Emerging News task: Detecting emerging events from social media and news feeds

Marc Gallofré Ocaña, Andreas L. Opdahl, Duc-Tien Dang-Nguyen
University of Bergen, Norway
{marc.gallofre,andreas.opdahl,ductien.dangnguyen}@uib.no

ABSTRACT

The Emerging News task at MediaEval 2021 involves finding emerging events in a real-time stream of news and Twitter messages, and providing relevant insights for journalists like identifying the tweets related to the news stories. Both news texts and tweets are represented as RDF-graphs following the Event Description Ontology. The RDF-graphs describe the named entities identified in the text using Linked Open Data resources from Wikiedata and DBpedia.

1 INTRODUCTION

For news organisations it is critical to identify events and emerging situations as soon as they appear, delays can cause economic and audience losses [13]. Keeping journalists and readers up-to-date is a highly demanding task. News agencies spend a lot of time and human power on continuously monitoring social media, TV programs, radio shows and blogs looking for new events. Artificial Intelligence (AI) and Big Data can assist news agencies and alleviate journalists on this tedious task by distilling events from the different media channels and assess their newsworthiness, while keeping journalists in the loop for human judgement.

The vast amount of information that is continuously broadcasted on the internet makes it significantly challenging for journalists to distil daily events [8]. For example, Twitter publishes more than 500 million tweets a day (i.e., an average of 5700 tweets per second) [10] and more than 10000 English news articles are published online every day worldwide [9]. Some news agencies already employ software solutions to support events detection and assessing their newsworthiness [6]. Automating the detection of emerging events from social media and news feeds can help news agencies discover new events when they are not the first ones to cover them.

2 TASK DESCRIPTION

The *Emerging News* task aims to explore novel ways to detect emerging events from a stream of social media messages and news feeds. We define emerging events as those newsworthy events [3] that can potentially be turned into news stories and have not been yet covered in the current mainstream. Therefore, the emerging events represent news that have not been widely covered or published yet.

Participants are expected to develop a real-time solution that identifies emerging events and relates social media messages to them. The solution must read from a stream of news-related items and output those stories that could be considered emerging events. These news-related items are semantically represented using RDF graphs following the Event Description Ontology [12]. We used DBpedia Spotlight [11] to identify the named entities from the items text and Linked Open Data (LOD) resources from Wikiedata and DBpedia to describe these entities [2, 4]. The expected solutions must work with these semantic representations to identify and provide the potential emerging events.

This task is proposed in context with the News Angler project [5] which uses AI and Big Data techniques to exploit social media and online sources. The project's purpose is to support journalists in finding new and unexpected angles and unfolding news stories, along with suitable background information. Central AI techniques so far are knowledge graphs, ontologies, LOD, natural language processing (NLP) and machine learning (ML). Knowledge graphs, ontologies and LOD offer a standard form for representing information and knowledge. In this way, the collected information can be analysed, retrieved, and shared more easily and precisely.

As part of the News Angler project, we developed an evolving big-data platform that harvests potentially news-related information in real time from textual sources, such as social media, websites, commercial news aggregators, and open reference sources [7]. These news-related items are semantically annotated using a NLP pipeline inspired by [1] that outputs RDF graphs following the Event Description Ontology [12]. As a result, the News Angler platform provides an real-time stream of RDF-graph representing news-related items. We want to extend the platform with new components for analysing news items representations and providing newsworthy information to journalists.

Participants at the Emerging News task have access to the RDF-graph stream of social media messages and news feeds produced by the News Angler platform. The data stream can be accessed through an API from where participants can get JSON-LD objects that contain semantic metadata along with strings representing the RDF graphs serialised in TURTLE. We expect participants to use the JSON-LD stream as input of their proposed solutions. Participants can choose to use either a continuous stream or time-windowed batches of, for example, 5, 10, 15, 20 minutes. We expect participants to discuss the most suitable set-up for their solutions. As an output, participants' solutions must provide a group of JSON-LD items that belongs to the emerging event or a single JSON-LD item that is an emerging event (we leave it to the participants' decision too). Optionally, participants can provide a user interface to better interpret and evaluate the results.

3 DATA DESCRIPTION

Data is accessed trough an API that returns one JSON-LD object from the current real-time stream on every call. The real-time stream provides tweets and news as soon as they are published. It outputs tweets from more than 70 accounts comprising different news agencies and journalists accounts, and a sample news articles

MediaEval'21, December 13-15 2021, Online

(i.e., between 100 and 500 articles every 15 minutes) from news sources and blogs over the web.

Each returned JSON-LD objects has an URI for the news-related item and a string representing one RDF graph in TURTLE notation. The URI contains an MD5 hash of the original news or tweet URL. Because the News Angler platform harvests news from news aggregators as well as from original sources, it may contain duplicates, hence, the URI can be employed to remove those duplicates. The RDF graph representing the news-related item is described following the Event Description Ontology [12] and annotated with the named entities found in the text. The RDF graph also contains the source text, to help participants and evaluators understand them.

