Using news articles for real-time cross-lingual event detection and filtering

Gregor Leban
Jožef Stefan Institute
Ljubljana, Slovenia
gregor.leban@ijs.si

Blaž Fortuna
Jožef Stefan Institute
Ljubljana, Slovenia
blaz.fortuna@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Ljubljana, Slovenia
marko.grobelnik@ijs.si

Abstract

The written medium through which we commonly learn about relevant news are news articles. Since there is an abundance of news articles that are written daily, the readers have a common problem of discovering the content of interest and still not be overwhelmed with the amount of it. In this paper we present a system called Event Registry which is able to group articles about an event across languages and extract from the articles core event information in a structured form. In this way, the amount of content that the reader has to check is significantly reduced while additionally providing the reader with a global coverage of each event. Since all event information is structured this also provides extensive and fine-grained options for information searching and filtering that are not available with current news aggregators.

1 Introduction

News publishers daily produce large numbers of news articles. Most of these articles describe happenings that are currently occurring in the world, such as natural disasters, meetings of important politicians, crime, business and sport events. Not all reported information is equally important – some events get higher media coverage, while other events get reported only by a small set of publishers.

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

In order to learn about current events, people nowadays usually either go to their favorite news publisher's web site and browse through the frontpage articles or they use of some type of aggregator, such as Flipboard or Bloomberg Terminal. Neither of the two approaches are optimal. By browsing a publisher's web site you typically learn about a small subset of current events (usually constrained to the geographic location of the news source) that are not necessarily unbiased and objective but instead implicitly promote political, social and religious views of the publisher/author. Using a news aggregator on the other hand can provide the readers with a coverage of the same events from multiple news sources, but unfortunately also overwhelms the reader with huge amounts of news articles (Bloomberg Terminal daily provides over 1 million articles). Using a news aggregator is also helpful since it usually allows one to specify a particular topic to follow, such as Business, Technology, Apple or Android. The list of topics is however quite narrow and does not allow one to specify long-tail interests.

In this paper we will describe a system called Event Registry [4] that tries to alleviate the aforementioned issues with news consumption and is freely available at [1]. Just as news aggregators it collects news articles published globally from more than 100,000 news sources in over 10 different languages. However, unlike the aggregators, Event Registry identifies from the articles the actual events that are being described in the articles. For Event Registry, an event is defined as any significant happening in the world that was reported in at least a few articles. Two examples of events are the death of David Bowie on Jan 11, 2016 that was reported in over 4,000 news articles as well as the news reported in 13 articles on Jan 23, 2016, that in Smithsonian's National Zoo, the Giant Panda was really enjoying the snow.

Grouping of news articles into events has several ad-

[1] http://eventregistry.org/

vantages. First, given an event, the reader can choose to read articles from various news sources that reported about the event. Providing the complete and global coverage of the event allows the reader to construct an unbiased view of the event and all related details. Secondly, when browsing through the current events, the reader does not have to go through hundreds of news articles, where several articles report about the same event. Instead, all articles about the same event are grouped together and shown only once, which easily reduces the amount of content for one or two orders of magnitude. Lastly, for each event in Event Registry there is also abundant semantic information that is extracted from the articles, such as the location of the event, date, who and what the event is about, etc. This semantic information allows the reader to determine very specifically what his interests are and get a custom-tailored feed of events and news.

The rest of the paper is organized as follows. We will first describe the process in which Event Registry identifies events from news articles. We will also describe in more details the process in which the articles about the same event can even be linked although they are written in different languages. Additionally we will also describe the concept of a topic page which can be used by readers to very specifically determine the news articles and events of interest. We end the paper with a conclusion and some ideas for future work.

2 Event Registry

Event Registry consists of a pipeline of services that collect, process and analyze news articles collected globally in different languages. We will now briefly describe the major components in the pipeline.

2.1 Collecting news

In order to collect the news we developed a service called Newsfeed [5] that monitors RSS feeds of over 100,000 news publishers. Whenever a new article is detected in a feed, we crawl the web page and extract from it the news article and the available metadata information. In this way we collect daily between 200,000 and 300,000 news articles in various languages.

2.2 Semantic enrichment

The collected news articles provide information in unstructured form which requires a human to interpret it.

One way in which we extract structured/semantic information from the articles is by identifying and disambiguating relevant entities (people, locations and organizations) and non-entities mentioned in the articles. Examples of relevant non-entities would be things, such as Zika virus, murder, movie, automobile, etc. Identification of concepts (entities + non-entities) is done by wikification, which is a process of entity linking that uses Wikipedia as the knowledge base. As a result, each mentioned concept is annotated with a URI that is the link to the corresponding Wikipedia page. Since Wikipedia provides pages for the same concept in several languages (Barack Obama has a Wikipedia page in 225 languages), the question is which URL to take as the concept URI. We use the link to the English Wikipedia, when it is available, and the link to original (article) language otherwise. "Normalizing" the concepts to the same URI is very important since it allows the readers to find content regardless of the language in which it is written. The URI for the concept of the Sun, for example, would be the same, regardless if it is found in an English, Slovene (as 'Sonce'), Italian (as 'Sole') or any other language. Along with the URI, we also compute the relevance of the concept for the article. The relevance is computed depending on the number of times the concept is mentioned as well as it's locations in text and can be in the range between 1 and 5.

