



VLX-Stories: Building an Online Event Knowledge Base with Emerging Entity Detection

Dèlia Fernàndez-Cañellas^{1,2(✉)}, Joan Espadaler¹, David Rodríguez¹,
Blai Garolera¹, Gemma Canet¹, Aleix Colom¹, Joan Marco Rimmek¹,
Xavier Giro-i-Nieto², Elisenda Bou¹, and Juan Carlos Riveiro¹

¹ Vilynx, Inc., Barcelona, Spain
delia@vilynx.com

² Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

Abstract. We present an online multilingual system for event detection and comprehension from media feeds. The system retrieves information from news sites, aggregates them into events (event detection), and summarizes them by extracting semantic labels of its most relevant entities (event representation) in order to answer the journalism Ws: who, what, when and where. The generated events populate VLX-Stories -an event ontology- transforming unstructured text data to a structured knowledge base representation. Our system exploits an external entity Knowledge Graph (VKG) to help populate VLX-Stories. At the same time, this external knowledge graph can also be extended with a Dynamic Entity Linking (DEL) module, which detects emerging entities (EE) on unstructured data. The system is currently deployed in production and used by media producers in the editorial process, providing real-time access to breaking news. Each month, VLX-Stories detects over 9000 events from over 4000 news feeds from seven different countries and in three different languages. At the same time, it detects over 1300 EE per month, which populate VKG.

Keywords: Knowledge base population · Knowledge graph ·
Event encoding · Entity linking · Emerging entities · Topic detection

1 Introduction

An increasing amount of news documents are published daily on the Web to cover important world events. News aggregators like *Google News*¹ or *Yahoo! News*² help users navigate by grouping this overwhelming amount of materials in event clusters. Such systems facilitate users to stay informed on current events and allow them to follow a news story as it evolves over time. This clustering

¹ <http://news.google.com>.

² <http://news.yahoo.com>.

task falls on the field of Topic Detection and Tracking (TDT), which aims to develop technologies that organize and structure news materials from a variety of broadcast news media. However, media professionals are in need of more advanced tools to describe, navigate and search specific pieces of information before writing their own piece of news. Semantic Web and Information Extraction (IE) technologies provide high level structured representations of information, which can help solving the mentioned problems. *Knowledge Graphs* (KGs), which store general knowledge represented by world entities and their relations, are currently seen as one of the most essential components of semantic technologies. They allow to generate linked data spaces and structuring information by linking entity mentions to KG entities. The most popular ones are Freebase [4], DBpedia [3], YAGO [36] or Wikidata [39]. Nevertheless, most of these existing KGs focus on traditionally encyclopedic facts like names of popular people, their birth date and place, job, etc. Dynamic information, such as events reported in the news, often involve short term relations and unknown people that are not captured by these resources, and are therefore missed by most KGs. Detecting these out-of-knowledge-base (OOKB) events and its related Emerging Entities (EE) is crucial for any KG maintenance process [14, 25, 30]. In particular, when willing to provide efficient tools for news description, search and analysis.

In this work, we describe VLX-Stories, a system under exploitation that alleviates the aforementioned issues from journalists teams. It consists of a unified online workflow of event detection, tracking, pattern extraction and Dynamic Entity Linking (DEL), with the aim of building an event-based KB. At the same time, the new EEs detected by VLX-Stories populate an external KG, called *Vilynx Knowledge Graph* (VKG), with background encyclopedic knowledge. In VLX-Stories, events are represented by means of an ontology inspired on the journalist Ws [33]: *what* is happening, *who* is involved, *where* and *when* it took place, and the general *topic* under discussion. The system is characterized by the adoption of semantic technologies, combined with Information Extraction techniques for event encoding. The extraction of mentions and its linkage to entities from an external multilingual KG generates an event linked space. This allows the multilingual linkage across stories, semantic search, and the linkage to customer contents by matching entities.

The goals and contributions of this work are: (a) the generation of an event KB from news stories, (b) the detection of EE from them and (c) the large-scale deployment of the system, which is currently consumed by several media companies to gather information more efficiently. To the best of our knowledge, this is the first system that uses the redundancy of aggregated news articles for a robust detection of EE, in an online manner. Performance statistics are given and evaluation is carried out on the two main contributions addressed (event KB construction and EE detection).

This work is structured as follows. Section 2 describes the industrial context in which this system was built. Section 3 presents the techniques used to detect events on news articles, and Sect. 4 describes how we use pattern mining and entity linking techniques to structure event information. Section 5 includes

performance statistics, and Sect. 6 demonstrates the quality of our system by evaluating the distinct modules of our pipeline. Finally, Sect. 7 presents the related work and Sect. 8 the final conclusions and future work.

POLITICS

Brexit: Donald Trump warns Theresa May EU deal would threaten trade

🕒 11/27/18 📁 Brexit 🌐 United Kingdom, United States 🗣️ Donald Trump, Theresa May

NEWS ABOUT THIS STORY

Trump says Brexit deal hampers U.S.-UK trade
THE AUSTRALIAN


May rebukes Trump as she bids to sell Brexit deal
SBS NEWS

May rebuts Trump's Brexit trade comments
HERALD SUN

Brexit: Donald Trump warns Theresa May EU deal would threaten trade
THE WEEKLY TIMES

Trump torpedoes close ally with new outburst
NEWS.COM.AU

Trump warns Brexit will harm UK-US trade
SBS NEWS



RELATED TAGS

Donald Trump

Brexit

Deal

Theresa May

United Kingdom

Trade

President

Future

European Union

United States

White House

Washington, D.C.

