Storybase: Towards Building a Knowledge Base for News Events

Zhaohui Wu[†], **Chen Liang**[‡], **C. Lee Giles**[‡†]
[†]Computer Science and Engineering, [‡]Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802, USA
zzw109@psu.edu, {cul226,giles}@ist.psu.edu

Abstract

To better organize and understand online news information, we propose Storybase[1], a knowledge base for news events that builds upon Wikipedia current events and daily Web news. It first constructs stories and their timelines based on Wikipedia current events and then detects and links daily news to enrich those Wikipedia stories with more comprehensive events. We encode events and develop efficient event clustering and chaining techniques in an event space. We demonstrate Storybase with a news events search engine that helps find historical and ongoing news stories and inspect their dynamic timelines.

1 Introduction

Users are often overwhelmed by the flood of information, especially frequently daily updated news. Search engines effectively find news snippets and related Web pages, or group similar pages in clusters. However, it remains difficult to coherently connect isolated information nuggets to form the big picture, or to accurately track the flow of information to show the evolution of events. For example, current news search engines or aggregation sites, such as Google or Yahoo news, show only isolated daily news events, without linking them to historical events or show storylines.

Most existing knowledge bases such as DBpedia and Freebase are designed for managing general named entities or concepts and often lack coverage or representation for temporally evolving news events. For example, as of this writing, Freebase has not treated "2014 Ferguson unrest" as an "event", let alone show its sub events or timelines. As such, we propose building a knowledge base, namely Storybase, that stores news events in a semantic coherent schema that could explicitly display their evolving timelines. We define a *story* as a set of topically or causally related and temporally ordered news events, usually corresponding to a Wikipedia article such as "Malaysia Airlines Flight 370". An event is defined as something important happening at some time in some place, reported by a set of news articles, which is encoded by named entities, actors and actions used as the main points in a plot.

Building an event knowledge base from scratch is challenging, since it is difficult to obtain a gold standard for events and their timelines. We found that Wikipedia current events[2] provide high-quality manually edited news events. To scale up, we link daily news sources and fit them into existing stories or create new stories, by efficient event detection and storyline construction techniques in a semantic space which is encoded with news events' entities, actors, and actions. From April 1, 2013 to March 1, 2015, we have collected 1,256 stories consisting of 35,362 news events from Wikipedia current events, and 35,166,735 daily news articles. Experimental evaluation compares our methods for event clustering and chaining with multiple baselines. We build a news event search engine based on Storybase to show news stories and their event chains.

Our main contributions include:

- A news event knowledge base, Storybase, with a search interface for news storylines;
- The introduction of Wikipedia current events as resources for building event knowledge bases and as datasets for event detection and storyline construction;
- New approaches for event clustering and chaining with experimental comparisons to other baselines.

[1] http://breckenridge.ist.psu.edu:8000/storybase

[2] http://en.wikipedia.org/wiki/Portal:Current_events

Proceedings of ACL-IJCNLP 2015 System Demonstrations, pages 133–138,
Beijing, China, July 26-31, 2015.

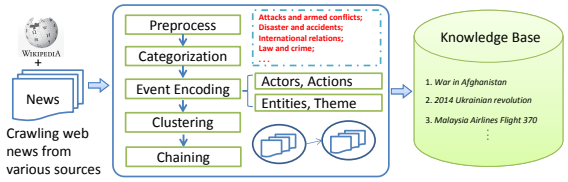

Figure 1: Overall process of building Storybase

Figure 2: Examples of Wikipedia current events

2 Overview and Definitions

Figure 1 shows the overall process for building Storybase. Input is daily crawled web news articles and Wikipedia current events. This generates the storylines and builds the Storybase using five steps system: preprocessing, categorization, event encoding, clustering, and chaining. Details are in Section 4. A news event search engine is built to provide a query based interface to search and visualize the Storybase, which is shown in Section 5.

We now define concepts that will be frequently referred to.

- A *event* identifies something (non-trivial) happening in a certain place at a certain time (Yang et al., 1999); it is a set of news articles on the same news report.
- A *story* is a set of topical related news events.
- A *storyline* is a series of temporally ordered events of the same story.
- *Actors* in an event are named entities that make or receive actions.
- *Actions* are verbs that connect actors.

