce them. The news sources first have to produce the news content, which usually passes through an editorial pipeline after which they are crawled and indexed by the news aggregators. This process takes time, which means that there is an unavoidable time lag between when the event occurs and when the news aggregators can display them. Also, most news aggregators only show a few important stories, where importance is usually assigned based on how many different news sources contribute to (i.e., report on) a certain *news topic*, which is obtained by *clustering* news articles so that similar articles from different sources are associated with the same news topic. This means that the news topic should contain enough news articles before it is prominently displayed, which further adds to this perceived time lag with conventional news aggregators.

A news aggregator using Tweets (e.g., TwitterStand) attempts to capture late breaking news entirely using Tweets written by the users of Twitter. The result is analogous to a distributed news wire service, where the identities of the contributors/reporters (i.e., analogous to news sources) are not known in advance and there may be many of them. Tweets are not sent according to a schedule and there are no reporters being assigned to cover stories. The challenge becomes how to separate the news Tweets from non-news Tweets, which is a hard problem. What makes Twitter attractive for capturing breaking news is that there is very little lag between the time that an event happens, or is first reported in the news media, and the time at which it is the subject of a posting on Twitter. The data is "pushed" by the content providers (i.e., people who send Tweets) and is delivered nearly instantaneously to the content consumers (i.e., people who receive Tweets and TwitterStand). In contrast, conventional news aggregators must constantly poll the content providers for updates with web spiders, which means that there could be a significant time lag between the time the news is published and is first picked up by the news aggregators. Thus we see that from the point of view of speed (i.e., the ability to generate a scoop), Twitter provides news aggregators such as TwitterStand an edge over conventional news aggregators such as Google News.

However, there are still issues with news aggregators that rely on Tweets for news gathering. In particular, such aggregators must deal with the inherent unreliability of the information carried by Tweets. As Tweets undergo no editorial control there is very little one can do to assure reliability of the information that they carry. Moreover, there is very little in the way of holding users accountable for publishing misinformation. In fact, Twitter users often indulge in misinformation campaigns. For example, consider the rumor campaign on Twitter reporting on the death of the actor Johnny Depp in a car accident [15] in January 2010. This means that what we perceive as news Tweets in Twitter could very well be part of a deliberate rumor campaign. So, there is a need to examine ways to instill trust and reliability in the news gathered by the way of Tweets, which is the goal of this paper. Please note that this issue of trust and reliability does not still diminish the utility of systems like TwitterStand. Even with these problems, the basic intent of the majority of Twitter users is not to indulge in misinformation campaigns, but one always has to be wary of the unreliability of this medium.

In general, assuring trust and reliability of news obtained from Twitter is a hard problem. A simple idea that we explore in this paper is to combine the reliability of conventional news media with the swiftness of Twitter for certain kinds of breaking news. In particular, we are interested in exploring a synergy between conventional news sources and Twitter when it comes to events that are prescheduled. In other words, we are not interested in unexpected events (e.g., Earthquake in Haiti), but in events that we already know are going to take place in the future, in a certain time window. In this regard, our ability to recognize future events in con-

ventional news (as described in section 5) comes in handy. Given that we know that a certain news event is going to occur at a certain time, can we now use Tweets posted in Twitter to provide live updates as the event is progressing. For example, considering that we know from processing conventional news that Superbowl 2011 will happen on Feb 6, 2011, can we use Twitter to post live updates as the game is progressing? The proposed system combines the reliability of conventional news media with the speed as Twitter. The assurance that the Superbowl will happen on Feb 6, 2011 is obtained from conventional news sources, while the swiftness of Tweets is used to provide real time updates as the game is taking place.

The setup of our system is as follows. We constantly process conventional news sources so that we can pick out any news topic n containing mentions of a future event. This is done under the auspices of our earlier system called NewsStand [28], which is a system that brings news reading experience to a map interface. Note that the extraction of future events from news articles is described in Section 5. A news topic n containing mention of a future event is denoted by its feature vector $\boldsymbol{TFV_n}$, which is the set of words or phrases that can be used to describe the news topic. The feature vectors are obtained using the TF-IDF [20] measure. Furthermore, each news topic n is associated with a time window $[t_s, t_e]$, which is the time period during which an event related to n will occur. Finally, the geographic region n_g serving as the focus of n is obtained through geotagging [11, 28] of the news articles associated with n. Furthermore, assuming that the current time is t, let N denote a set of news topics such that the time window associated with the events in N contains t (a "time window," defined by a start and end time, is just a period of time). In other words, N denotes the set of news topics that are happening at moment t. We refer to N as the set of active news topics, which is constantly updated as news topics are added and removed from it. Let F be the set of feature vector terms obtained by pooling the feature vectors of all the news topics in N. Now, the set of keywords F forms the input of the *track* API of Twitter, which takes up to 200,000 keywords as input, and obtains Tweets in Twitter containing one or more of the keywords in F.