Fig. 1. Example of the resulting event information displayed on Vilyn Dashboard. In the top we display the article category, the title summarizing *what* happens, and the other properties: *when*, *topic*, *where* and *who*. Titles from articles clustered give context and additional information on the story. At the bottom the entities in the event semantic pattern are displayed as related tags and sorted according to their relevance describing the event.

2 Industrial Use of VLX-Stories

VLX-Stories is a product developed and commercialized by Vilyn³, an AI company specialized in analyzing media contents. The system is deployed in production worldwide and is currently used by US networks and expanding to Europe and South America. Journalists and editorial teams consume the rich structured news data offered by VLX-Stories to explore how a story relates to other news and detect about which topics they should be writing. This information is served through API calls and a dashboard, providing a general view of what is happening in the world. In Fig. 1, we present an example of a detected event displayed on our dashboard. Notice how the different Ws are addressed and additional context on the news story is given through clustered articles and the related tags (entities). Customers' contents are also linked to detected events using entities, complementing their content information. Thanks to the entities linkage, it is possible to offer a practical interface for navigation and exploring

³ <https://www.vilyn.com>.

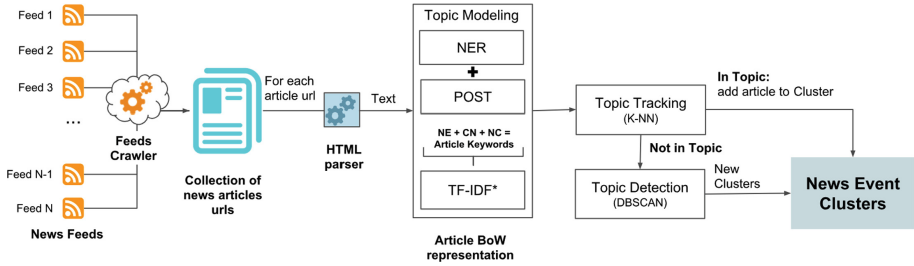


Fig. 2. Schema of the news event detection pipeline.

news. Moreover, VLX-Stories extracts temporal relations and information on temporal trends, which are internally used for other products, e.g. recommendations, trends detection and disambiguation.

Apart from the customer services which can be offered through VLX-Stories, the contributions from this work are also an essential internal tool in Vilynx. The core of Vilynx’s technology is the *Vilynx Knowledge Graph (VKG)* [8], which is used as a semantic base for indexing media documents. This KG is constructed by merging different public knowledge resources: Freebase, Wikidata and Wikipedia. It provides multilingual aliases for over 3M entities, and 5M relations between entities based on properties. VKG is required to be dynamic because it must support real-time indexation of media documents. Is thus in need of an online detection and population of EE. To provide new knowledge to VKG we use structured data, which is updated periodically by querying the three mentioned public KBs. However, media news often talk about unknown people, who are not indexed on these public knowledge resources or that have not yet been updated [30]. Indexing these novel entities requires extracting EE from non-structured data, e.g. news articles. The VLX-Stories system, presented in this article, will provide the information and dynamics required for VKG maintenance with OOKB entities, while detecting events.

3 Event Detection

This section describes the three parts of the News Event Detection pipeline, outlined in Fig. 2. First, a collection of news articles is built by crawling RSS feeds and parsed (Sect. 3.1). Afterwards, in the topic modeling block (Sect. 3.2), articles are represented with bag-of-words (BoW). Finally, in the topic tracking module, each article vector is compared with articles grouped in event clusters: if matching the event, it is assigned to the cluster; if not, it is added to a pool of articles that will be processed in the topic detection module (Sect. 3.3) in order to detect newly emerging events.

3.1 News Feeds Crawler

News articles are collected by an RSS feeds crawler, which processes 1500 news feeds every 30 min. The RSS feeds come from a manually generated list of 4162 feeds from the main media sources of seven countries: United States, Australia, Spain, Canada, United Kingdom, Portugal and Ireland. Feeds are also manually categorized in seven category groups: *politics*, *sports*, *general news*, *lifestyle and hobbies*, *science and technology*, *business and finance*, and *entertainment*. The feeds crawler visits each feed, crawls it, and stores in the DB each article URL, publication date, title and description if provided. In a second step, whenever a new article is detected in a feed, we crawl the URL and parse the article using a customized HTML parser to extract all its text data and images.

3.2 Topic Modeling

Topic modeling (TM) consists of representing the abstract matter that occurs in a collection of documents. To do this, we will rely on a BoW representation of the articles. As news stories typically revolve around people, places and other named entities (NE), some works [12, 32] use mentions of NE instead of all words in the document. However, some news do not turn around NE, e.g. weather news or events related to anonymous people. Therefore, other information, such as common nouns (CN) or noun chunks (NC), is needed to distinguish this kind of events [27]. Combining these three works, we will use named entities, common nouns and noun chunks in the BoW representation, instead of all words in the text corpus. We will call this collection of mentions and nouns *article keywords*. These keywords are extracted from the article's text by a Named Entity Recognition (NER) module and Part of Speech Tagger (POST). We use the Spacy's⁴ library and multilingual models for these tasks. For performance reasons, we constraint the articles to be represented for at least 8 keywords, and a maximum of 80 keywords.