Another type of semantic enrichment we perform is categorization of the news articles based on the article's content. Currently we categorize news articles into a DMOZ [1] taxonomy. This taxonomy contains over a million categories, but we only consider top 3 levels, which amounts to 5,000 categories. The taxonomy was built for organizing web pages so it is not the optimal fit for categorizing news content. A more appropriate categorization would be to the IPTC's Media Topics taxonomy [2], which contains about 1.400 topics structured into 3 levels. Unfortunately we have not yet been able to obtain an annotated corpus of articles that we could use to train the models for this taxonomy.

Additionally we also extract from news articles all mentions of dates. Extracting dates is relevant for the following steps when we want to determine when the event described in the text occurred. In order to extract the dates we created an extensive set of regular expressions for individual languages that can detect date mentions in various forms.

2.3 Clustering of news articles

In order to group all articles that describe the same event we use an online clustering algorithm. The clustering is applied on each language separately and in short works as follows. Each collected article is first represented as bag-of-words – a representation in which we only keep an unordered list of words from the article and the number of times they occurred in the article. After applying TF-IDF weighting we compute

the similarity of the article with centroids of existing clusters. The criteria that is used when computing similarity between the article and the cluster centroid are the cosine similarity of the text, similarity of the mentioned concepts and the date difference. If computed similarity of the most similar cluster is above the threshold, the article is put into the cluster, otherwise a new (micro) cluster is created, containing only the single article. Micro clusters are not considered to be events until they reach a certain number of articles. The threshold value for becoming an event depends on the language and was empirically determined to be between 3 – 6 articles.

News about an event are typically reported only for a limited amount of time. For this reason we also want to remove clusters after they reach a certain age. Currently, when a cluster becomes 5 days old we remove it, which means that new articles can not be assigned to it anymore. In this way we can maintain high performance of the system as well as prevent incorrect assignments of new events to old clusters.

2.4 Construction of events

Each time a micro-cluster of articles reaches a certain size, we form in Event Registry an event and associate it with the cluster of articles. Clustering has to be done for each language separately so each event is initially mono-lingual. Most relevant world events are however covered by various publishers globally that report in various languages. To represent such clusters as a single event we use a machine learning approach that will be described in more details in the next section.

Each created event is represented in Event Registry with a unique identifier that can be used to reference it. For each event we also want to extract it's core information – what occurred, where, who as involved, etc. To determine these details we use the available semantic and meta information provided by the articles assigned to the event.

To determine the date of the event, we can analyze the publishing date of the articles in the clusters. The naive approach would be to use the date of the first article as the date of the event. In practice this approach generates erroneous results for events that are reported in advance (such as various meetings of politicians, product announcements, etc.) as well as when the collected publishing dates of the articles are inaccurate. A more error prone approach that we use is to analyze the density of reporting and use the time point where the reporting intensified as the date of the event. Additional input can be provided by the mentioned date references – a particular date that is consistently mentioned across the articles most likely the correct date of the event.

In order to determine who is involved in the event we can analyze and aggregate the entities mentioned in the articles. A list of entities and their associated relevance can be obtained by analyzing the frequency of their occurrence in the articles as well as their assigned scores. Entities can be scored and ranked according to this criterion which provides an accurate aggregated view on what and who is the event about.

Location of the event is another important property. Since the event location is commonly mentioned in the articles, we can identify it by analyzing the frequently mentioned entities that are of type location. Additional signal for determining the event location can be obtained by inspecting the datelines of the articles. A dateline is a brief piece of text at the beginning of the news article that describes where and when the described story happened. The datelines are unfortunately not present in all news articles and even when they are, they sometimes represent the location where the story was written and not the actual location of the event. To determine which location, if any, is the event location, we apply an SVM classifier. Each mentioned city is considered to be a candidate for the event location and we generate for it a set of learning features. The features we use are based on the number times the city is mentioned in the articles and the number of times it is mentioned in the dateline. The SVM model that we use was trained on 200 events for which location was manually determined. Using 5-fold cross validation on this training data we found that the achieved classification accuracy of the model is 98%.

3 Cross-lingual linking of clusters

Since same events can be reported in multiple languages we need a way for identifying clusters in different languages that are discussing the same event so that they can be merged and represented as a single event. In short, we need an approach that given two clusters of articles determines if they describe the same event or not.