Our input is a stream of Tweets from Twitter track API containing one or more keywords in common with one or more news topics in N. As most of the Tweets obtained using this method from Twitter are not related to news, we first apply a coarse filtering on incoming Tweets, which classifies incoming tweets as either *junk* or *news*, where junk tweets have a good chance of not being related to the news and hence, are discarded, while the news tweets have a good chance of being related to news. Note that our goal is not to completely get rid of noise, which may not even be possible given the uncertain boundary between news and noise, but instead to find a way of discarding tweets that clearly cannot be news. So, the goal here is to throw away as many tweets as possible without losing many news tweets. Note that we are not really worried about misclassifying a small percentage of *news* Tweets as junk and not even processing them. This is due to the incredibly high amounts of Tweet data that is available to us from Twitter. However, our aim is to ensure that we provide a very good quality output, one that does not contain too many noise Tweets. For the purpose of separating junk Tweets from news Tweets, we use a naive Bayes classifier [16] that is trained on a training corpus of tweets that have already been marked as either news or junk. Interested readers are referred to [23] for a brief review of the classifier used to separate out news Tweets. At this stage, the Tweets, which are still noisy, have a higher percentage of news Tweets, which is good enough for our purposes.

We now have to associate an input Tweet t from Twitter with a news topic n in N, or possibly discard it. For this purpose, we maintain an active set N of news topics, such that each news topic n in N is denoted by its *feature vector* $\boldsymbol{TFV_n}$. When an input news Tweet t is obtained, we first represent t by its feature vector representation $\boldsymbol{TFV_t}$ using the TF-IDF measure. We then use a variant of the *cosine similarity measure* [26] to compute the distance between t and a candidate news topic n in N, which is defined as follows:

$$\delta(t, n) = \frac{\boldsymbol{TFV_t} \bullet \boldsymbol{TFV_n}}{||\boldsymbol{TFV_t}||\,||\boldsymbol{TFV_n}||}$$

where $\boldsymbol{TFV_t}$, $\boldsymbol{TFV_n}$ are the feature vectors of t and n, respectively. t is considered a news update of the closest news topic n in N as long as $\delta(t, n) \leq \epsilon$, where ϵ is a pre-specified constant. In addition to a distance based constraint, we also stipulated that there be at least γ features between t and n, where γ is a small constant. If no such news topic exists in N, then we simply discard the Tweet t and proceed to the next Tweet in the input stream.

To expedite the search for a news topic n in N that is nearest to t as well as within a distance of ϵ from t, we maintain an inverted index on the feature vectors of the news topics in N. That is, for each feature f in $\boldsymbol{TFV_n}$ of a news topic n in N, the index stores pointers to all news topics in N that contain f. We use this index to reduce the number of distance computations required for associating a Tweet t with a news topic in N. When a new Tweet t is encountered, we only compute the distances to those news topics in N that have at least one feature in common with t. This optimization enables our algorithm to minimize the number of distance computations necessary for associating a Tweet with a news topic. Tweets that are updates to a news topic in N are directly posted to the user interface. The result is that users can see live updates of ongoing events, which is now possible due to our understanding of future event references in conventional news.

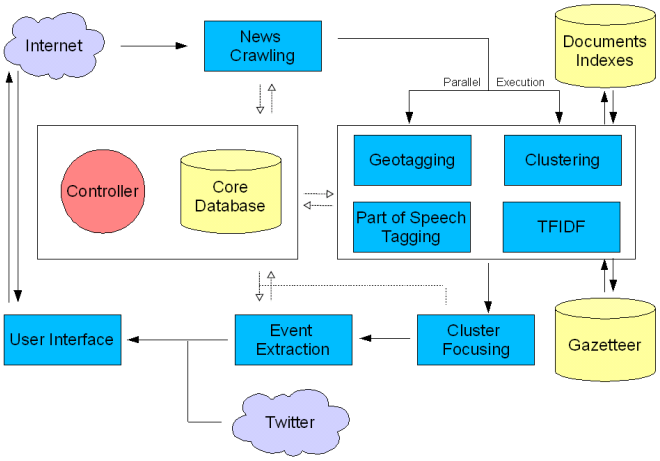

Figure 1: A high level overview diagram of NewsStand [28] with future events added to the system. News articles move through the system, which is orchestrated by the controller.