BoW keyword's frequencies are weighted by a slightly modified TF-IDF (term frequency - inverse document frequency), which reflects how important a word is to a document in a collection or corpus. TF-IDF is computed as the product of the term frequency (f_k) by the inverse document frequency (idf_k). However, we bias the TF with a weight to give more relevance to those keywords appearing on the title (α), description (α), or that are NE (β). Finally, inspired by [12], we apply a time factor with a linear function, which favors news documents to be assigned to more recent events.

3.3 Topic Detection and Tracking

Once a new article is ingested by the system, we must detect if it is associated to an already detected event (topic tracking) or it describes a new event (topic detection). For the topic tracking, we will use the *k-Nearest Neighbours* (k-NN)

⁴ <https://spacy.io/>.

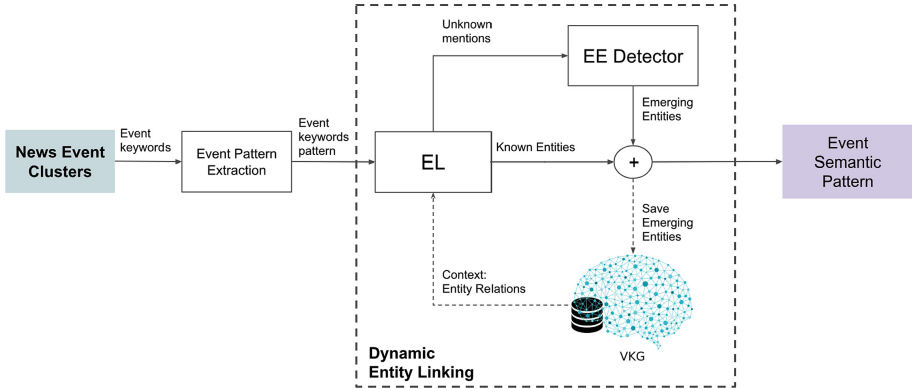


Fig. 3. Pipeline schema of the Event Semantic Pattern extraction module, composed by the Event Pattern Extraction and the Dynamic Entity Linking modules.

algorithm. Thus, a_i being a new article, we will associate the article with an event cluster if there are more than k articles in the cluster with a similarity higher than a given threshold γ . If the incoming article is not associated to any event cluster, we will try to build a new cluster with other articles not yet related to any event. This is the task of topic detection. The chosen clustering technique for topic detection is DBSCAN [7], which is an unsupervised density-based algorithm that provides robustness against the presence of noise. This method requires the estimation of two hyper-parameters: *min samples*, which is the minimum number of samples needed to generate a new cluster, and *eps*, the maximum distance allowed within its samples. We decided to fix the *minsamples* = 5, thus all events will be represented with at least five articles, and we optimize *eps* in order to have high precision without missing many events. We use *cosine similarity* as the distance metric for both tasks.

Moreover, some design decisions were made in order to compensate some of the problems of dealing with an online and large-scale deployment application with noisy Web data. In order to prevent wrong event detections due to web parser errors, we added two extra conditions on the cluster generation: the clustered articles need to be from at least three different news publishers, and one media publisher can not own over 50% of the articles in a cluster. Values were chose after manually analyzing several detection errors. Also, speed issues had to be considered to provide real-time trackin on news events, as the amount of comparisons between articles grows quadratically with the number of articles, slowing the whole article comparison. We decided to cluster articles by country, and for those countries with more feeds, we use a category-based comparison between articles. The category of the feed is used for this split, and in case the feed provides general news from any category, we trained a deep classifier based on a one layer LSTM [13] to predict the article category from its title. The training dataset was constructed by merging the category titles from the

UCI-ML News Aggregator Dataset [6] and titles from the manually labeled RSS news feeds.

4 Event Representation

Event representation tries to synthesize the agents, locations and actions involved in an event in a formal machine understandable way, but still natural for humans. This is achieved by extracting semantics from the event articles and structuring the knowledge. Our approach provides an event semantic pattern by combining pattern mining with a Dynamic Entity Linking module (Sect. 4.1). This module uses VKG for the entity disambiguation, which will also be populated with EE found in the process. Finally, this semantic pattern is post-processed to structure the entities into our event ontology (Sect. 4.2).

4.1 Event Semantic Pattern

The extraction of the event semantic pattern is achieved thanks to the two modules depicted in Fig. 3: Event Pattern Extraction (EPE) and Dynamic Entity Linking (DEL). The EPE module finds the keywords pattern describing the event, and the DEL links these keywords to entities from VKG, while detecting candidates of new entities which populate VKG. The details follow.

