Detecting environmental disasters in digital news archives

Amelia Yzaguirre, Robert Warren, Mike Smit
Dalhousie University
Halifax, Canada
Email: yzaguirre,rhwarren,msmit@dal.ca

***Abstract*—Automatically extracting events from large, unstructured/semi-structured textual data requires a mechanism for identifying the event, abstracting it from the text, validating the event's occurrence against some known values, and sharing the event with users effectively. Inherent in the challenge of Big Data is that it often exceeds a scale at which humans can effectively operate. In this paper, we focus on the domain of archived newspaper articles, and describe a system that generates a collection of event summaries from unstructured text, extracts a geographic marker for the event, and stores both in an on-line database that can be searched and/or visualized using an interactive map. The system relies on text mining techniques to filter out a dataset of news stories from a digital news archive source and extracts 1-2 sentences from each event to be stored in the database. We illustrate this approach using a flood database case study, automatically extracting descriptions of past flooding events occurring in Nova Scotia, Canada from a 20-year archive of regional newspaper articles. We validate our event extraction in two dimensions (identification of articles mentioning flood events; identification of accurate geographic markers from articles about flood events) using Amazon's Mechanical Turk (MTurk) to obtain human assessments at scale.**

***Keywords*-natural language processing; information extraction; mechanical turk; geographic information systems; on-line visualization;**

I. Introduction

Digital newspaper archives are rich sources of cultural, social, and historical information where news stories are unstructured text and organization is by date rather than by topic or location. Since a newspaper's general priority is the timely and accurate reporting of relevant news, rather than long-term secondary use of their articles, their archival value is not immediately accessible or valuable to a general user.

Manual examination of past newspaper articles is a common methodology for building databases of past events. As such, there is a need for methods that not only exploit accurate detail in newspaper articles to identify events of interest to an analyst, but also to a) summarize these events into a rich data set, and b) effectively visualize this data set with a tool that facilitates exploration and discovery for a general user.

We describe an approach using filtering and extraction (a form of summarization) of events from news stories taken from a large data set containing unstructured textual data. The filtering step uses keywords specific to the chosen event to pre-filter the article set, and to assess the remaining articles based on the vocabulary of the chosen event. The articles remaining after filtering are assessed based on geography, focusing on articles that mention place names in a gazetteer of toponyms from the geographic area of focus. Summarization techniques are used to (1) extract text describing the event from the article, which is then placed into the final database, and (2) to (when applicable) manually verify the relevance of the selected article from the full dataset and (3) to validate the accuracy of the geo-location for the extracted event.

To illustrate our approach, we describe a case study in which we seek to extract flooding events that occurred in Nova Scotia, Canada, from 2 million news articles dating from 1992 to 2015 created and maintained by a regional newspaper, *The Chronicle Herald*. Flood databases are typically created manually from news articles and insurance reports; Nova Scotia's had not been updated since 1987. A geo-referenced database of flood events helps persons working on flood impact mitigation, urban planning, emergency management, and more ([1], [2], [3], [4]). The standard for other provinces in this region is a basic textual interface to retrieve descriptions of flood events from a database (e.g., The Province of New Brunswick offers http://www.elgegl.gnb.ca/0001/en/Home/Main and Public Safety Canada http://cdd.publicsafety.gc.ca/). To inform our filtering step, we focused first on generating true positives for news stories about relevant flood events and second on linking the article with a mappable set of coordinates for a place in Nova Scotia. The summarization step was facilitated by a vocabulary generated from a pre-compiled collection of flood events from 1759-1987 [5] and relied on a modification of the term frequency-inverse document frequency (TF-IDF) metric. For the case study, we developed a small web application to provide a visualization accompaniment to facilitate use by experts and non-experts alike. A map provides a natural and interactive way to visualize the database through space and time.

We validated our approach, and the data set produced in our case study, by having a randomly chosen selection of newspaper articles manually catalogued by human assessors. We automatically generated human intelligence tasks (HITs) asking users of Amazon Mechanical Turk (MTurk) to assess the relevance of a pool of articles (half identified by our

approach as flood-relevant; half as not flood relevant but matching some keywords). We then created a second set of HITs for articles identified as flood-relevant, where we extracted toponyms from throughout the article and asked human assessors to identify which was the actual location of the flood, comparing the result to the one produced by our algorithm. Finally, we manually assessed a selection of articles to better understand how human assessors and our algorithm compared when geo-locating articles. Our approach achieved accuracy as high as 95% for the first tasks, and in general met or exceeded typical rates for similar automated tasks.

The three primary contributions of this paper are a) an approach to extracting event details from an unstructured corpus of newspaper articles (described in Section III); b) a case study illustrating the use of event extraction and geo-location to produce a usable flood database and visualization tool (used in Section III to illustrate our approach; and c) an approach to validating Big Data event extraction using human assessors at scale (Section IV). We begin with background and related work in Section II, and conclude the paper in Section V.