4. REVIEW OF GEOTAGGING

Geotagging [1] is the process of identifying locations in text and assigning them latitude/longitude coordinate values. Once text has been assigned coordinate values in this way, common spatial queries can be applied to it, including

both feature-based and location-based queries. This allows a search to find both where something referenced in the text is happening and what texts discuss a certain place. Geotagging systems have been constructed for a variety of domains such as blogs [31], encyclopedia articles [7], news articles [5, 28], Twitter messages [23], and more.

The textual references to geographic locations identified by geotagging are known as *toponyms* [8]. Since most toponyms are somewhat ambiguous (e.g., over 60 places around the world are named "Paris"), simple textual matching is not sufficient to identify the latitude/longitude of a textual entity. Moreover, there is also the issue of toponym recognition (e.g., "Jordan" can refer to a person as well as a county). The system we used for geotagging news articles combined of data in the geonames database [29] with the location of the publisher of the news article while inferring a reader's local lexicon [10, 11, 17]. There are many other geotagging systems as well [1, 8, 12, 18, 28].

The geotagging used in the system we describe is fully explained in [9, 11], and results in geotagging locations in news stories with a good degree of accuracy (precision of at least 0.800). These locations are used to determine the areas about which a Twitter user is knowledgeable, so the accuracy of the geotagger is important to ensuring the accuracy of our determinations about Twitter users.

5. IDENTIFICATION OF FUTURE EVENTS

The ability to identify future events using a large news corpus is the crux of our system for finding reliable Twitter posts. Future events, along with their keywords and locations, are used to find Twitter posts that could be interesting to a general audience and would certainly be interesting to those who are following the event in question. Once events are identified, we can find Twitter posts around the time of the event and, based on information gleaned from a news database, determine which posts are most likely to be reliable. This helps to identify which users on Twitter can be trusted to post about certain geographical regions or types of events. The process of identifying dates and times in news stories has also been attempted by both Mani [14] and Schilder [24]. Both used quite different approaches from ours and achieved similar results. Mani's work used a few starting rules (similar to the tense identifiers described in Section 5.2 and Section 5.3), and then used machine learning techniques to improve tense identification. Schilder, who worked with German text, determined "semantic attributes" related to the date expressions found in the text. These attributes included the text itself (e.g. "on Monday"), the published date and time of the article, and words like "last," "next," and so on. By contrast, our methods for identifying tense are focused on using the verbs surrounding the date to identify tense.

By adding the ability to identify future events and integrating this service with Twitter, we can extend our previous works NewsStand [22, 28] and TwitterStand [23] by adding a temporal component to their spatial displays. The resulting system is diagrammed in Figure 1 and a screenshot of the resulting application is presented in Figure 2. A longer description of the resulting application is presented in Section 7.

The identification of future events proceeds as follows. First, news articles are collected from many news sources across the Internet. Their text is extracted, cleaned, and tagged by geotagging, named entity recognition, and part of speech tagging. Articles are clustered at this time by a process described in [28]. The resulting clusters of articles are all about the same story, so a "cluster" is just a collection of articles about the same news story. Keywords are determined for each article using a standard TFIDF analysis. Next, the text is scanned for any word that could possibly indicate a date (as in [14]). Each of these potential dates must be identified as future or non-future. To achieve this, a tense detector processes each sentence with a potential date and marks the date as future or non-future. This yields a list of future (and past) dates for the article. These dates, coupled with the story around them, are termed *events* for an article.

Since the process by which future dates are determined is imperfect, the process of clustering the articles is used to improve accuracy. In particular, once a cluster's size exceeds a certain threshold value (determined by the accuracy of tense identification and the desired accuracy of the resulting events), the cluster is scanned to find dates that are reported in a suitable number of the cluster's articles (this number must also be determined by desired accuracy of the result). Clustering assists the identification process because it helps to eliminate incorrectly identified events. Since events are only reported if they appear in a certain percentage of the cluster's articles, it is very unlikely that the tense of an event can be reported incorrectly. Nevertheless, for this to happen, the event would have to be erroneously identified in a large number of articles, which is very unlikely given the accuracy of the tense identifiers described later in this section. If a future event appears in enough articles in a cluster, then the system reports that a future event for the cluster is happening, and information such as keywords (identified by TFIDF) and location (identified by geotagging) can be used to determine the nature of the event.

