VLX-Stories: A Semantically Linked Event Platform for Media Publishers

Dèlia Fernàndez-Cañellas[1,2], Joan Espadaler[1], Blai Garolera[1], David Rodriguez[1], Gemma Canet[1], Aleix Colom[1], Joan Marco Rimmek[1], Xavier Giro-i-Nieto[2], Elisenda Bou[1], and Juan Carlos Riveiro[1]

[1] Vilynx, Inc.
[2] Universitat Politecnica de Catalunya (UPC)

Abstract. In this article we present a web platform used by media producers to monitor world events, detected by VLX-Stories. The event detector system retrieves multi-regional articles from news sites, aggregates them by topic, and summarizes them by disambiguating and structuring their most relevant entities in order to answer the journalism W's: who, what, when and where. These events populate VLX-Stories -an event ontology- transforming unstructured text data to a structured knowledge base representation. The dashboard displays online detected events in a semantically linked space which allows navigation among trending news stories on distinct countries, categories and time. Moreover, detected events are linked to costumer contents, helping editorial process by providing real-time access to breaking news related to their contents.

Keywords: Knowledge Graph · Event Representation · Linked Data.

1 Introduction

Media publishers such as news broadcasters, magazines, media companies and bloggers have the need to monitor world events in order to be aware of trends and events worldwide. News aggregators have provided a solution to navigate and consume news by grouping the overwhelming amount of articles published on event clusters. However, these tools are designed for general users and do not provide several crucial capabilities for media publishers: a long term view and context on the news stories, multi-regional and multi-lingual information or linkage to their contents or those of their competitors. Semantic Web and Linked Data technologies provide solutions which can be applied to the mentioned problems by using Knowledge Graph (KG) and Ontologies to link, structure and serve this information. Many works have already applied these semantic solutions to monitor and structure news contents, e.g. GDELT [3], IJS newsfeed[4], iDiversiNews[5] and AFP4W[6]. These systems provide summarization, semantically enriched news and search engines based on the journalist W's[3]. However,
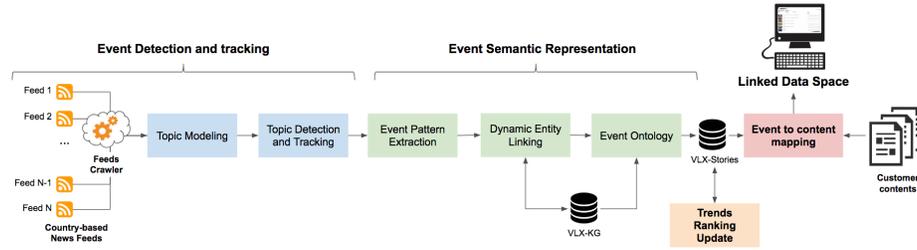
[3] https://www.gdeltproject.org
[4] http://newsfeed.ijs.si/
[5] http://ailab.ijs.si/tools/idiversinews/
[6] http://medialab.afp.com/afp4w/index.php

Fig. 1. Pipeline schema of VLX-Stories framework.

none of them provides integration in the editorial process and capabilities to link its information with other external sources provided by the costumer.

In this work we present VLX-Stories[2] interface and application in the editorial process. This framework detects multi-regional world events based on aggregated articles. Unlike previously mentioned works, VLX-Stories not only structures events, but also links it to customers contents. This is done by labeling both detected events and customers contents with multilingual KG entities. The presented demo shows the complete end-to-end process to generate the events ontology (Section 2) and displays the dashboard user interface (Section 3) to access and query VLX-Stories, currently encoding over 9000 events per month. This interface leverages semantic technologies to provide a complete linked space which allows navigation among time, categories, regions, publishers, topics, places or personalities. It is deployed in production and is being used by major media networks, accelerating the editorial process and improving their operational efficiency by helping on content discovery, search, content generation and exploring which stories will have the most impact on their audience.

2   System Overview

This section provides a general overview on VLX-Stories [2] pipeline, which is split in four major blocks as displayed in Fig.1.

The first block of the pipeline (Event Detection and Tracking) extracts news articles from media feeds and aggregates them into clusters describing the same event. Articles are collected by an RSS feeds crawler, and represented as vectors in the Topic Modeling module. Then, these articles are associated to an already detected event (Topic Tracking) or used as a seed for a new event (Topic Detection). The output of this block are clusters of aggregated news articles representing distinct world events.

The Second block (Event Semantic Representation) represents the events by synthesizing the agents, locations and actions involved in each cluster. This is achieved by extracting the entities involved in each event and structuring the knowledge in an ontology which answers the journalist W's [3] and the Topic. To do that, a keyword-based pattern is extracted for each detected event. Then, in the Dynamic Entity Linking module, mentions from the pattern are mapped to entities from an external KG, called Vilynx KG (VLX-KG) [1, 2]. This KG was initially constructed by merging different public KGs, like Freebase and Wikidata. However, as news often refer to people that have never been mentioned
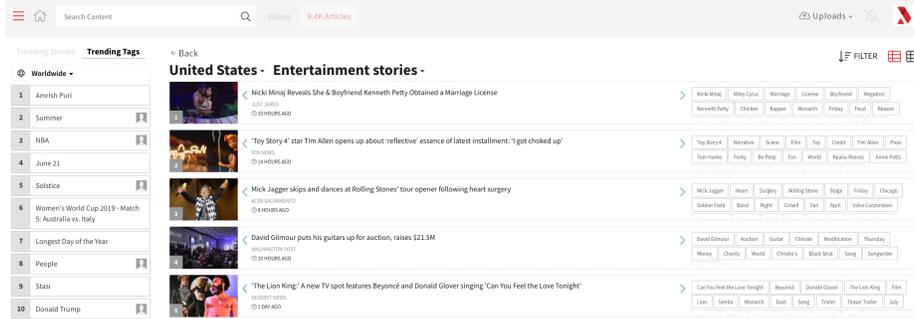
Fig. 2. Events landing page.