Pattern Mining: Data mining techniques search for patterns in data that are representative and discriminative. We define our pattern mining task with an *association rule* approach [2], such that our pattern corresponds to a set of association rules, $t^* \rightarrow y$, that optimize the *support* and *confidence* constraints for each event. Let n be the set of all keywords in the corpus $C = \{k_1, k_2, \dots, k_n\}$; and a *transaction* A be the set of keywords from a given article, such that $A \subseteq C$. Given a set of m transactions belonging to the same event $T = \{A_1, A_2, \dots, A_m\}$, we want to find the subset of C , say t^* , which can accurately predict the belonging to a target event $y \in E$. The *support* of t^* is an indicator of how often t^* appears in T , and it is defined as the proportion of transactions in the transaction set T that contain the itemset t^* :

$$s(t^*) = \frac{|\{A_a | t^* \subseteq A_a, A_a \in T\}|}{m} \quad (1)$$

Our goal is to find association rules that accurately predict the belonging to an event, given a set of keywords. Therefore, we want to find a pattern such that if t^* appears in a transaction, there is a high likelihood that y , which represents an event category, appears in that transaction as well. We define the *confidence* as the likelihood that if $t^* \subseteq A$ then $y \in A$, which can be expressed as:

$$c(t^* \rightarrow y) = \frac{s(t^* \cup y)}{s(t^*)} \quad (2)$$

Inspired by [22] we use the popular apriori algorithm [1] to find patterns within the transactions. We only keep the association rules with confidence $c_{min} \geq 0.8$ and calculate the support threshold (s_{min}) that ensures at least 10 keywords in the rule. Finally, we select the rule t^* with more keywords associated. This keywords are the ones that will be linked to VKG entities.

Dynamic Entity Linking: The event keywords in the pattern will be mapped to entities in VKG. This task is called Entity Linking (EL) or disambiguation. Our EL module gets entity candidates from VKG for each incoming mention. Entities are retrieved based on similarity matching between the text mention and the entities alias. Then, disambiguation is applied by scoring each candidate. Following the work in [8], an *intra-score* and an *inter-score* are computed for each candidate. On one hand the intra-score is computed by using the information of the mention and entity itself, combining the following metrics: word similarity, entity usability and entity type. On the other hand, the inter-score exploits the contextual information between entities by using distances between entities in a concept embedding space. The combination of all these metrics gives a confidence score for a given mention to be disambiguated to an entity. If this score is higher than a predefined threshold, the mention and entities are linked.

However, news often refer to people that has never been mentioned before, and thus, are not indexed in VKG. In order to populate VKG we added dynamics into the EL module, calling it a Dynamic Entity Linking (DEL). This module maintains EE that refer to unknown people as they appear on the news, and integrates it into VKG. This EE detector filters the unknown mentions, keeping only those that have been recognized as *persons* by the NER module and that are at least composed by two words (name and surname). The detection is highly robust because the EE come from the previously extracted event pattern, which means the entity has appeared in a high amount of articles from different publishers, in the same context, and is thus relevant when describing the event. Once an EE is detected, the system starts being capable of using it for tagging and linking documents, and it is already used to describe new events. However, it will require for a human validation in order to become a proper concept in the KG. This validation is needed because sometimes the names detected are spelling variations of entities already in the KG, or mistakes from the NER module. An independent system takes care of the EEs by searching for entity matching suggestions in external KBs (Google Knowledge Graph, Wikipedia and Wikidata), as well as entities in VKG. Suggestion results are displayed in an internal dashboard, together with context from the sentences where the EE has been seen, where a human makes the final decision. Thanks to previous process and auto-complete tools, the human intervention is minimal and very fast decisions can be made.

Multi-regional Event Matching: Before the final event modeling, the semantic pattern of the events detected for each country are compared and merged in case of match. To do that, we first rank the entities in the Event Semantic

Pattern by relevancy describing the event. The ranking is based on re-scoring entities based on its original keywords appearance frequency and origin (title, description or text body). As we solved the entity disambiguation we recompute the entity frequency taking into account co-references. Origins are taken into account by weighting the frequency of appearance by the origin. Afterwards, country-events are represented with a bag of concepts BoC where entity relevancies are the weights. Cosine similarity is computed between country-events and these events are merged into worldwide-events if its similarity is higher than a manually defined threshold.

4.2 Event Model

Both semantic and contextual event information extracted on previous steps are processed in order to represent the collected data in an ontological manner. This ontological information is stored in VLX-Stories KB, which keeps growing with the multiregional news information provided by the feeds. In this section, we first motivate the modeling decisions we took designing the ontology and we continue by describing the information extraction process.