Once a future event has been identified by finding it in articles, it can be used to find Twitter posts about the event. In particular, associated with each article is a list of keywords that are determined using a TFIDF analysis. These keywords are then compared between articles in the cluster so that if a keyword occurs in enough articles, then it is designated as a keyword for the cluster itself. Now, when one of the future dates for the cluster approaches, Twitter traffic that contains keywords associated with the cluster is likely to be about the event. This is effective because the clustering identifies keywords that pertain to the article, and future event identification is used to indicate when people are most likely to be discussing an article (really a cluster topic) on Twitter. Additionally, each article is run through a geotagging process, giving locations at which the event may occur or locations that are involved in the event (such as a sports event between two cities). Using Twitter's information about the location of its posters (or inferring it as in [23]) can also help identify traffic about an event that is legitimate, as users closer the the event geographically may be more likely to discuss it.

5.1 Tense Identification

To determine whether a date in a sentence corresponds to a future date is not always a simple task, and to date, little research has been done on the topic (though part-of-speech taggers have greatly improved, determining the tense of verbs well and some work has been done on automatically learning tenses of verbs [3, 13, 19], but these efforts have only been applied to past tense constructions). Some dates in text, like May 19, 2010 or 2010-05-19 or even "next Friday" can be identified unambiguously, many dates cannot be uniquely determined. For example, consider the difference between the sentences, "He appears in court Tuesday" and, "He appears to have fled on Tuesday." The first occurrence of "Tuesday" is clearly a future occurrence, while the second occurrence is clearly a past occurrence. In order to distinguish between examples such as these, we devised four different tense identification methods, termed *tense identifiers*. The first two methods are very simplistic and are

presented here only to be used as a baseline to evaluate the last two methods, which are more complex and more effective. It should be noted that the last two methods require part of speech tagging to be applied to the news articles, while the first two methods do not. In systems handling huge numbers of articles at a high speed, the part of speech tagging requirement may be an obstacle, but (as reported in Section 6) the last two methods performed so much better than the first two that it is probably worth applying part of speech tagging to at least the sentences that contain date words for most systems.

5.2 Naive with "will" Method

The first method, termed the *naive with "will" method*, starts by searching for the word "will", the most common word marker of the future tense in the English language, and tags all of its occurrences as "future." Next, sentences containing a pronoun (identified by a list) or a proper noun (which can be identified by capitalization or TFIDF) followed immediately by a verb in the past tense, are tagged as "past." Past tense verbs were determined with the aid of a short list of common irregular past tense verbs (had, was, were, etc.) and any word ending in "-ed." Otherwise, the sentence was tagged as "not future" (but not necessarily past).

When this method determined that sentences were in the future tense, it was generally correct, as the word "will" is a fairly reliable marker of tense. However, in news articles, the future is often indicated by a present tense construction, such as the sentence "She hopes to resolve the matter on her July 6 court date." The fact that the "present" is constantly changing and "becoming the past", makes it nearly impossible to write a news story about the present, and thus it is usually the case that present tense verbs are used to indicate a future event in news articles.

5.3 Naive with "-s" Method

To address the issue of misinterpretation of "present tense" constructions, by being unable to classify a sentence like "She hopes to resolve the matter on her July 6 court date" as future, the naive method was modified to create a second method termed the *naive with "-s" method*. It functions the same way as the naive with "will" method with an additional check at the end for the occurrence of a a pronoun or proper noun being immediately followed by a word ending in "-s," in which case the sentence and all dates occurring in it are identified as "future." This was found to be effective as news stories are often written in third person, and the third person singular ending "-s" for present tense verbs can therefore be used to indicate a future event. As described in the Experimental Results section below, the naive with "-s" method had a substantially better recall value but a far worse precision value in comparison to the naive with "will" method. The reason for the dramatic fall in the precision value is the fact that most plural nouns (and many other words) also end in "-s," so sentences that contained a proper noun or a pronoun followed by a plural noun were mistakenly identified as "future." However, the advance in recall was significant enough that this method is worth considering for systems that cannot apply part of speech tagging.