4 EVALUATION AND RESULTS

In order to improve and facilitate the replication and dissemination of the results, we ask participants to (a) publish the proposed solution, associated files, AI/ML models and documentation files using the MIT copyright licence; (b) release the proposed solution as a dockerized API with its associated Dockerfile, and instructions on how to run it (optionally, it can be accompanied by another code or docker image that simulates the data ingestion); and (c) design a solution that does not require more than 16GB RAM and a quad-core CPU to run. We believe these constraints promote the accessibility of the proposed solutions increasing the research dissemination and contribution to a large audience, as well as facilitating the evaluation of the proposed solutions.

The proposed solutions will be evaluated based on their relevance by a panel of experts with relevant background in journalism and media. The experts will use the developed solution to judge if the information provided by the solutions can be considered as an emerging event or not and how useful the information is. The panel will focus on (a) the newsworthiness and completeness of the reported event, if the information related to the event provides enough insights to conform a news story; and (b) the relevance, if the emerging event reports a new aspect of an existing story or a new story. During the evaluation, all participants will use the same data set. Participants must not assume experts know RDF, therefore experts will base their decisions on the source text.

5 DISCUSSION

This was the first edition of News Emerging task and it lacked a well-established standard data-set for evaluation. Because of this, we designed an evaluation method based on experts to judge the quality of the outputs. At the same time, it forced participants to explore solutions that work on real-time data, instead of being tuned to particular data-set characteristics.

The *Emerging News* task is focused on RDF-graph representations instead of using text-based approaches. While this decision reduces information from text, it also reduces ambiguities, facilitates data enrichment from external sources and enables reasoning and structural marching. The resulting solutions can perhaps be applicable to other types of news-related items like images, where there is no implicit text.

ACKNOWLEDGMENTS

Supported by the News Angler project funded by the Norwegian Research Council's IKTPLUSS programme as project 275872.

REFERENCES

[1] Tareq Al-Moslmi and Marc Gallofré Ocaña. 2020. Lifting News into a Journalistic Knowledge Platform. In *Proceedings of the CIKM 2020 Workshops*. Galway, Ireland.

[2] Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L. Opdahl, and Csaba Veres. 2020. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access* 8 (2020), 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

[3] Tareq Abdo Abdullah Al-Moslmi, Marc Gallofré Ocaña, Andreas Lothe Opdahl, and Bjørnar Tessem. 2019. Detecting Newsworthy Events in a Journalistic Platform. In *The 3rd European Data and Computational Journalism Conference*.

[4] Mohammed Albared, Marc Gallofré Ocaña, Abdullah Ghareb, and Tareq Al-Moslmi. 2019. Recent Progress of Named Entity Recognition over the Most Popular Datasets. In *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*. https://doi.org/10.1109/ICOICE48418.2019.9035170

[5] Marc Gallofré Ocaña, Lars Nyre, Andreas L Opdahl, Bjørnar Tessem, Christoph Trattner, and Csaba Veres. 2018. Towards a Big Data Platform for News Angles. In *4th Norwegian Big Data Symposium (NOBIDS) 2018* (Tondheim, Norway). http://ceur-ws.org/Vol-2316/paper1.pdf

[6] Marc Gallofré Ocaña and Andreas Lothe Opdahl. 2020. Challenges and Opportunities for Journalistic Knowledge Platforms. In *Conference on Information and Knowledge Management (CIKM 2020 Workshops)*. http://ceur-ws.org/Vol-2699/paper43.pdf

[7] Marc Gallofré Ocaña and Andreas L. Opdahl. To appear. Developing a Software Reference Architecture for Journalistic Knowledge Platforms. In *European Conference on Software Architecture (ECSA2021 Companion Volume)*.

[8] Ulrich Germann, Renārs Liepins, Guntis Barzdins, Didzis Gosko, Sebastião Miranda, and David Nogueira. 2018. The SUMMA Platform: A Scalable Infrastructure for Multi-lingual Multi-media Monitoring. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-4017

[9] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2020. Bias-aware news analysis using matrix-based news aggregation. *International Journal on Digital Libraries* 21 (2020).

[10] Raffi Krikorian. 2013. New Tweets per second record, and how! (2013). Retrieved September 16, 2021 from https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how

[11] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11)*. Association for Computing Machinery, 1–8. https://doi.org/10.1145/2063518.2063519

[12] Andreas L Opdahl and Bjørnar Tessem. 2020. Ontologies for finding journalistic angles. *Software and Systems Modeling* (2020).

[13] Jorge Vázquez-Herrero, Sabela Direito-Rebollal, Alba Silva-Rodríguez, and Xosé López-García. 2020. *Journalistic Metamorphosis: Media Transformation in the Digital Age*. Springer, Cham. https://doi.org/10.1007/978-3-030-36315-4