For example, as shown in Figure 2, "Pro-Russian militants seize the regional prosecutor's office in the eastern Ukrainian city of Donetsk" is an event reported by a set of articles from different news sites. "2014 pro-Russian unrest in Ukraine" represents a story that consists of temporally evolving events, which forms a storyline. "Pro-Russian militants" and "the regional prosecutor's office" are actors while "seize" is the action.

3 Data Collection

Wikipedia current events list manually edited daily news events since 1998, which provide rich semantics and structure for news stories and events such as story names, event categories (*not* Wikipedia categories), and links to Wikipedia concepts, as shown by Figure 2. For example, we can observe that the event "Pro-Russian militants seize the regional prosecutor's office in the eastern Ukrainian city of Donetsk" belongs to the story "2014 pro-Russian unrest in Ukraine" and the category "Armed conflicts and attacks", containing links to Wikipedia concepts "Eastern Ukrainian" and "Donetsk". Thus, we construct a storyline for "2014 pro-Russian unrest in Ukraine" by connecting all events under it.

The category labels provide a natural way to classify news events. However, since the Wikipedia events are edited by various users, the category labels are not always consistent. For example, one may use "Armed conflicts and attacks" while others might use "Attack and conflict". After canonicalization using Levenshtein distance and grouping similar labels using word based Jaccard similarity, we manually clean all the labels into 12 categories, as shown in Table 1.

Although Wikipedia provides high quality manually edited news events, it covers only a small number of events every day, usually less than 30. Thus, to scale up Storybase and make the stories more comprehensive, starting from April 1, 2013, we crawl daily news articles from a large number of sources from various news publishers, provided by GDELT[3] project (Leetaru and Schrodt, 2013).

4 Building Storybase

4.1 Preprocess and Categorization

To extract and parse Wikipedia current events, we implement two template based extractors for events between January 2003 and April 2006 and those events after April 2006 respectively due to their difference in templates. The news articles linked at the end of each event description are also crawled. We use boilerpipe[4] to extract the title and main text content of each news article. We extract the first three sentences in the main content for summarization.

[3] http://www.gdeltproject.org/data.html
[4] https://code.google.com/p/boilerpipe/

ID	Category
1	conflict, attack
2	disaster, accident
3	international relations
4	politics and elections
5	law and crime
6	business and economy
7	science and technology
8	sports
9	arts and culture
10	health, medicine, environment
11	education
12	deaths

Table 1: Categories of events in Storybase

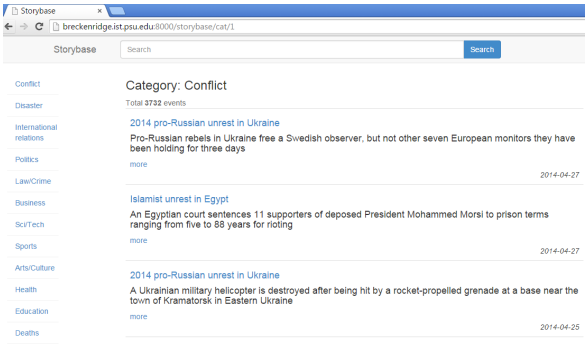

Figure 3: Screenshot of category "Conflict"

We maintain an N-to-1 mapping for each category listed in Table 1. For example, any category label in {"Armed conflicts and attacks", "conflicts and attacks", "Armed conflicts", "Attacks and conflicts", "Attacks and armed conflicts"} will be mapped to Category 1. For an event not belonging to existing stories, we label its category using the majority of their k-nearest (k=10) neighbors based on the cosine similarity of event descriptions.

4.2 Event Encoding

We encode an event as a vector containing named entities, actors and actions. Named entities such as people and locations in news reports contain important information of the event. Core entities that play important roles in an event are called actors, which are usually people or organizations that make or receive actions. We use the Stanford CoreNLP (Manning et al., 2014) for the named entity recognition and extract all Wikipedia concepts appearing in news content. Entities that are subjects or objects in the title and description are treated as actors. If no entities are found, we then use the CAMEO dictionaries[5] for actor and action extraction.