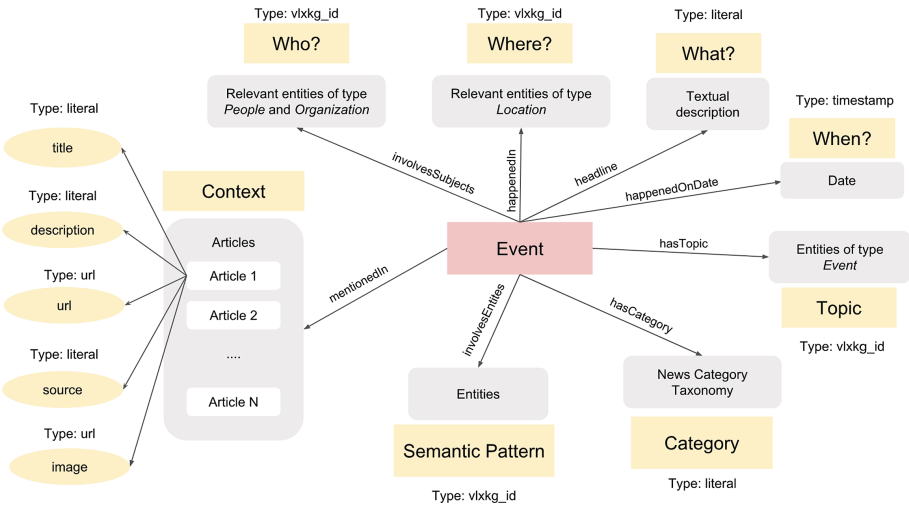


Fig. 4. Event ontology schema.

Modeling the Event Ontology: The main requirement of our event ontology is that it has to synthesize in a both machine and human readable way the unstructured semantic pattern extracted. Events are usually defined by its agents, locations and actions occurring and the moment when the event takes

place. Journalistic articles are typically structured to answer four of the journalist 5W-questions: e.g. *what* happened, *who* is involved, *where* and *when* did it happen. These questions should be addressed within the first few sentences of an article to quickly inform readers. The fifth W, *Why* it happened, is often addressed in opinion columns or post-event coverage [16]. Moreover, news stories often fall into bigger topics which are composed by several events, like *Brexit*, *Academy Awards*, *Olympic Games*, etc. This information, if present, offers the possibility of tracking long story lines and to provide a complete context on the development of an event in a point in time.

Considering the above mentioned 4Ws and the topic, we defined our ontology with the next *properties* or *core classes* for each event: Who, What, When, Where and Topic. This properties will be extracted from the event semantic pattern and the titles and descriptions from the event articles. Moreover, as shown in the ontology schema in Fig. 4, all entities in the semantic pattern, answering or not the 4Ws or topic, will be included in the ontology within the Semantic Pattern class. The event Category, e.g. sports, politics, entertainment, etc.; is also included as a property. Additional context on the event is given by the clustered articles, from which we store the title, description, URL, news source and image, if present.

Event Properties Extraction: The Who, Where and Topic are extracted from the Event Semantic Pattern using a set of filters based on entity types. Entities in VKG have a schema⁵ type associated, which denotes if the entity is a person, place, organization, etc. These types will be used to classify the entities in the pattern together with the type of metadata field from which they were extracted. For this task only the entities from the title and description are used. Moreover, the same entity relevance scores computed for multi-regional event matching will be used to pick those most relevant entities. The Who property needs to be an entity of type *Person*, *Organization* or *CreativeWork*; the Where is a property of type *Place* and the topic of type *Event*. We define the What of an event with a sentence (literal) summarizing the main event occurring. As news article’s titles should give this information, we answer the What with the most descriptive title between the clustered articles. This is selected using a voting algorithm where each word in the title sums its given score in the semantic entities ranking. This approach favors longer titles, which also tend to be semantically richer. To answer When, we take the date of the first article published related to an event. We plan on improving it on next versions by analyzing time expressions in the text. Finally, we complete the ontology by adding the event Category. Categories come from our pre-defined set of categories, e.g. *Sport*, *Entertainment*, *Finances*, *Politics*, etc. The categories assigned to the RSS feeds are used to extract this information. One event may belong to more than one category.

⁵ <https://schema.org/>.

5 System Analytics

VLX-Stories was first deployed on July 2018 for its use in the United States. Since then, the system has been growing, adding new world regions and languages. Table 1 contains the activation date of VLX-Stories for each country, the language it processes, the number of feeds activated on each country, the average number of events detected each day and the daily number of articles processed and associated to events. Results are provided country by country and also on the worldwide event aggregation. According to these statistics, VLX-Stories grows in a speed average above 300 news events/day, classifying an average of over 17k multilingual articles from seven different countries. Since we activated the multi-regional event aggregation module on November 2018, the system includes the option of analyzing how an event is reported in different world regions and in different languages. Semantic disambiguation is essential for this multilingual aggregation task.

Table 1. Statistics on VLX-stories population.

Country	Activation date	Language	#Feeds	Events/Day	Articles/day
USA	07/2018	English	952	96.70	4,745.59
SP	08/2018	Spanish	918	90.61	4,929.01
CA	09/2018	English	551	27.15	990.04
AU	09/2018	English	893	53.19	3,223.09
IR	09/2018	English	121	20.46	654.25
UK	09/2018	English	442	38.57	1,518.63
PT	09/2018	Portuguese	285	35.80	1,235.76
Total	-	-	4,162	362.48	17,296.37
World Wide	11/2018	Multilingual	4,162	301.84	17,296.37

6 Evaluation

VLX-Stories was evaluated by considering each block separately. This approach provides a detailed analysis and facilitates the identification of bottlenecks.