II. Background and Related Work

News articles are not a new data source being used for this purpose, and their credibility as a data source continues to be studied in various contexts (see 1986 – [6]; 2005 – [7]; 2007 – [8]), nevertheless they continue to provide a large and always growing collection of historical and modern-day documents containing information that aims to be precise and accurate in descriptions and reports on disastrous events.

The geo-tagging of news articles is commonly faced with the problem of homographs (see e.g. Table I), wherein references to cities and landmarks become ambiguous due to sharing names with people (e.g., Chelsea and Murray are towns in Nova Scotia), common words (e.g., Victory is a town in Nova Scotia), or different locations (e.g., Sydney, Oxford, Brooklyn, Cleveland are all in Nova Scotia, too). Language used to describe an event also poses another challenge for event extraction. For example, a simple keyword search for the appearance of the word "flood" in a news article can dramatically reduce the sample size of all news articles, though it only yields a set of possibly relevant articles since "flood" can often be used to describe a variety of non-environmental scenarios.

In 2012 Leetaru [9] discussed the rise of "born geographic" information, stating that full-text geo-coding of large document collections to create new spatial visualization, interaction, and search capabilities is a powerful new way to understand and use information. Algorithms that automatically extract geo-spatial terms from text to support geo-referenced document indexing and retrieval have had successes, for example the Geo-Referenced Information Processing System (GIPSY) described by [10] matches geographic names in text to all possible spatial coordinates without disambiguation and attempts to match such phrases as "south of Lake Tahoe" with fuzzy polygons that delineate the region. An interesting overview of such geo-coding systems is provided in the related work section of [11].

Continent	% places w/ multiple names	% names w/ multiple places
North & Central America	11.5	57.1
Oceania	6.9	29.2
South America	11.6	25.0
Asia	32.7	20.3
Africa	27.0	18.2
Europe	18.2	16.6

Table I
From the *Getty Thesaurus of Geographic Names*: places with multiple names and names applied to more than one place

The value of flood databases and (primarily, manual) methods for creating them, has been established in case studies for Athens, Greece [1] and Catalonia [12]. Researchers have also employed flood databases and GIS tools for predicting floods [3], [13] and assessing flood hazard [4], building hydrodynamic and hydrological models [2], [14], flood susceptibility mapping [2], emergency management preparation [15], and flood plain modelling [16].

Sentence extraction can be traced back to Luhn's 1958 paper [17] on automatic methods for creating abstracts of technical papers and magazine articles, and continues today in studies ranging from searching and indexing historical handwritten collections [18] to extracting customers' concerns via summaries of their product reviews [19].

Our contribution in this realm of work was to focus on extracting a succinct and relevant snippet for each of the news articles that would make up the flood database. The snippet would be a relevant and interesting summarization of the news article for users interested in the database's collection of events.

III. Approach

Our approach to identifying, geo-referencing, and storing events of interest from an unstructured corpus of news articles will be described using a running case study to illustrate the steps. We begin by introducing the context of the case study and the sample data set, then describe the two key steps of the approach (filtering and extraction) in detail. We conclude this section by describing our web application for exploring the dataset.

A. Case Study

The flood database was commissioned by Nova Scotia Environment, a department of the Province of Nova Scotia, Canada, as part of their Climate Change Adaptation project. Understanding the history of floods in the province informs efforts to understand and measure the frequency of natural

disasters which are often attributed to climate change. Manual approaches are time-consuming and expensive, and they were interested in employing a text analytics approach.

The major regional newspaper, *The Chronicle Herald*, agreed to provide access to their digitized news articles for this project. This provided us with a text corpus of 2 million articles; after an initial data cleaning/filtering step, removing non-news articles from the database, our raw data set included $1,210,476$ digital articles in XML format dating from January 1992 to February 2015. Some articles had malformed XML (less than $.001\%$ of the data set) and were excluded. The XML documents had some structure, but most of the metadata was not relevant to this task; we used the text of the news article, the headline, the publication date, and a unique ID of the article to generate appropriate back references.

Our goal was to identify articles that described flooding events in Nova Scotia, and produce a structured dataset that included a series of flooding events, complete with a date, the geo-tagged location of the event, a snippet from the full story text, a list of other places mentioned in the article, and a full citation of the article (author, headline, publication date).

B. Filtering news articles

The first stage of our approach requires several filtering steps. The first is to retain articles that include one or more relevant keywords. In our case study, we first filtered for articles which included the *flood* keyword in their story text; this keyword requirement reduced the data set to $13,659$ articles. Keyword selection is important to ensure that all relevant articles are recalled; at this stage false positives are unimportant as long as all true positives are included (that is, recall is far more important than precision). Based on our manual review of several hundred articles, and domain knowledge, it would have been difficult and wordy to describe a natural disaster involving water without using the word 'flood', and given the domain of newspaper articles, this keyword was deemed likely to capture all true positives.