4.3 Event Clustering and Chaining

Event clustering groups together news on the same event. Locality-Sensitive Hashing (LSH) (Van Durme and Lall, 2010) is used for fast similarity comparison. We first do deduplication on all articles on the same date using 84 bits sim-Hashing (Charikar, 2002). We then use modified sim-Hashing on the vector space of event described in Section 4.2, rather than shingling or bag-of-words (Paulev et al., 2010). A new article is encoded into the event space with the content of its title and description. Its LSH key k (84 bits binary code) is computed and compared to keys of other articles. Articles whose keys have hamming distances smaller than a threshold θ among each other will be clustered as an event. We then check all events of the previous date and merge two events into one if their distance (average hamming distances of key pairs) is smaller than θ and their categories are the same.

Event chaining links an event to an existing story or determines if it is the starting event of a new story. While LSH could give high-purity event clusters, it might not be able to determine whether two events with distance larger than θ are topically related, or belong to the same story. Intuitively, an event should bring some novelty and preserve some common information compared with the previous ones in the story, causing a trade-off between relevance and novelty, which could be measured by some textual similarity. Adding an event should also keep the storyline coherent. To model coherence, we investigate two features, the Connecting-Dots coherence score (Shahaf and Guestrin, 2010) and KL-divergence. We use the gradient boosting tree (Friedman, 2001) to learn if an event belongs to a story by using the above features of relevance/novelty and coherence, all based on storylines constructed from Wikipedia current events. For a story $\{e_1,...,e_m\}$, $(e_i, \{e_1, ..., e_{i-1}\})$ are positive pairs; $(e_-, \{e_1, ..., e_{i-1}\})$ are negative pairs, $i = 2, ..., m$, where e_- is an event randomly sampled from other stories in the same date of e_i.

For all GDELT news on date t, we first detect all events using event clustering. For an event that has not been merged into events of the previous date, we use the model to decide which story it belongs to. If none, the event will be served as the first event of a new story with an empty story name.

[5]http://eventdata.parusanalytics.com/data.dir/cameo.html

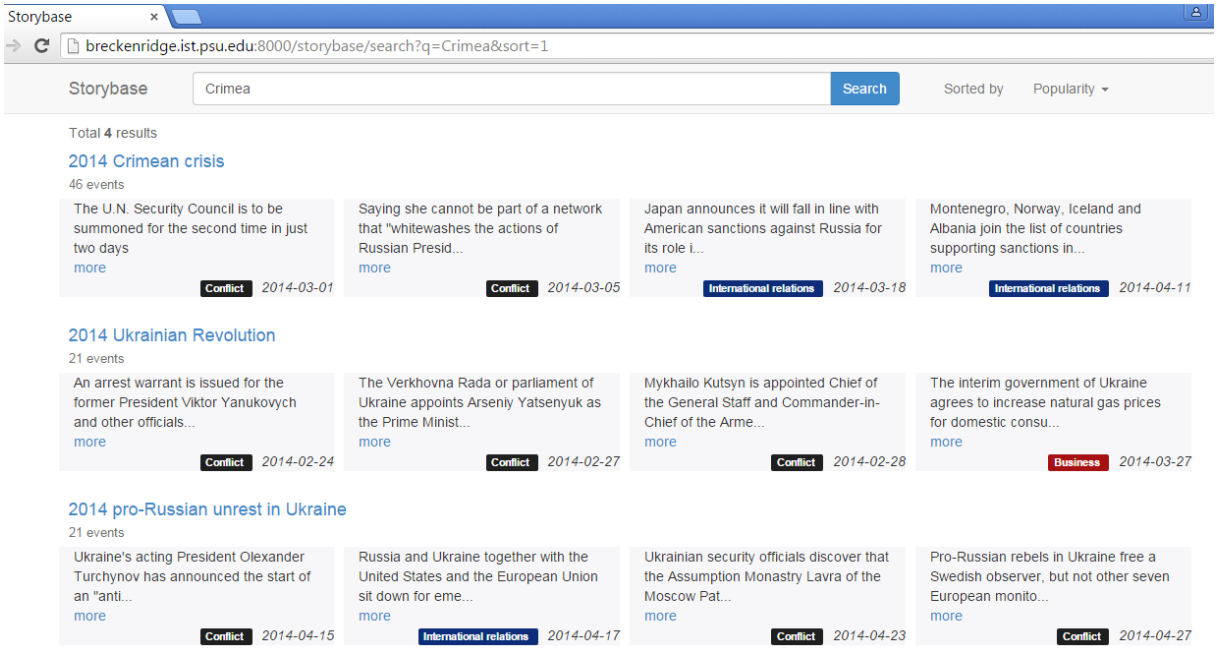

Figure 4: Screenshot results for the query "Crimea"