6.1 News Event Detection

Regarding news aggregator or event detection evaluation, we used a subset of the UCI-ML News Aggregator Dataset [6]. This dataset consists of 7,231 news stories, constructed from the aggregation of news web pages collected from March 10 to August 10 of 2014. The resources are grouped into clusters that represent pages discussing the same news story or event. Events are categorized into four

different categories: entertainment, health, business and technology. For each news article, its title, URL, news story id and category are given. However, we had to discard a 13% of the events on the dataset because of the following reasons: 36% of the URL link's were broken, our system could not extract enough keywords from 17% of the articles, and finally, we had to discard some of the remaining news stories because they were not represented by enough articles to be detected by our system (each of our events needs to be represented by at least 5 news articles). The final dataset subset consists on 6,289 events constructed by 91,728 news articles. For our experiments the DBSCAN parameters were set to $eps = 0.65$ and $minsamples = 5$. Table 2 presents the news event detection results on the dataset. Most of the events are correctly detected (81.58%), however a lot of articles are not associated to any event. This is reflected by the high average precision (94.57%) but a poor recall (58.54%). This is mostly because of the restrictive parameters set in the system in order to make sure that aggregated news served are correct. Quality of aggregated news is similar across news categories. The lowest quality is found in the business category, because most of the business news share financial terms which are repeated in many articles, even not related ones. Best results are for entertainment, the type of news with more named entities, which improves the representation.

Table 2. Results of the news event detection on UCI Dataset subset

	#articles	#events	Event P.	Article P.	Article R.	F1
Business	21,535	1,796	78.28%	91.44%	51.76%	66.11%
Entertainment	36,790	1,673	88.76%	96.43%	64.44%	77.26%
Technology	23,921	1,811	85.69%	94.38%	57.37%	71.36%
Health	9,482	1,009	68.18%	94.86%	56.85%	71.09%
GLOBAL	91,72	6,28	81.58%	94.57%	58.54%	72.32%

6.2 Dynamic Entity Linking

The mapping of event keyword patterns to event semantic patterns includes the EL task and the EE detection. However, this tasks also depends on the quality of the NER and POS modules from Spacy. According to Spacy's documentation⁶ the quality of these modules is state of the art performance, with accuracies close to the 90%. To evaluate the quality on the EL and EE detection we ran two experiments.

Entity Linking: The first experiment evaluated EL of the keywords on the event pattern to semantic labels. The experiments were conducted over a corpus of 100 semantic event patterns randomly selected from the United States

⁶ <https://spacy.io/usage/facts-figures>.

events, detected by our news aggregator module during the week from the 1st of January to the 7th of January 2019. The keywords from the patterns were mapped to entities from VKG using the dynamic EL module. The correctness of the mapping was evaluated with *TP* when the semantic entity related to the keyword was correct, *FP* when the semantic entity related was wrong, *TN* when no entity was related and it is not an existing entity or it is an error from NER, and *FN* if no entity was mapped but there is an entity in VKG for it. Results are shown in Table 3, with a total accuracy of the 86%. However some mentions do not disambiguate to its correct entities. This is specially common when finding homonym words or unknown contexts. Further research should be developed to improve these ambiguous cases.

Table 3. Results on entity linking

#Event patterns	TP	TN	FP	FN	Precision	Recall	F1	Accuracy
100	966	329	52	156	0.86	0.94	0.90	0.86

Emerging Entities: The capacity of detecting EE was evaluated by deleting existing entities from our VKG and testing the capacity of the system to create them again. We initially built a dataset of 648 news events detected during the week from the 1st to the 7th of January 2019. The multilingual capabilities of the system was tested by choosing events from three regions with different languages: the United States, Spain and Portugal. The dataset was generated by running the Event Semantic Pattern module, removing the corresponding person’s entities from VKG, and extracting again the Event Semantic Pattern, expecting for the EE detector to re-generate the deleted entities. As shown in Table 4, an average of 78.86% of the deleted entities were recovered. Some of the missing entities were composed by just one word, like *Rhianna* or *Shakira*. Our system did not detect them because it constrains person entities to be described with two words (name and surname). Other errors were caused by the similarity between entities, which are wrongly disambiguate to existing entities; e.g. when deleting the *Donald Trump* entity, the EL disambiguated to *Donald Trump Jr.* because of a perfect match between the alias and the similar usage context.

Finally, a statistical study of the created entities and their quality is done by analyzing data between 12th December 2018 and 15th March 2019. Table 5 presents the average number of EE detected every day in each language. However, not all the detected EE become new entities in VKG. After the human supervision we extracted the next metrics: 75.45% of the detected EE become new entities, 22.15% are alias of already existing entities and 9.7% are wrong candidates because of NER errors.

Table 4. Results on emerging entities detection

Country	Language	#Stories	#Deleted entities	%EE Recovered
United States	en	282	373	80.16%
Spain	es	251	299	74.91%
Portugal	pt	115	104	85.57%
Total	-	648	776	78.86%

Table 5. Statistics on emerging entities detection by VLX-stories

	EN	ES	PT	Total
Avg. EE detected/day	41.18	20.08	9.27	67.88

7 Related Work

The system presented in this work tries to solve the *semantic gap* between the coverage of structured and unstructured data available on the Web [28], in order to provide journalistic tools for event analyzing. In the past decades, a great amount of research efforts has been devoted to text understanding and Information Extraction (IE). Many research projects have entangled with the different problems described in this work, i.e. news aggregation [5, 9, 12, 20, 35], event pattern extraction [15, 43], entity linking [12, 19, 20, 23], emerging entity detection [14, 17, 25, 30, 34], event ontology population [29, 42] and automatically answering journalist Ws [10, 11]. However, only a few big projects are comparable to our system as end-to-end online pipelines for event detection and encoding. In this section we will focus on reviewing these large-scale systems.