It should be clear that both this method and the naive with "will" method (referred collectively as the *naive methods*) are inadequate for our purpose. First, neither method attempts to definitively identify the parts of speech of the words in the sentence. Aside from a small list of verbs, these naive methods attempt to infer which words are verbs and their tense by looking for pronouns and proper nouns followed by words with verb-like endings. While they work well for basic constructions such as "he was there Thursday," "she performs Wednesday," "they scored seven runs last Tuesday," etc., they are ineffective for more complicated constructions where the verb is split from the subject such as in the sentence "Jed York, president and CEO of the San Francisco 49ers, announced plans for the new stadium on Wednesday." Another problem with the two naive methods is that in news articles, future events are often described using infinitives (e.g. "The report, expected to arrive on Tuesday..."). Finally, some sentences have multiple cases, as in "He said Thursday that he will be playing in Sunday's game" In this example, the date word "Thursday" is in the past tense, while the date word "Sunday" refers to the future.

5.4 Verbs Method

The third tense identification method, termed the *verbs method*, addressed the major issues with the naive methods. In particular, in this method, each article is processed with a part of speech tagger, which identifies both verbs and their cases. The result is that all verbs are tagged, eliminating the need to infer which words were verbs. Additionally, sentences were analyzed at a phrase level instead of at sentence level, so that sentences with multiple tenses can be analyzed correctly.

The algorithm works as follows. Once the part of speech tagging process is done, all date words in the article are identified. Next, the tense of each date word is determined by analyzing the sentence in which it occurs. Each such sentence is decomposed into phrases (using punctuation), so that only the part of the sentence nearest the date word is used in the date word's identification. Initially, all verbs in the phrase containing the date word (including helping verbs) are assigned a score. The scores were determined heuristically and seem to work well in the systems that will be described. Both the verbs method of tense identification and the next tense identifier used the same scores.

High positive scores (+10) are assigned to verbs tagged as being in the future tense, which also included the modal verb "will". The high score represents the frequency with which these verbs are found in future constructions. In the sentences that were identified manually, every sentence containing the word "will" was at least partially in the future tense (some had a clause in future tense and a clause in past).

Verbs tagged as being in the past tense are assigned a score of -2. Past tense verbs are a good, but not absolute (as is the case of "will"), indicator of a past construction. Most sentences with past tense verbs in the sample set were past tense, but past tense verbs also occurred in appositive phrases, such as "The team, which won last week, plays tomorrow" in which the past tense "won" does not make the sentence past tense. Constructions like "it is expected to arrive Friday" also include past tense verbs despite being in future tense. However, in our corpus, past tense verbs were highly correlated with past constructions, so a score of -2 is assigned.

Verbs tagged as being in the present tense are assigned a score of +1. The rationale for this assignment in terms of the weakness of the evidence of the future action (i.e., assigning a -2 score for past events vis-a-vis assigning +1 to present events) was based on our experimental observation that the correlation of past tense verbs with past sentences was much higher than the correlation of present tense verbs were with future events. That is, in our corpus, present tense verbs were more correlated with future constructions than they were with past, but that correlation was not as strong as the correlation between past tense verbs and past constructions.

Most modal and helping verbs (such as "can," "be," and "may" but NOT "will") are assigned a score of 0. In our corpus, there was no clear correlation between modal verbs

Table 1: Comparative Results of Tense Identification

Identifier	Naive with "will"	Naive with "-s"	Verbs	Phrases
Correct Future	33	63	65	78
Number Future	97	97	97	97
Correct Past	240	189	240	237
Number Past	253	253	253	253
Incorrect	77 (22.0%)	98 (28.0%)	45 (12.9%)	35 (9.7%)
Precision (Future)	70.2%	49.2%	86.7%	84.8%
Recall (Future)	34.0%	49.2%	67.0%	80.4%
Precision (Past)	78.9%	84.8%	87.0%	91.5%
Recall (Past)	94.8%	74.7%	94.8%	93.7%

and tense.

Infinitive verbs, like present tense verbs are given a score of +1 since they often indicate future events, especially in news sources. They can be in all three tenses and are best determined by the verbs around them. For example, consider the three sentences, due to [27], which discusses the tense of infinitives at length:

- "I expect John to win the race" (future, determined by the present tense verb "expect"),
- "The president is believed to be guilty" (present, determined by present "is" with past participle "believed"),
- "I remember John to be the smartest" (past, determined by the present tense verb "remember").

The tense identifier accumulates all the scores for the verbs in the phrase containing the date word. If the phrase is assigned a non-zero score, then the score and its tense interpretation are returned and the process terminates. Otherwise (i.e., the phrase containing the date word has a score of 0), adjacent phrases are processed and assigned scores to be added to the score for the original phrase. Once a non-zero result is found or the method runs out of phrases, then the score is returned. A more formal description of the algorithm is given below, where scoreVerbs is a method that adds up the scores of all verbs in a given phrase.