5 Storybase Demonstration

We demonstrate Storybase by building a news event search engine that can retrieve and visualize the stories. In the backend, we implemented various facilities, such as ranking functions (BM25, cosine similarity, and inner product) and refining metrics (popularity and recency). The ranking functions compute relevance between queries and stories while a story is represented by the story name and all event descriptions. *Popularity* measures the impact of stories on Web. For simplicity, we implement popularity as the accumulative number of unique news reports for all events of a story. *Recency* measures the timeliness or freshness, which is an important and helpful feature for sorting and filtering news stories, and is implemented by simply sorting stories based on the date of their latest event.

The front page gives a category navigation list in the left, a search box in the middle, and the recent stories behind the box. A category links to the recent events from the category, as shown by Figure 3. The demo contains three views: storyline, story, and event. Figure 4 shows a screenshot of the storyline view returned by querying "Crimea". The results are organized at the story level, where we show a thumbnail of the event chain for each story. The description, category, and date of an event are presented in the event box. By clicking the story name, it will direct to a story view page that chronologically lists all its events where the story name links to the corresponding Wikipedia article. Clicking "more" for each event links to the event view page that lists all the news articles of the event. At the upper right corner there is dropdown menu which allow users to set the ranking functions and refine metrics.

6 Experiments

We evaluate the event clustering and chaining in an experimental dataset constructed using Wikipedia current events from 01/01/2013 to 01/31/2015, which contains 652 stories covering 9004 events with 8,944 news articles.

We first explore whether our event clustering can effectively and efficiently cluster news articles of the same event. To construct the dataset, we select the events that link to more than 4 news articles, which in total gives us 55 events from 229 news articles. We then compare our method with the state-of-art clustering algorithms including K-means (Hartigan and Wong, 1979) and DBSCAN (Ester et al., 1996), and the state-of-art LSH methods including min-Hashing (Broder, 1997) and sim-Hashing (Charikar, 2002). We use the cluster module provided by *sklearn*[6]. For both K-means and DBSCAN, we use TFIDF based Euclidean distance in bag-of-word space. For K-means, we set the number of clusters to 55. For

[6]http://scikit-learn.org/stable/modules/clustering.html

Methods	Precision	Recall	F1
K-means	76.2%	73.1%	74.6%
DBSCAN	77.9%	74.6%	76.2%
Min-Hashing	**82.1%**	51.2%	63.1%
Sim-Hashing	80.1%	50.2%	61.7%
Event-Hashing	79.6%	**76.8%**	**78.2%**

Table 2: Event clustering comparisons

Methods	Avg. Accuracy
Cosine	66.7%
Connecting-Dots Coherence	45.2%
KL Coherence	43.3%
Learning based Model	**71.5%**

Table 3: Comparisons of event chaining

DBSCAN, we set the neighborhood size (the minimum number of points required to form a dense region) as 1. Both min-Hashing and sim-Hashing generate an 84 bits binary code to represent an event. We set θ as 5.

Table 2 shows the average precision, recall, and F1 scores over all clusters. Our method (Event-Hashing) outperforms both distance-based and LSH based clustering algorithms in terms of effectiveness, suggesting that our event representation using entities, actors, and actions is a more promising approach than bag-of-word ones. Our method is somewhat slower than min-Hashing and sim-Hashing because of the extra computing on the event space. It is worth noting that min-Hashing and sim-Hashing have higher precisions than ours, but at the cost of a big loss in recall.

We then evaluate the effectiveness of the event chaining for constructing storylines. We use the 458 stories starting in range [01/01/2013, 02/28/2014] for training and the other 194 stories for testing. We define *accuracy* of a constructed storyline as the fraction of the correctly linked events. For testing, each story is initialized by its first event. Thresholds of the three baseline measures are tuned in the training set. As shown by Table 3, our learning based model combining the three features significantly outperforms the baselines in average accuracy over the testing stories.

A small scale evaluation on the effectiveness and efficiency of the news event search engine is also performed. First, we evaluate the ranking performance for different ranking functions on a test query set including 10 different queries using precision at k (P@k). The query set contains "United States", "Russia", "China", "Barack Obama", "European Union", "President of the United States", "Car bomb", "North Korea", "South Korea", "President of Russia". We choose these queries because they appear frequently in the news articles and are very likely to be searched by users. Table 4 shows the performance of three ranking functions. The P@k scores for BM25 and cosine similarity is higher than inner product. This happens because the inner product does not do normalization thus favors the longer documents which should be less relevant in our setting.