Two well-known event-encoding systems are the *Integrated Crisis Early Warning Systems*⁷ (ICEWS) and the *Global Database of Events, Language and Tone*⁸ (GDELT). This two projects have been developed to automatically extract international political incidents such as protests, assaults and mass violence from news media. These datasets are updated online, making them useful for real-time conflict analysis. ICEWS is a project supported by the Defense Advanced Research Projects Agency (DARPA), to be used for US analyst. Its data has recently been made public through Harvard’s Dataverse⁹, however events are posted with a 1 year delay and the techniques and code utilized are not open source. GDELT was build as a public and more transparent version of ICEWS. Its data is freely available and includes over 200 million events since 1979, with daily updates. However, legal controversies over how data resources were obtained distanced it from research. It is currently incorporated into Google’s services and its data is utilized for analysis of international events [18, 21]. As the two more spread news databases, several comparison studies have been made

⁷ <https://www.icews.com/>.

⁸ <https://www.gdeltproject.org/>.

⁹ <https://dataverse.harvard.edu/dataverse/icews>.

between ICEWS and GDELT. Even though no conclusion could be extracted on the superiority of any system, GDELT overstates the number of events by a substantial margin, but ICEWS misses some events as well [40,41].

A more recent event data program is the *Open Event Data Alliance*¹⁰ (OEDA). This organization provides public multi-sourced political event datasets, which are weekly updated [31]. All the data is transparent and they provide open code of the ontologies supported. They use Stanford CoreNLP tools [24] and WordNet [26] dictionaries. However, OEDA's efforts still have not reached the scale of the other two mentioned projects.

Another well-known project is the *NewsReader*¹¹ [38]. This system is a big collaborative research project, which constructs an Event-Centric Knowledge Base (ECKB) based on financial and economic news articles. They take advantage of several public knowledge resources to provide multilingual understanding and use DBpedia [3] as KG for EL. They define their own event ontology, the Simple Event Model (SEM) [37], which is designed to be versatile in different event domains allowing cross-source interoperability. To deal with entities not properly represented in the knowledge resources, they introduce the concept of *dark entities*. Although these detected dark entities are used for event representation, they are not used to populate the background KG.

From the works presented, ICEWS, GDELT and OEDA are focused on political data for the analysis of conflicts, and NewsReader generates an ECKB from financial data. Notice there is still a big coverage gap when it comes to media event encoding. In this sense, VLX-Stories offers a wider service for journalistic purposes, as it covers, as well as politics and finances, many other categories, like sports, entertainment, lifestyle, science and technology.

8 Conclusions

We presented an online event encoding system which aggregates news articles from RSS feeds, and encodes this information using semantic entities from an external KG. These entities populate an event ontology which answers the journalistic Ws. On the process, discovered EE complete the external KG (VKG) with OOKB entities. VLX-Stories realizes thus a twofold functionality: (a) generating an event KB, and (b) maintaining a KG with EEs. The detected events are served through API calls and a dashboard to media producers and other global media companies. These companies use VLX-Stories in the editorial process to identify which topics are gaining momentum, find news related to their contents, and searching for background information on trending stories.

The system matches unstructured text with Semantic Web resources, by exploiting Information Extraction techniques and external knowledge resources. This makes possible the multilingual linkage across events, semantic search, and the linkage to customer contents by matching entities. Moreover the ontological structure behind it facilitates event comprehension, search and navigation. Our

¹⁰ <http://openeventdata.org/>.

¹¹ <http://www.newsreader-project.eu/>.

engine processes an average of 17,000 articles/day, and detects an average 300 worldwide events/day from seven different countries and three languages. Our experimental results show an F-1 score of 72.32% for event detection, and a high capacity of detecting EE of people, with an average of 78.86% of the deleted entities being detected again. EE detection statistics show that the system detects an average of almost 68 EE/day, the 75.45% of which become new entities, and 22.15% are used to populate VKG entities with new alias.

We plan to continue this work by adding more countries, and improving the event representation by extracting semantic triplets that would describe the relations between the entities on the event. Regarding the KG maintenance process we will include the detection of other types of EE.

Acknowledgments. Dèlia Fernández-Cañellas is funded by contract 2017-DI-011 of the Industrial Doctorate Program of the Government of Catalonia.

References

1. Agrawal, R.S., Srikant, P.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD Record, vol. 22, pp. 207–216. ACM (1993)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC-2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
4. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. ACM (2008)
5. Conrad, J.G., Bender, M.: Semi-supervised events clustering in news retrieval. In: NewsIR@ ECIR, pp. 21–26 (2016)
6. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017). <http://archive.ics.uci.edu/ml>
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231 (1996)
8. Fernández, D., et al.: Vits: video tagging system from massive web multimedia collections. In: Proceedings of the 5th Workshop on Web-scale Vision and Social Media (VSM), pp. 337–346. IEEE Press (2017)
9. Guo, X., Gao, L., Liu, X., Yin, J.: Improved deep embedded clustering with local structure preservation. In: IJCAI, pp. 1753–1759 (2017)
10. Hamborg, F., Breiting, C., Schubotz, M., Lachnit, S., Gipp, B.: Extraction of main event descriptors from news articles by answering the journalistic five W and one H questions. In: JCDL, pp. 339–340 (2018)
11. Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., Gipp, B.: Giveme5W: main event retrieval from news articles by extraction of the five journalistic W questions.