In the second filtering step, we use a domain-specific vocabulary to identify words and phrases we expect true positives to include. For our case study, in order to rank the set of $13,659$ articles as true positives for a flood event we used a flood database compiled out of [5] that included causes, description, magnitude, region, river, and damage information for a collection of 183 flood events in Nova Scotia prior to 1987 (note there is no overlap with our dataset). We obtained frequency counts for each word in this text, excluding stop words; we also excluded named entities from this list. The top 100 words are shown in Figure 1. We converted frequency scores into relative weights; see the top ten words and their weights in Table II. This vocabulary was then used to remove false positives from our data set.

Figure 1. Building a vocabulary using word counts (visualization from TagCrowd [20])

weight	word	weight	word
10	river	6	damage
8	flood	5	heavy
8	rain	4	jam
6	inch	4	recorded
6	ice	3	water

Table II
TOP TEN VOCABULARY WORDS FOR TEXT CATEGORIZATION ALONG WITH THEIR RELATIVE WEIGHTING.

We counted occurrences of each vocabulary word in a news article's story text to obtain a score for that article which was normalized against its total words to give us a percentage. A *relevance threshold* of $\geq 3\%$ was used to determine which articles to keep in the data set (this threshold was arbitrary and tuned based on manual review of a selection of articles; a more robust method for determining a threshold is part of our future work).

In the third step all articles that meet the relevance threshold undergo a third filtering process where we remove the stories that have one vocabulary word that appears more often than the others. This meant that articles that scored high due to a single word such as 'damage' or 'heavy' would not be retained after setting an arbitrary threshold tuned for the dataset. In our case study this tuning was used to exclude

MM/DD	Section	Headline
10/18	World	Wilmington, N.C. - Hurricane Irene...
09/22	Observer	Jacksonville, N.C. - Most residents stranded...
09/24	FrontPage	Oxford - Torrential rains...
12/28	NovaScotia	Look out, Mother Nature!...
08/08	World	Seoul - Koreans called it The Water Demon

Table III
A SAMPLE OF NEWS ARTICLES FROM 1999 CATEGORIZED BY OUR SYSTEM AS FLOOD EVENTS.

stories about damages from other types of violent events, such as home invasions or criminal reports. These two steps reduced our data set to a collection of $1,223$ news articles, or about 9% of the original data set; see Table III for a very brief sample of the retained articles.

The fourth and final filtering step selects only the articles in the geographic area of interest to us; in the case of our study, Nova Scotia. We anticipate often not having training data appropriate for news articles until further extension of our approach to larger data sets. Thus, a heuristic approach is employed, which is effective as our interest is primarily in identifying geographic names and not in the broader task of named entity recognition and categorization.

The Geonames gazetteer restricted to places in the province of Nova Scotia only is used to identify articles with toponyms (place names) in our area of study. We used the Natural Language Toolkit (NLTK) [21], a suite of program modules to support natural language processing in Python, in order to extract proper names from our news articles. A helpful feature of news articles is that they generally mention multiple places of interest in a single article; this allowed us to produce a list of extracted toponyms and score an article based on the number of places it contained that were places found in Nova Scotia. We require at least two non-ambiguous location matches for the article to be retained. We also removed articles whose lists contain more geographic locations that are also place names outside of Nova Scotia than those place names that are only in Nova Scotia, which (for example) helped to eliminate news articles that took place in Sydney, Australia and not Sydney, Nova Scotia. Lastly, we disambiguated locations from those that were person's names by excluding the place if the entity was preceded in the article with a title (e.g. Col., Miss, Mr., Ms., etc.).

The list of toponyms is retained for the database, but for visualization purposes we preferred a single geospatial marker. To pick one location we explored several options; ultimately, we take the toponym most closely collocated to the word 'flood' in the article's text.

C. Snippet extraction from news articles

In addition to identifying a set of flood-relevant articles (the filtering stage), and producing a structured dataset with dates, geo-locations, a list of toponyms, and an article citation, we also wished to produce a snippet from the full story text. A well-chosen snippet helps users of the database quickly understand the substance of the article, which will improve usability. Additionally, the news articles are valuable intellectual property and cannot be displayed in their complete form. The summarization step identifies the most important portions of the article to produce these snippets (and thus is key to the event extraction).