Algorithm 1 The verbs method of tense identification

```
Break the sentence s into phrases p_i.
Let d be such that the date word is in phrase p_d.
Let score ← 0
Let window ← 0
while score = 0 do
  for i ← d − window to d + window do
    score ← score + scoreVerbs(p_i)
  end for
  window ← window + 1
end while
Return score
```

As an example, consider two applications of this method to the sentence "He will speak Tuesday night about the storm, which destroyed many houses in the Midwest on Friday." When the method is applied to the word "Tuesday," the first phrase is analyzed for a total score of 11 (10 for "will" and 1 for "speak"). This score is non-zero, so "Tuesday" is declared to be in the future tense. When the method is applied to "Friday," the first phrase analyzed only has the past tense "destroyed," so a score of -2 is returned and "Friday" is declared to be in the past tense.

Observe that the verbs tense identification method makes two major improvements over both of the naive methods. First, by examining phrases instead of sentences, it eliminates the main source of erroneously-identified future events in the two naive methods, which is caused by sentences in which a date word appears in one part in one tense, while another part of the sentence is in a different tense. For example, in the above sentence involving the "storm," the naive with "will" method identifies "Friday" as a future word because of the presence of the word "will," but the verbs method correctly identifies the more proximate use of "destroyed" as determining the tense. Second, by using the output of a part of speech tagger, it eliminates errors caused by poorly guessing which words in the sentence are verbs. Because there are many verbs with past tenses not ending in "-ed" and there are many sentence constructions other than "proper noun or pronoun followed immediately by verb" this can greatly improve output. For example, neither naive method can handle a sentence as simple as "Smith, a former running back, arrived on Tuesday" because the verb is split from the proper noun, while the verbs method will easily identify "arrived" as a verb and Tuesday as being in the past.

5.5 Phrases Method

The fourth and final tense identification method, termed the *phrases method*, is very similar to the verbs method, except that it is also aware of verb phrases. In particular, verbs occurring within two words of each other are treated as a single phrase, and the phrase is assigned a score of -2, +1, or +10, much like a verb in the verbs method. The score assigned to a phrase is simply the score of the first word in the phrase. For example, in the sentence, "the storm is expected to arrive Tuesday" the verbs method would assign it a score of 0 (1 for the present "is," -2 for the past participle "expected," and 1 for the infinitive "arrive"). However, the phrases method assigns the phrase "is expected to arrive" a score of 1 since this is the score for the leading word in the phrase. The result is that the phrases method correctly identifies "Tuesday" as a future occurrence, while the verbs method does not. As another example, consider the sentence "But a change of plea hearing is scheduled for Wednesday in federal court." The verbs method identifies the word "Wednesday" as being in the past because the past tense tagging of "scheduled" (-2) outweighs the present tense word "is" (1). On the other hand, the phrases method by being aware of verb phrases can recognize that "is" is the determining word for the tense, and correctly identifies that "Wednesday" is in the future.

6. EXPERIMENTAL RESULTS

We evaluated the different tense identification methods by manually tagging 350 different dates from news as being future dates or not future dates, and running the tense identifiers on them. In this sample set, 97 dates were in the future, and the remaining 253 were in the past (generally there is no present tense in news stories, as the news cannot be read at the same time it is written). We report precision and recall values for both the determination that an event was occurring in the future and the determination that an event was not occurring in the future. In addition,

Figure 2: Screenshot of Spatio-Temporal Twitter Demo

we report correct and incorrect numbers for both future and non-future events. so as to show how often the events were correctly identified. Unlike the analyses by [14] and [24], we do not report accuracy on "TIMEX" expressions; we only report the results of the complete temporal identification of dates in news. The results can be found in Table 1.

In terms of the number of correct identifications, the phrases method did slightly better than the verbs method (i.e., the method that only used verb tags but was not aware of verb phrases). For future events, the recall score of the phrases method was also better, although its precision was slightly worse. Both naive methods are very bad, and should only be used as a baseline for the purpose of a comparison to the other two methods. The naive with "-s" method has very bad precision but decent recall, while the naive with "will" method has very bad recall with decent precision. Although, in the context of a large news database, precision should be more important (because there are enough articles to determine the future event in another source if it is missed in one), the increase in recall was sufficiently great when using the phrases method that the outcome of its use was superior for determining future events in the corpus as a whole.