before, an Emerging Entity (EE) detection step has been added into VLX-Stories pipeline to extract *out-of-knowledge-base* (OOKB) entities and dynamically populate VLX-KG. Finally, entities are structured into the Event Ontology using their respective types and relevancy.

The third block involves the ranking of events by trends. This module computes a trending score for each detected event by linearly combining the trendiness of its associated entities with the number of articles discussing the same topic. Entity trendiness is computed in an external module which monitors Twitter and Google Search. To compute the article-based trendiness we model how the number of articles being published evolves over time and determine an event being trending when the number of articles associated is higher than expected.

Finally, detected events are associated to customer's contents generating a big linked data space in the Event to Content Mapping block. To do that, customer contents are tagged using the same Dynamic Entity Linking system used to extract the entities representing an event. Once the contents are tagged, the linkage is done by comparing its associated entities.

3 User Interface

Our event-navigation landing page is captured in Fig.2. It displays stories on the news with a list of the events detected for each county and category, ranked by trendiness. The country and category can be chosen through drop-down menus. Current categories include: *top stories*, *latest stories*, *politics*, *sports*, *entertainment*, *general news*, *business and finance*, *science and technology*, and *lifestyle and hobbies*. The menu also provides the following filtering capabilities: a) sort events by date or trending score; b) display events filtering by source: any source or only publisher related contents; and c) temporal navigation by date range.

The list of detected events is displayed behind the filters menu. Clicking on a news story takes the user to the individual story page, where the full list of articles that were identified as being related to the topic is displayed. In Fig.3 we present an example of the resulting event menu, which structures the information to answer the journalist W's. In the top of the menu, the event category is displayed. The title summarizing *what* happens, and the other properties: *when*, *topic*, *where* and *who* are shown behind. Titles from articles clustered give context and additional information on the story, having linkage to the source page.
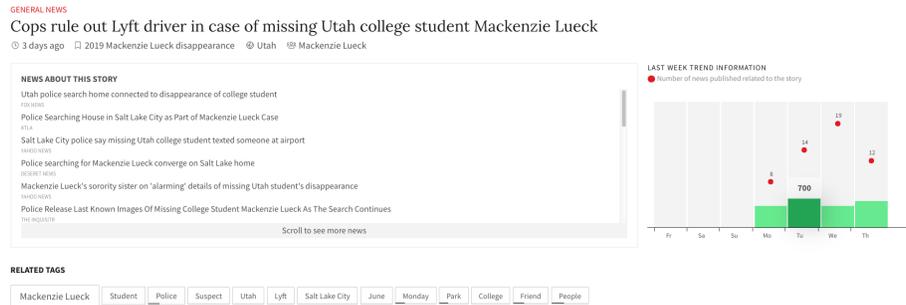
Fig. 3. Example of the event display menu. Notice that '*Mackenzie Lueck*' is an EE detected by VLX-Stories.

Social impact throughout the week is displayed in the Trending Chart. In this chart the daily number of articles clustered related to the event is shown with a red dot, and the green bars represent the trending score daily evolution. In the bottom, the entities in the event semantic pattern are displayed as related tags. Notice these entities are sorted according to their relevance describing the event and the bar behind the entity box represents the current trendiness of the entity. Lastly, when available, we aggregate all existing content about the story created by the individual publisher, as well as related content from their library. Moreover, all events and customer contents are linked through the tags. Clicking on a tag brings the user to the entity page where all news stories and contents labeled with a given entity are displayed together with a trending chart and map which displays tag trendiness on social networks. Find the link [7] to the demo video.

4   Conclusions

We have presented a dashboard which displays real-time information about semantically linked events, based on aggregated multilingual news articles and linkage to customer contents. This interface is a commercial tool used by media producers and other global media companies in the editorial process to identify which topics are gaining momentum and they should be writing about, as well as the trending stories they are already covering. Moreover, the presented system allows the visualization and navigation among countries, categories, time and related entities through a friendly and intuitive dashboard.

References

1. Fernández, Delia, e.a.: Vits: Video tagging system from massive web multimedia collections. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 337–346 (2017)
2. Fernández, Delia, e.a.: Vlx-stories: building an online event knowledge base with emerging entity detection. In: International Semantic Web Conference. Springer (2019)
3. Singer, J.B.: Five ws and an h: Digital challenges in newspaper newsrooms and boardrooms. The International Journal on Media Management (2008)

---

[7] https://drive.google.com/file/d/1yULvPs9AmJ449PwQ0A1-v5FuYGyEuMAs