We first manually selected 18 articles from 1992 that clearly mention events of interest (in our case study, flood events) from our data set. These articles form an *exemplar sub-corpus* which is used as the document base relative to which we calculate a similarity measure for sentences in our filtered set of news articles (a variant of a term-frequency inverse document-frequency (TF-IDF) measure, using cosine similarity). For each of these documents, we identify flood-relevant sentences automatically and generate a document frequency value. To generate an article snippet, we calculate the term frequency for every sentence in the news article relative to the document frequency of a randomly chosen snippet from the sub-corpus, and thus produce a modified TF-IDF score for each sentence. The sentence with the highest TF-IDF score was taken as that article's snippet. The sub-corpus helps ensure there is some variety in the snippet database, while identifying sentences that are relevant.

It is worth noting that before finding a method that worked, we tried to calculate the TF-IDF using two different methods: 1.) Cosine similarity between the full story text for a given article and each full story text of a sub-corpus article was calculated; these were summed to generate a total score for that article's relevance as a flood article. 2.) Cosine similarity between the sentences that contained the word 'flood' of a given article and those of the sub-corpus articles was calculated; these were summed to yield a flood relevance score. The scores obtained through these methods did not disambiguate articles in a useful way.

D. Database and Web Application

The final database of news articles contains 595 records of flood events in Nova Scotia from 1992 to 2015. Since it is common for events to take place at the same location, we opted to display each record at a unique point by algorithmically spreading out each point in circles centred at the original location's longitude and latitude coordinates. Figure 4 illustrates this spreading out for stories with a location of Halifax or Dartmouth. Clicking on a marker opens an info window for that record with the story snippet, a list of other locations mentioned in the article, the season that the story occurred in, a full citation for the record, and a link to the errata page to report the record for re-evaluation – see Figure 3. The List View button directs the user to a full list of all records, which can be sorted at the top by Year, Location, Citation, Snippet, or Other Places.

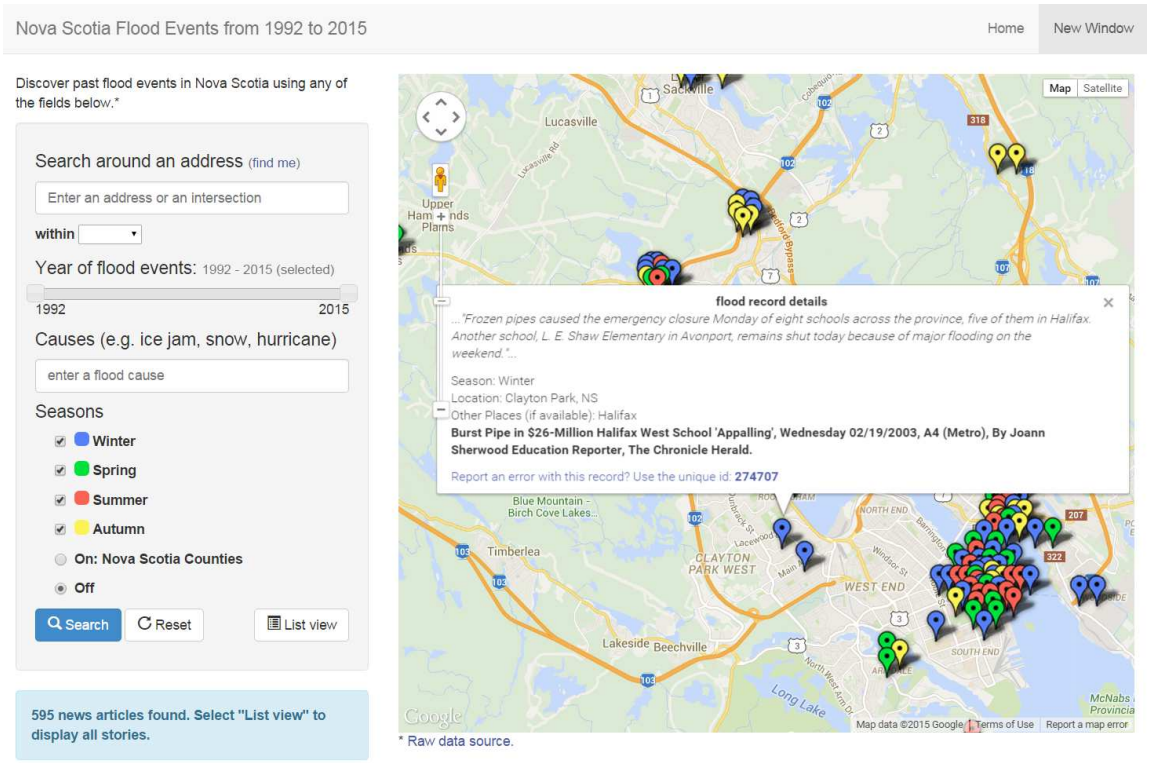

Figure 3. The web application for the final database. An info window appears for each marker on the map; it contains the story snippet, season and citation information.

Random seed snippet:

> The weekend's heavy rain caused a sewer backup which flooded the basements of an apartment building and about 10 area homes on Perron Street.