To perform the task we again represent it as a learning problem. From the two tested clusters we extract a set of learning features that can be used for training a classification model. There are three groups of learning features that we use:

Cross-lingual article similarity. Using an approach based on CCA [3] we can compute an estimated similarity between articles in different languages. Given this measure we can compute how similar individual articles in one cluster are to the individual articles in the other. From these results we can generate a number of learning features such as the maximum similarity, the average similarity, standard deviation, etc.

Concept-related features. Articles in Event Registry are annotated with concepts that have language independent URIs. For each cluster, we can analyze the associated articles and determine the top concepts based on how frequently they appear in these articles and what are their assigned scores. Using two such weighted vectors, one for each cluster, we can compute a list of informative features. Examples of these features include cosine and Jaccard similarities of the two vectors. Additional features can also be computed separately for the entities and non-entities in the vectors.

Miscellaneous features. Additional set of features can be computed reporting (a) whether the event locations found for the two clusters are the same or not, (b) the absolute difference in hours between the events in the two clusters and (c) the similarity of the dates that are being mentioned in the articles in the two clusters.

To evaluate how accurately we can, given these features, predict whether two clusters are about the same event or not, we performed the following experiment. Using two human experts we have manually annotated 808 pairs of clusters in English, Spanish and German language. The dataset contained 402 examples of cluster pairs that report about the same event and 406 examples where they do not. By training a linear SVM model and by using 10-fold cross validation schema we were able to achieve 89.2% classification accuracy.

4 Topic pages

Whenever an event is identified or updated, the information is stored in the Event Registry. Currently, Event Registry holds information about 3.6 million events that it identified from 88 million news articles, which were collected since January 2014. The users can use the web interface to search for events based on various criteria, such as relevant concepts, news sources that reported about it, location of the event, category, date, size and others. The users can also simply observe the stream of new/updated events as they are shown on the Event Registry home page.

An even more useful functionality than observing the whole feed of events, is the option for the users to create their own feed of articles and events based on their own interests. We call this functionality a topic page, where a topic can be defined using a set of relevant concepts, keywords, news sources and/or categories. The user can define the topic page using an interface shown in the top part of Figure 1. To each specified concept, keyword, news source and category, the user also assigns a weight of relevance for the topic. Each article and event that is processed by Event Registry is then scored according to the specified criteria and only those that achieve high enough score (a parameter specified by the user) are then shown to the user in the feed of the topic page.

More specifically, the scoring is done as follows. Let's assume that the user defines a topic T using a set of conditions $c_i, i = 1..n$ and their associated weights w_i, where conditions consist of one or more concepts, keywords, news sources and/or categories. For each new event e, a score $S_T(e)$ is computed as

$$S_T(e) = \sum_{i=1}^{n} w_i \cdot in(c_i, e) \cdot val(c_i, e)$$

$$in(c_i, e) = \begin{cases} 1 & c_i \in e \\ 0 & otherwise \end{cases}$$

$$val(c_i, e) = \begin{cases} e_{c_i}/100 & c_i \text{ is a concept} \\ 1 & otherwise \end{cases}$$

The score $S_T(e)$ is therefore a simple sum over all conditions, where for each condition c_i we multiply the associated weight w_i with a Boolean function $in(c_i, e)$ and a scoring function $val(c_i, e)$. Function $in(e, c_i)$ simply determines if the condition c_i matches the event e or not. In case the condition is a concept or a category, the function is true when the event is annotated with it. In case the condition is a news source, the function is true if the event contains an article written by the news source. Lastly, in case the condition is a keyword, the function is true if the keyword appears in any of the articles assigned to the event. The scoring function $val(c_i, e)$ is trivial, except in the cases when c_i is a concept. When concepts c_j are associated with an event e, they are assigned a score e_{c_j} that is in range between 1 and 100, which represents how important the concept is to the event. The function $val(c_i, e)$ therefore simply ensures that for all conditions, the returned value is in range between 0 and 1. The scoring function for scoring articles is almost the same, except that the normalization constant in function $val()$ is 5, each concept in an article is assigned a score between 1 and 5. The events and articles that match the topic page can be then visualized on a map or displayed in a feed. An example topic page for USA presidential elections is available at Figure 1.