- In: Chowdhury, G., McLeod, J., Gillet, V., Willett, P. (eds.) iConference 2018. LNCS, vol. 10766, pp. 356–366. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78105-1_39
12. Hennig, L., et al.: SPIGA-a multilingual news aggregator. In: Proceedings of GSCL 2011 (2011)
 13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
 14. Hoffart, J., Milchevski, D., Weikum, G., Anand, A., Singh, J.: The knowledge awakens: keeping knowledge bases fresh with emerging entities. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 203–206. International World Wide Web Conferences Steering Committee (2016)
 15. Ji, H., Grishman, R.: Refining event extraction through cross-document inference. In: Proceedings of ACL 2008: HLT, pp. 254–262 (2008)
 16. Jou, B., Li, H., Ellis, J.G., Morozoff-Abegauz, D., Chang, S.F.: Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 357–360. ACM (2013)
 17. Kuzey, E., Vreeken, J., Weikum, G.: A fresh look on knowledge bases: distilling named events from news. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1689–1698. ACM (2014)
 18. Kwak, H., An, J.: A first look at global news coverage of disasters by using the GDELT dataset. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 300–308. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13734-6_22
 19. Le, P., Titov, I.: Improving entity linking by modeling latent relations between mentions. arXiv preprint [arXiv:1804.10637](https://arxiv.org/abs/1804.10637) (2018)
 20. Leban, G., Fortuna, B., Grobelnik, M.: Using news articles for real-time cross-lingual event detection and filtering. In: NewsIR@ ECIR, pp. 33–38 (2016)
 21. Leetaru, K., Schrodt, P.A.: GDELT: global data on events, location, and tone, 1979–2012. In: ISA Annual Convention, vol. 2, pp. 1–49. Citeseer (2013)
 22. Li, H., Ellis, J.G., Ji, H., Chang, S.F.: Event specific multimodal pattern mining for knowledge base construction. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 821–830. ACM (2016)
 23. Luo, G., Huang, X., Lin, C.Y., Nie, Z.: Joint entity recognition and disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 879–888 (2015)
 24. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
 25. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction-meets the semantic web: a survey. *Semant. Web* (Preprint), 1–81 (2018)
 26. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
 27. Ng, K.W., Tsai, F.S., Chen, L., Goh, K.C.: Novelty detection for text documents using named entity recognition. In: 2007 6th International Conference on Information, Communications & Signal Processing, pp. 1–5. IEEE (2007)
 28. Polleres, A., Hogan, A., Harth, A., Decker, S.: Can we ever catch up with the web? *Semant. Web* **1**(1, 2), 45–52 (2010)

29. Rospocher, M., et al.: Building event-centric knowledge graphs from news. *J. Web Semant.* **37**, 132–151 (2016)
30. Sagi, T., Wolf, Y., Hose, K.: How new is the (RDF) news? In: *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 714–721. ACM (2019)
31. Schrodtt, P.A., Beielser, J., Idris, M.: Three’s a charm?: open event data coding with EL: DIABLO, PETRARCH, and the open event data alliance. In: *ISA Annual Convention* (2014)
32. Shah, C., Croft, W.B., Jensen, D.: Representing documents with named entities for story link detection (SLD). In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, pp. 868–869. ACM (2006)
33. Singer, J.B.: Five Ws and an H: digital challenges in newspaper newsrooms and boardrooms. *Int. J. Media Manag.* **10**, 122–129 (2008)
34. Singh, J., Hoffart, J., Anand, A.: Discovering entities with just a little help from you. In: *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pp. 1331–1340. ACM (2016)
35. Steinberger, J.: MediaGist: a cross-lingual analyser of aggregated news and commentaries. In: *Proceedings of ACL-2016 System Demonstrations*, pp. 145–150 (2016)
36. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: *Proceedings of the 16th International Conference on World Wide Web*, pp. 697–706. ACM (2007)
37. Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (SEM). *Web Semant.: Sci. Serv. Agents World Wide Web* **9**(2), 128–136 (2011)
38. Vossen, P., et al.: Newsreader: using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowl.-Based Syst.* **110**, 60–85 (2016)
39. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. *Commun. ACM* **57**(10), 78–85 (2014)
40. Wang, W.: Event detection and extraction from news articles. Ph.D. thesis, Virginia Tech (2018)
41. Ward, M.D., Beger, A., Cutler, J., Dickenson, M., Dorff, C., Radford, B.: Comparing GDELT and ICEWS event data. *Analysis* **21**(1), 267–297 (2013)
42. Wu, Z., Liang, C., Giles, C.L.: Storybase: towards building a knowledge base for news events. In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pp. 133–138 (2015)
43. Zhang, T., et al.: Improving event extraction via multimodal integration. In: *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 270–278. ACM (2017)