Extracted snippet from news article:

"... Wild winds overnight Saturday caused a power outage in about 22,500 Halifax households. ... A wire fastened to the pole broke free in the wind, connected with another wire and shorted out, cutting power to homes in Clayton Park, Fairview and Spryfield. Heavy rain that accompanied the high winds also caused problems for some metro residents by flooding their basements. ... A store clerk said customers had been coming in all morning, complaining of basements flooded by cracks in foundations, roof leaks or sewer backups. Several people also had frozen pipes in their homes burst. ..."

Figure 2. Using cosine similarity to extract a sentence snippet from a news article.

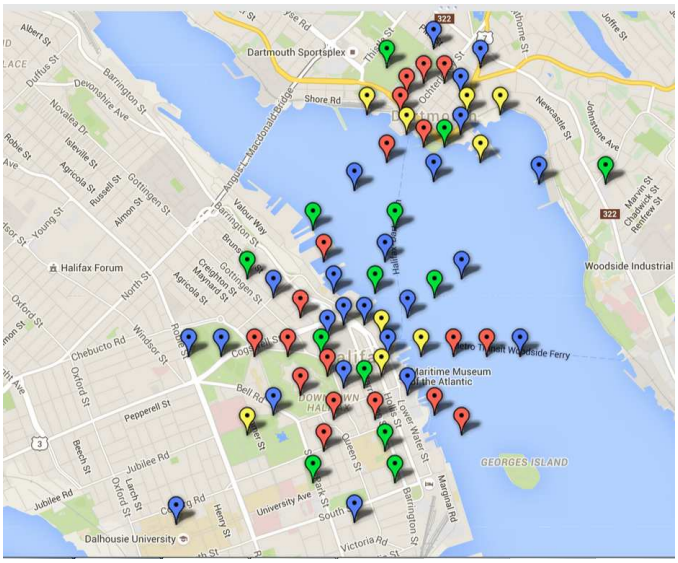

Figure 4. A screenshot of the map zoomed in near two common geo-location tags, Halifax and Dartmouth.

The database is searchable using any of the radio buttons in the left sidebar: address, year, keyword (e.g. causes of flood events), and seasons. For instance, a user can decide to view all articles that have occurred 10 km of Cole Harbour over the last 10 years during the Summer, of which there are 11 news articles (Figure 5).

Primarily the web application focused on the community level location of the event using the geonames.org gazetteer. Alternative databases, such as the Open Street Map gazetteer, could allow us to report a refined geometry of the locations and features (river, street, etc.) being flooded

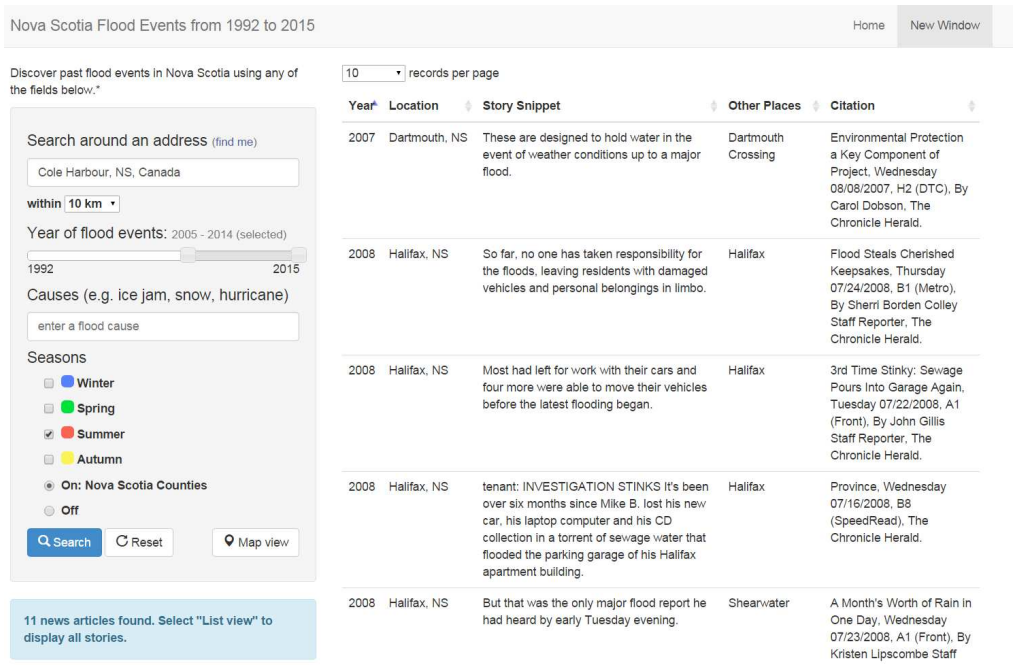

Figure 5. Querying for flood-relevant news articles 10 km of Cole Harbour over the last 10 years that have occurred during the Summer.

or marked as the cause of the flood.