5 Conclusion

In this paper we have presented a system called Event Registry with fixes several shortcomings in the ways how news content is currently being consumed. Firstly, it is able to aggregate large amounts of news articles into actual events. Instead of flipping through tens or hundreds of articles about the same event in your

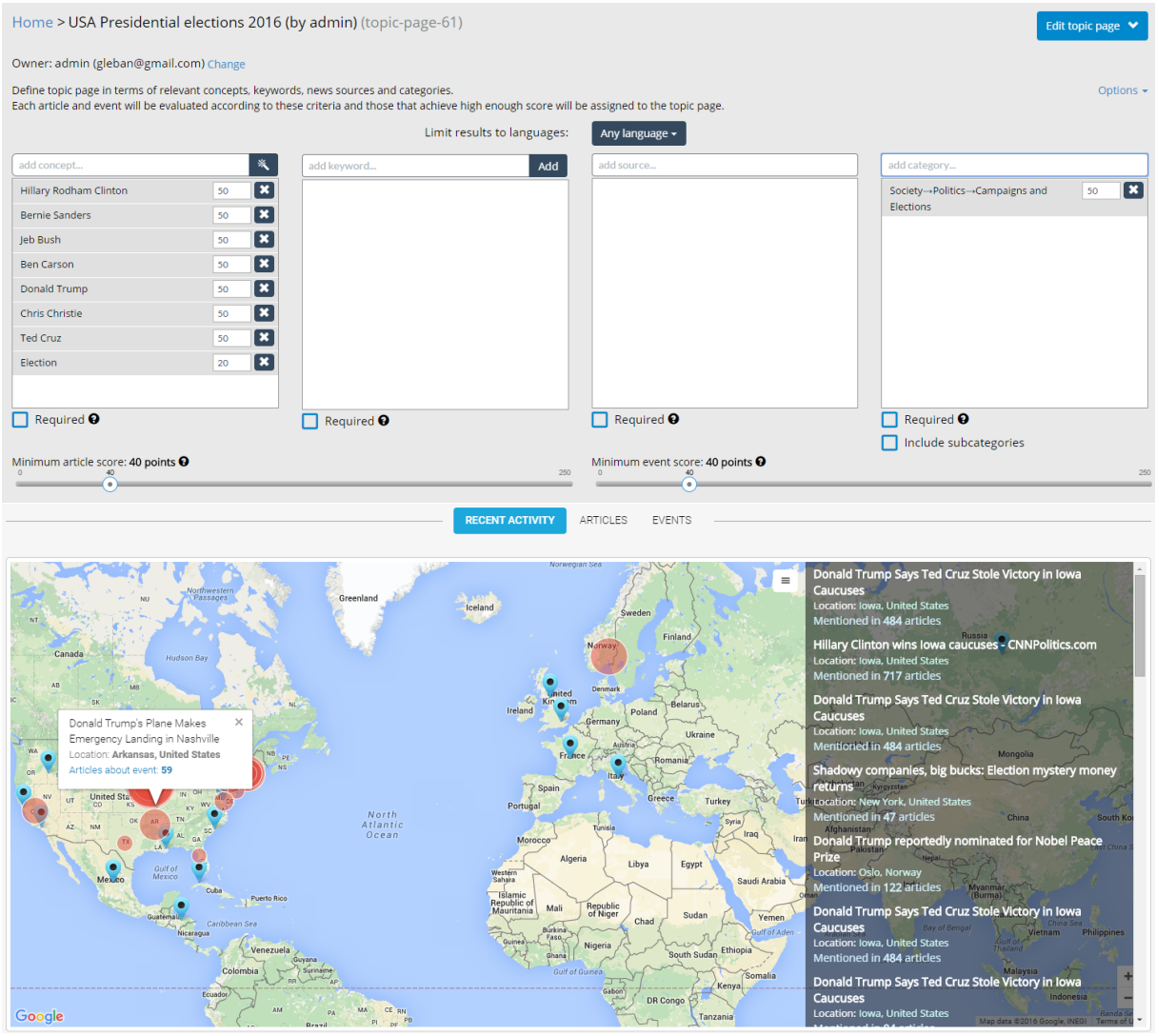

Figure 1: The interface for defining the topic page (top) and the feed of current events that match the criteria (bottom). The feed can be displayed on a map or as a list of matching articles and events.

news aggregator, a single item can be shown, together with the structured information about the event (who, what, when, where,...). If interested in the event, the user can then open the details of it and read individual articles (even in different languages) about it. By reading multiple articles, the user can form a more complete and unbiased view of the event as if he would be able to by just reading about it from a single news publisher. Having extensive structured information about the events allows the users of Event Registry to also create custom feeds based on a combination of general or long-tail topics of interest.

6 Acknowledgments

This work was supported by the Slovenian Research Agency as well as X-Like (ICT-288342-STREP) and xLime (ICT-611346-STREP) projects.

References

[1] DMoz, open directory project, *http://www.dmoz.org/*.

[2] Media topics, *https://iptc.org/standards/media-topics/*.

[3] S. T. Dumais, T. A. Letsche, M. L. Littman, and T. K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21, 1997.

[4] G. Leban and et. al. Event registry – learning about world events from news. In *Proceedings of*

23rd International World Wide Web Conference, 2014.

[5] M. Trampus and B. Novak. Internals of an aggregated web news feed. In *Proceedings of 15th Multiconference on Information Society 2012 (IS-2012)*, 2012.