Method	P@3	P@5	P@10	AvgTimePerQuery
Inn. Pro.	57	66	69	133ms
BM25	100	94	92	104ms
Cosine	100	94	96	136ms

Table 4: Performance comparisons of ranking methods on event search

7 Related Work

Little work has been reported on the building of event knowledge bases with the exception of EVIN (Kuzey and Weikum, 2014). However, their main focus is on extracting named events from news articles in an offline setting for knowledge base population (Ji and Grishman, 2011), but not building storylines for new events from large scale daily news streams.

Topic detection and tracking (TDT) that addresses event-based organization of news has been widely studied (Yang et al., 1999; Allan, 2002; Petrović et al., 2012). Furthermore, there is a rich literature on bursty event detection (Kleinberg, 2002; Fung et al., 2005; He et al., 2007), where an "event" is a set of word features that co-occur in certain time windows in text streams. There is also an emerging interest in building news timelines (Li and Li, 2013; Yan et al., 2011), event chains (Chambers and Jurafsky, 2008; Shahaf and Guestrin, 2010; Tannier and Moriceau, 2013), or topic model based storylines (Ahmed et al., 2011). It is worth noting that some work uses similar event encoding based on actors and actions for political events (O'Connor et al., 2013). Our work is different from existing work in both the representation of an "event" and event detection techniques. We use a three-layer (story-event-article) representation to organize the storylines and develop efficient clustering and chaining methods on the event space.

8 Conclusion and Future Work

We presented Storybase, an event knowledge base for news stories containing rich temporal and semantic information and described a storyline based news event search engine. Experimental results demonstrated that our proposed methods are effective and efficient for event detection and storyline based search. Future work could include enriching properties of a story using Wikipedia infobox and better summarizing events and stories.

9 Acknowledgements

We acknowledge partial support from Raytheon and the National Science Foundation, useful discussions with B.J. Simpson, Robert Cole, Philip A. Schrodt, and Muhammed Y. Idris, and technical support from Jian Wu, Kyle Williams, and the CiteseerX team.

References

Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J Smola, and Choon Hui Teo. 2011. Unified analysis of streaming news. In *WWW*, pages 267–276.

James Allan. 2002. Introduction to topic detection and tracking. In James Allan, editor, *Topic Detection and Tracking*, volume 12 of *The Information Retrieval Series*, pages 1–16.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997.*, pages 21–29. IEEE.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, pages 789–797.

Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *VLDB*, pages 181–192.

J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108.

Qi He, Kuiyu Chang, and Ee-Peng Lim. 2007. Analyzing feature trajectories for event detection. In *SIGIR*, pages 207–214.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *ACL*, pages 1148–1158.

Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *KDD*, pages 91–101.

Erdal Kuzey and Gerhard Weikum. 2014. Evin: building a knowledge base of events. In *WWW companion*, pages 103–106.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *Paper presented at the ISA Annual Convention*, volume 2, page 4.

Jiwei Li and Sujian Li. 2013. Evolutionary hierarchical dirichlet process for timeline summarization. In *ACL*, pages 556–560.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL*, pages 55–60.

Brendan O'Connor, Brandon M Stewart, and Noah A Smith. 2013. Learning to extract international relations from political context. In *ACL (1)*, pages 1094–1104.

Loc Paulev, Herv Jgou, and Laurent Amsaleg. 2010. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters*, 31(11):1348 – 1358.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2012. Using paraphrases for improving first story detection in news and twitter. In *NAACL*, pages 338–346.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *KDD*, pages 623–632.

Xavier Tannier and Véronique Moriceau. 2013. Building event threads out of multiple news articles. In *EMNLP*, pages 958–967.

Benjamin Van Durme and Ashwin Lall. 2010. Online generation of locality sensitive hash signatures. In *ACL*, pages 231–235.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline generation through evolutionary trans-temporal summarization. In *EMNLP*, pages 433–443.

Yiming Yang, Jaime G Carbonell, Ralf D Brown, Thomas Pierce, Brian T Archibald, and Xin Liu. 1999. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43.