Locating larger-scale disasters that cross borders, such as floods, remains a challenge.

IV. Evaluation and Results

To evaluate the performance of our event extraction approach, we compared the results of human assessment to the results of our approach. We used Amazon's Mechanical Turk service (MTurk) to recruit human assessors, provide them with tasks, and capture their responses. We then compared the human assessment to the results of our approach and to our own manual assessment.

We evaluated our approach in two key dimensions: 1. did it accurately filter the news articles to identify only those articles referring to flood events, and 2. did it accurately identify the most important toponym in the article during the geo-location step. We describe each stage of the evaluation in the following subsections; we begin with a brief description of MTurk.

A. Amazon Mechanical Turk

MTurk provides programmatic/batch access to human intelligence. A registered pool of Workers sign up to complete human intelligence tasks (HITs), which they can choose freely, complete at their own pace and in their own time, and are typically small short-duration tasks. These HITs are provided by Requesters, who have tasks that require human intelligence to complete. HITs can request content creation or transcription, but most commonly require the Worker to choose one or multiple responses from a set of options. The typical value of a multiple-choice HIT is around 2-5 cents USD.

We employed categorization-style HITs, which ask Workers to examine content and choose the correct category. The HITs each required the content to examine, and a list of possible categories. Additionally, MTurk advises that Requesters have some mechanism for ensuring that Workers responded accurately and did not simply choose responses at random. We mitigated this risk by requesting that 5 Workers evaluate each HIT, and rejecting responses that deviated from the majority too frequently.

There are several threats to validity incurred by employing Mechanical Turk. First, the selection of available answers to multiple choice questions can result in the only reasonable option being the choice that aligns with the approach we are testing. For example, when testing the geo-location step, the available responses should all be plausible given the article, and users must be provided with the option to identify that none of the responses match the article. Second, if a majority of users are not answering honestly or accurately, our approach will reject the (minority) of correct answers and accept the majority. To mitigate this threat, we rely on Amazon's use of SSN's to uniquely identify Workers, and their policy of deleting users who repeatedly or intentionally make errors.

B. Evaluating Filtering

To evaluate the accuracy of the filtering of news articles, we started with the 13,659 articles after the first filtering step (which simply looked for the 'flood' keyword). The term

'flood' is often used to refer to things other than flooding events, so we considered this the filtering step with the highest potential for error. We randomly selected 864 of these 13,659 news articles, half articles that appeared in our final database, and half articles that did not. We established five categories (based loosely on MTurks sentiment categorization HIT template); the positive categories reference flood events, the negative categories reference floods in some way but not events. Users were provided with this training (which was repeated as a prompt for each HIT):

Strongly Positive
: The article snippet is clearly discussing the occurrence of a flood event.
Example: "...other towns in the area reported minor street and parking lot flooding..."

Positive
: The article snippet mentions a flood but does not discuss any details about it. Other news is discussed besides the flood event.
Example: "...coal miners were killed in the first nine months of this year in fires, floods and explosions..."

Neutral
: The article snippet is educational. Its purpose is not to report a flood event.
Example: "...volcanic eruptions in Iceland often spark flash flooding from melting glacier ice but rarely cause deaths."

Negative
: The article's focus is unclear from the snippet.
Example: "one person's warming sunshine is another's killer drought, and one's gentle rainfall is another's disastrous flood."

Strongly negative
: The article snippet is not about an environmental flood event.
Example: "...when they were told she was a match, they were flooded with mixed emotions..."

For each news article five MTurk Workers were assigned to score the article from strongly positive to strongly negative, see Figure 6; this produced 4,320 categorizations. There were a total of 35 MTurk workers who participated in this task. We aggregated their scores for each HIT to determine if the score was positive or negative. The Workers collectively identified 412 articles as non-negative, all of them identified by our approach (no false negatives). Our approach identified 432 articles in total as being about flood events. This represents a false positive error rate of about 4.9%, for a total accuracy of over 95%. (One could narrow the evaluation to require that the aggregate score be at minimum "Positive", but in the MTurk context, this is ambitious; this score would not be achieved even if four rate it Flood Positive, but one is neutral. However, we can note that the rate would be 92% with this much stronger expectation, and still a strong performance).