Experiments to evaluate the quality of the breaking news obtained from combining Twitter with our future detection module were performed over a two week period starting Oct 10, 2010. In particular, our event detecting module identified 680 events in the two week period. Recall, that each event is represented by a set of feature vectors, which form the input for the *track* API of Twitter. In total we obtained about 96.5 million Tweets, which we continuously processed for breaking news. We discarded all Tweets containing less than 3 words as not being of sufficient length to convey meaningful information. Next, after passing the remaining Tweets through our classifier that classifies Tweets into *spam* or *news*, we discarded 92.7 million Tweets as spam. Although this may seem too draconian, such a measure is necessary to ensure that our output is of a high quality. We obtained 3.8 million Tweets, which were classified as news Tweets by the classifier. Now, we tried to associate these Tweets with an ongoing event, using the technique described in Section 3. Recall, that a Tweet t is considered part of the breaking news if the distance between an event n and t is less than ϵ, when both of the events are represented in terms of their feature vectors as well as when both t and n contain γ features in common between them. In our experiments, we set ϵ to be 0.5 and γ to be 3. Our system identified 76,098 Tweets as breaking news. In order to evaluate the quality of the output, we chose 100 events at random and chose up to 100 Tweets for each of the events. A human evaluator determined if a Tweet t associated with news event n actually belonged to n. The output was computed in terms of precision, which in this case is defined as the number of incorrect Tweets associated with a news event n. Note that computing the recall of our system is not interesting as we discard a large percentage of the input Tweets. However, we do argue that such an aggressive approach to noise reduction is probably the only option available to us given that most of the Tweets are not related to news. The precision of our method was found to be 93.80%, which means that the output of our system is of a high quality, which has always been our goal. Moreover, almost all of the errors occurred in Tweets associated with events drawn from relatively *local* events, which were not of broad interest to the users of Twitter. A simple way to eliminate these errors would be to discard events drawn from news clusters, which have relatively few news articles in them.

7. CONCLUDING REMARKS

Although the precision of our event detection methods is too low to definitively mark future events on an article-by-article basis, future events can be identified with a high degree of accuracy by comparing the classification results with other articles in the cluster. News stories about future events both in the near future (a few days) and the more distant future (a few months) were routinely identified. The resulting articles can be used as a way of determining what is currently happening in the world or as a seed to find reliable information on Twitter.

Once events are identified along with location and keywords, it is fairly simple to filter Twitter traffic around the time or place of the event based on those keywords. This identifies posts about the event, helping indicate which posts are reliable and which users are knowledgeable about certain areas or events. The resulting system allows a display of news that is both spatial and temporal, with input on ongoing events provided by Twitter. A screenshot of this application, based on our NewsStand [28] and TwitterStand [28] applications is provided in Figure 2.

Our system is designed to answer the following three questions "What is happening at X?" (a location-based query [2]), "Where is topic T or article Y happening?" (a feature-based query as in spatial data mining [2]), and "When is a topic T or article Y happening?". Consequently, our user interface has two selection modes corresponding to the "What" and "Where" modes and a slider that allows selection of news articles or topics based on the "When" criterion. The controls for the "What" and "Where" are radio buttons on the top right of the screenshot in Figure 2, while the slider on top left corner of Figure 2 allows filtering of news topics based on the "When" criterion. The users can interact with the map using *pan* and *zoom* to retrieve additional news articles. As users pan and zoom on the map, the map is constantly updated to retrieve new topics for the viewing window, thus keeping the window filled with topics, regardless of position or zoom level. A given view of the map attempts to produce a summary of the news topics in the view, providing a mixture of topic significance and geographic spread of the topics. Users interested in a smaller or larger geographic region than the map shows can zoom in or out to retrieve more topics involving that region. A slider provides dynamic control in restricting article based on the reference to the future event they contain. In particular, our system improves on both NewsStand and TwitterStand in the sense that it exposes the user to the temporal aspect of events in news. Moreover, for news stories in our news interface that belong to the set of active stories, we constantly post updates as we find them in the form of Tweets posted in Twitter.

Future work involves using more advanced techniques from computational linguistics to increase the rate of correct tense identification. In particular, improvements to the analysis

of verb phrases and modifications of the scoring parameters for sentences could be very helpful. Also, combining our techniques for tense identification with those of [14] or [24] could be productive. In addition, the linking of articles to future events allows the creation of a temporal news network, providing insight into how a story is reported before and after it occurs, and making it easier to determine how a story compared with its expectations. A temporal news network could also provide insights into the causes of events by providing a clear sequence. Placing this sequence on a map as in Figure 2 enables easy knowledge discovery about the temporal and spatial relationship between events. This could be aided by making use of spatial browsers and libraries [21, 25]. Finally, since the methods described were more effective at finding past events than future events, past events could also be cataloged. This could help to both identify reliable Twitter traffic (by looking through posts around the time and place of the event) and to support queries about what was happening on a given day. For example, if enough news stories were captured and analyzed, one could identify a number of events that were happening on a given day in any country, state, or even major city. By interacting with Twitter, historical researchers could find out how a group of people reacted to a given event (or set of events, type of event, etc.) with a high degree of accuracy. Additionally, finding past events as well as future events would enhance the usefulness of any application dedicated to helping people learn about events. By making it easy to compare news coverage to Twitter posts about an event, our system offers both up-to-the-minute information and valuable insight into past events.

8. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging web content. In *SIGIR*, pp. 273–280, Sheffield, UK, July 2004.

[2] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *PODS*, pp. 265–272, Nashville, TN, April 1990.

[3] J. L. Bybee and D. I. Slobin. Rules and schemas in the development and use of the English past tense. *Language*, 58(2):265–289, 1982.

[4] A. Chowdhury. State of Twitter spam. `http://blog.twitter.com/2010/03/state-of-twitter-spam.html`, March 2010.

[5] E. Garbin and I. Mani. Disambiguating toponyms in news. In *HLT/EMNLP*, pp. 363–370, Vancouver, Canada, October 2005.

[6] M. Helft. At Google, slow growth in news site. `http://www.nytimes.com/2008/06/24/technology/24google.html`, June 2008.

[7] W. Kienreich, M. Granitzer, and M. Lux. Geospatial anchoring of encyclopedia articles. In *InfoVis*, pp. 211–215, London, July 2006.

[8] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, University of Edinburgh, Edinburgh, Scotland, 2007.

[9] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *SIGIR*, pp. 843–852, Beijing, China, July 2011.

[10] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *GIR*, Zurich, Switzerland, February 2010.

[11] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pp. 201–212, Long Beach, CA, March 2010.

[12] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: architecture of a spatio-textual search engine. In *ACM GIS* , pp. 186–193, Seattle, WA, Nov. 2007.

[13] C. X. Ling. Learning the past tense of English verbs: The symbolic pattern associator vs. connectionist models. *J. of Arti. Intll. Res.*, 1:209–229, 1994.

[14] I. Mani and G. Wilson. Robust temporal processing of news. In *ACL*, pp. 69–76, Morristown, NJ, USA, 2000.

[15] R. Mansfield. Johnny Depp alive after twitter hoax. `http://news.sky.com/home/technology/article/15535854`, January 2010.

[16] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, NY, 1997.

[17] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *ACM GIS*, pp. 43–52, San Jose, CA, Nov. 2010.

[18] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *HLT-NAACL*, pp. 50–54, Edmonton, Canada, May 2003.

[19] D. E. Rumelhart and J. L. McClelland. *On Learning the Past Tenses of English Verbs*. MIT Press, Cambridge, MA, USA, 1986.

[20] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Proc. & Mngt.*, 24(5):513–523, 1988.

[21] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *CACM*, 46(1):63–66, January 2003.

[22] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. Adapting a map query interface for a gesturing touch screen interface. In *WWW (Companion Volume)*, pp. 257–260, Hyderabad, India, March-April 2011.

[23] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling. Twitterstand: News in tweets. In *ACM GIS*, pp. 42–51, Seattle, WA, November 2009.

[24] F. Schilder and C. Habel. From temporal expressions to temporal information: semantic tagging of news messages. In *Proc. on Temporal and Spatial Inf. Proc.*, pp. 1–8, Morristown, NJ, USA, 2001.

[25] C. A. Shaffer, H. Samet, and R. C. Nelson. QUILT: a geographic information system based on quadtrees. *Int. J. of Geo. Inf. Sys.*, 4(2):103–131, April–June 1990.

[26] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, pp. 1–20, Boston, MA, Aug. 2000.

[27] T. Stowell. The tense of infinitives. *Linguistic Inquiry*, 13(2):561–570, Summer 1982.

[28] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: A new view on news. In *ACM GIS*, pp. 144–153, Irvine, CA, November 2008.

[29] M. Wick and B. Vatant. The geonames geographical database. `http://www.geonames.org/`.

[30] J. Yarow. Twitter finally reveals all its secret stats. http://www.businessinsider.com/twitter-stats-2010-4, April 2010.

[31] N. Yasuda, T. Hirao, J. Suzuki, and H. Isozaki. Identifying bloggers' residential areas. In *AAAI-CAAW*, Palo Alto, CA, March 2006.