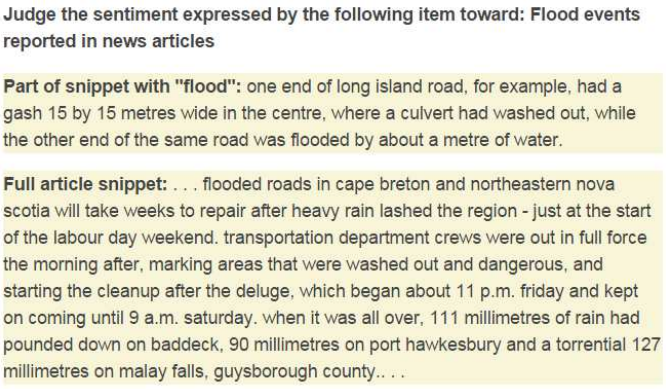

Figure 6. A sample MTurk Human Intelligence Task (HIT) to rank a news article.

C. Evaluating Geo-location

To evaluate the accuracy of the geo-location step of our approach, we focused on the news articles that MTurk Workers in the earlier evaluation identified as describing flood events. We asked them to tag each article with a GeoNames toponym. To prepare this HIT, for each news article we extracted a list of named places that had a match in the GeoNames gazetteer for Nova Scotia. MTurk's template for text categorization limited the amount of information and categories we could offer workers. We decided on a snippet of 5 sentences from the article, selected by the following procedure:

1) Index all sentences that include the appearance of a GeoNames toponym and the word 'flood'.
2) Index all sentences that include the word 'flood'.
3) For each sentence with a GeoNames toponym, calculate its nearness to the 'flood' sentences by taking the absolute value of the difference between the indices. Order the sentences from least to greatest difference.
4) Include all sentences from the first step, if there are less than five sentences include ordered sentences from step 3 up until there are five sentences.
5) If there are still less than five sentences, include all sentences indexed before and after those in step 2.

MTurk workers were then asked to categorize the location for the flood event by reading the snippets and selecting the best choice from seven categories: 1-4 were single locations, 5 listed remaining locations from the toponym extraction, 6 was 'none of the above' and the 7th category was 'no locations are provided', see Figure 7. Two workers were asked to answer for each news article. There were a total of 25 MTurk workers who participated in these HITs, 6 of which also participated in the filtering task.

First sentence: truro highway crews, police officers and firefighters were kept busy in colchester county saturday as the worst flooding in 25 years hit the area.

Article snippet: . . . longtime truro area residents are recalling the infamous valentine's day flood in 1971 as the benchmark for rating this weekend's rising water. the combination of warm air, heavy rain and melting snow that hit nova scotia on the weekend also caused flooding in other parts of the province, particularly bedford, sackville and parts of the south shore and cape breton. flooding in bible hill resulted in a call to the fire brigade which helped three residents of a main street trailer get to higher ground. the kiwanis park pond on truro's juniper street rose to road level with water flooding the nearby mt & t building and vehicles. homeowners on the south shore also had to get out their mops and buckets, with flooded basements reported in liverpool and first south, near lunenburg. . . .

A: bedford

B: truro

C: south shore

D: cape breton

E: nova scotia, bible hill, first south

None of the above: other

No locations are provided: none

Figure 7. A sample MTurk HIT to geo-tag a news article.

Comparator	Accuracy (%)	Geo-tagging Comparator	Accuracy (%)
Positive	92%	MTurk	52%
Non-negative	95%	Manual ($n = 20$)	60%
		MTurk (consistent)	64%

Table IV
EVALUATION OF EVENT EXTRACTION PERFORMANCE

Our system agreed with the geo-tagging from MTurk workers 52% of the time. The two MTurk workers agreed on a geo-tag for a news article 60% of the time. When MTurk workers were in agreement, our system agreed with them 64% of the time. As a third measure for the accuracy of our geo-tagging system, we randomly selected 20 news articles and hand-coded their location; interestingly, both our system and MTurk had a 60% accuracy against this data set. A summary of our results is provided in Table IV. These results suggest that location extraction is difficult even for human readers.

These performance results are the average for event extraction: tasks that rely on identifying named entities and relations held between them can be achieved with an accuracy varying from 80 to 90% [22], obtaining precision/recall figures oscillating around 60% for event extraction. For news corpora, studies have reported an 80% accuracy to classify an article as being about a global crisis event and a 60% accuracy in identifying a geo-spatial marker for the event [23].

V. CONCLUSION

We have described an approach to analyzing unstructured newspaper article text to identify events of interest, and illustrated this approach with a case study demonstrating the creation of a flood event database for a Canadian province. Our evaluation of this approach, in the context of the case study, demonstrated that identifying relevant articles was accurate (95%), but that geo-tagging these articles is difficult for humans and for automated methods (both achieving 60% accuracy against our oracle). Both accuracy levels meet or exceed typical performance in the literature. Our approach is general, and can be used in any context where it is useful to automatically explore large text corpora for events of interest.

We remain interested in the development of systems that allow users to access information through geographic attributes of documents. Our focus is on improving the categorization system and the extension of our system to other environmental disasters in Canada. Building a text corpus for such events would also facilitate work in this area, but there remain intellectual property challenges. We are interested in exploring mechanisms for providing segmented access to Big Data in newspaper contexts, where analysts can submit jobs and be charged only for the data they use (e.g. [24]).

ACKNOWLEDGMENT

The support of Nova Scotia Environment and The Chronicle Herald is gratefully acknowledged.

REFERENCES

[1] M. Diakakis, A. Pallikarakis, and K. Katsetsiadou, "Using a spatio temporal GIS database to monitor the spatial evolution of urban flooding phenomena. the case of Athens metropolitan area in Greece," *ISPRS International Journal of GeoInformation*, vol. 3, no. 1, pp. 96–109, 2014.

[2] B. Pradhan, "Flood susceptible mapping and risk area delineation using logistic regression, GIS and remote sensing," *Journal of Spatial Hydrology*, vol. 9, no. 2, 2010.

[3] A. Deschamps, D. Greenlee, T. Pultz, and R. Saper, "Geospatial data integration for applications in flood prediction and management in the red river basin," *IEEE International, Geoscience and Remote Sensing*, vol. 6, pp. 3338–3340, 2002.

[4] N. N. Kourgialas and G. P. Karatzas, "Flood management and a GIS modelling method to assess flood-hazard areas—a case study," *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, vol. 56, no. 2, pp. 212–225, 2011.

[5] A. D. Kindervater, *Flooding events in Nova Scotia : a historical perspective*. Inland Waters Directorate, Atlantic Region, 1977.

[6] C. Gaziano and K. McGrath, "Measuring the concept of credibility," *Journalism Quarterly*, vol. 63, no. 3, pp. 451–462, 1986.

[7] M. W. Downton and R. A. Pielke, "How accurate are disaster loss data? the case of u.s. flood damage," *Natural Hazards*, vol. 35, no. 2, pp. 211–228, 2005. [Online]. Available: http://dx.doi.org/10.1007/s11069-004-4808-4

[8] F. Pasquarè and M. Pozzetti, "Geological hazards, disasters and the media: The italian case study," *Quaternary International*, vol. 173, pp. 166–171, 2007.

[9] K. H. Leetaru, "Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia," *D-Lib Magazine*, vol. 18, no. 9, p. 5, 2012.

[10] A. Woodruff and C. Plaunt, "Gipsy: Automated geographic indexing of text documents," *JASIS*, vol. 45, no. 9, pp. 645–655, 1994.

[11] G. Crane and A. Jones, "The challenge of virginia banks: An evaluation of named entity analysis in a 19th-century newspaper collection," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '06. New York, NY, USA: ACM, 2006, pp. 31–40. [Online]. Available: http://doi.acm.org/10.1145/1141753.1141759

[12] M. Barnolas, L. Botija, and M. del Carmen, "A flood geodatabase and its climatological applications: the case of catalonia for the last century," 2007. [Online]. Available: http://dx.doi.org/10.5194/nhess-7-271-2007

[13] X. Yang, A. Grönlund, and S. Tanzilli, "Predicting flood inundation and risk using geographic information system and hydrodynamic model," *Geographic Information Sciences*, vol. 8, no. 1, pp. 48–57, 2002.

[14] J.-P. Fortin, R. Turcotte, S. Massicotte, R. Moussa, J. Fitzback, and J.-P. Villeneuve, "Distributed watershed model compatible with remote sensing and GIS data. i: Description of model," *Journal of Hydrologic Engineering*, vol. 6, no. 2, pp. 91–99, 2001.

[15] A. E. Gunes and J. P. Kovel, "Using GIS in emergency management operations," *Journal of Urban Planning and Development*, vol. 126, no. 3, pp. 136–149, 2000.

[16] C. Yang and C.-T. Tsai, "Development of a GIS-based flood information system for floodplain modeling and damage calculation1," *JAWRA Journal of the American Water Resources Association*, vol. 36, no. 3, pp. 567–577, 2000.

[17] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, Apr. 1958. [Online]. Available: http://dx.doi.org/10.1147/rd.22.0159

[18] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 139–152, 2007.

[19] J. Zhan, H. T. Loh, and Y. Liu, "Gather customer concerns from online product reviews–a text summarization approach," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2107–2115, 2009.

[20] D. Steinbock. Tagcrowd (2008). [Online]. Available: http://www.tagcrowd.com

[21] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[22] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[23] H. Tanev, J. Piskorski, and M. Atkinson, "Real-time news event extraction for global crisis monitoring," in *Natural Language and Information Systems*. Springer, 2008, pp. 207–218.

[24] M. Shtern, B. Simmons, M. Smit, and M. Litoiu, "Toward an ecosystem for precision sharing of segmented big data," in *Cloud Computing (CLOUD), 2013 IEEE Sixth International Conference on*. IEEE, 2013, pp. 335–342.