Interactive System for Automatically Generating Temporal Narratives

Arian Pasquali[1,2(✉)], Vítor Mangaravite[1], Ricardo Campos[1,3], Alípio Mário Jorge[1,2], and Adam Jatowt[4]

[1] LIAAD – INESCTEC, Porto, Portugal
{arrp,vima}@inesctec.pt
[2] FCUP, University of Porto, Porto, Portugal
amjorge@fc.up.pt
[3] Polytechnic Institute of Tomar - Smart Cities Research Center, Tomar, Portugal
ricardo.campos@ipt.pt
[4] Kyoto University, Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Abstract. In this demo, we present a tool that allows to automatically generate temporal summarization of news collections. Conta-me Histórias (Tell me stories) is a friendly user interface that enables users to explore and revisit events in the past. To select relevant stories and temporal periods, we rely on a key-phrase extraction algorithm developed by our research team, and event detection methods made available by the research community. Additionally, we offer the engine as an open source package that can be extended to support different datasets or languages. The work described here stems from our participation at the Arquivo.pt 2018 competition, where we have been awarded the first prize.

Keywords: Information retrieval · Temporal summarization

1 Introduction

During the last decade, we have been witnessing an ever-growing number of online content posing new challenges for those who aim to understand a given event. This exponential growth of the volume of data, together with the phenomenon of media bias, fake news and filter bubbles, has contributed to the creation of new challenges in information access and transparency. For instance, following the media coverage of long-lasting events like wars, migration or economic crises can be oftentimes confusing and demanding. One possible solution is the adoption of timelines to support story-telling as a way to organize the different phases of complex events. Media outlets use this type of solution very often. However, manually building such timelines can be very laborious and time-consuming for journalists. Besides, it simply does not scale. One possible approach to overcome this problem is to automatically summarize large amount of news into consistent narratives, an active topic of research in the academic community [1,6] with the proposals of innovative and creative solutions [7,8].

L. Azzopardi et al. (Eds.): ECIR 2019, LNCS 11438, pp. 251–255, 2019.
https://doi.org/10.1007/978-3-030-15719-7_34

2 Proposed Solution

In this demo paper, we propose a user-friendly interface that allows running queries on news sources and exploring the results in a summarized and temporally organized manner with the help of an interactive timeline. Our project named Conta-me Histórias (Tell me stories) results from the participation on the Arquivo.pt 2018 contest, where we have achieved the first rank among 27 competing teams[1].

Given a user query, our system automatically identifies relevant dates and the most important headlines to illustrate the story. Figure 1 gives an overview of the framework, which can be described in 5 simple steps: (1) News Retrieval; (2) Term Weighting; (3) Identifying Relevant Time Intervals; (4) Computing Headline Scores; (5) Deduplication.

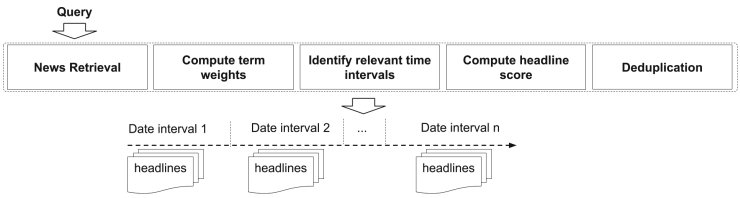

Fig. 1. System architecture

News Retrieval. The first step in the pipeline is to run the query against any data source of interest. The only requirements are that the result set must contain a list of items with a headline, timestamp, URL and optionally a source name. The source code for our temporal summarization framework, as well as examples on how to adapt for different data sources are available online[2].

Term Weighting. We then calculate each term weight through an adapted version of YAKE! [2,3] keyword extractor method (Best Short Paper at ECIR'18), which relies on statistical features to select the most important key-phrases of a document. In our approach, every headline is treated as an independent document. We calculate a number of term statistics, such as frequency within the entire result set, average frequency, standard deviation and positional features. All these features enable the identification of common and rare terms with much higher accuracy than TF-IDF. This step produces a term dictionary that will be used later in the pipeline to identify n-grams candidates. A more thorough discussion of the details of YAKE!, may be found in the above referred papers.

Identifying Relevant Time Intervals. To select relevant time periods, we applied a strategy that forces the system to select intervals with at least one

[1] http://sobre.arquivo.pt/en/arquivo-pt-2018-award-winners/.
[2] https://github.com/LIAAD/TemporalSummarizationFramework.

peak of occurrence. This strategy tries to identify main events related to the query assuming such events result in many headlines in a short period of time. We begin by dividing the timespan into 60 equi-width intervals (partitions). These partitions are used to aggregate the frequencies in order to find peaks of occurrences. The interval boundaries are then given by the fewer partition (*smallest peak*) among each pair of peaks; We apply the function *argrelextrema* from the Scipy's signal processing module[3] in order to find each relative peak of occurrences. Figure 2 illustrates time interval selection for the query *"Guerra na Siria"* (War at Syria). In this case, the system identified seven important time intervals (between 2010 and 2016): the red lines represent interval boundaries, while the blue ones highlight number of news aggregated by date.

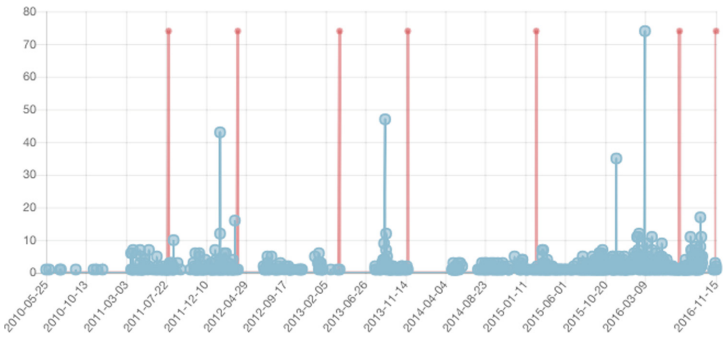

Fig. 2. Relevant time intervals detection for the query *"War at Syria"*

Computing Headline Scores. After determining the relevant time periods, we now aim to determine the most important headlines for each temporal interval. It is important to mention that it is not a simple concatenation of headlines, as we will only present the most important ones. In order to proceed with this summarization process, we look at each individual term of the headline and compute an aggregated value based on its individual term weight. A fully detailed description of the underlying scientific approach and the evaluation methodology on this particular aspect can be found in our recent work [3].

De-duplication. Finally, we eliminate similar key-phrases based on Levenshtein similarity measure. We specify a threshold where we ignore key-phrases that are more than 80% similar. When comparing a pair of strings we keep the longer one, assuming it carries more information than the shorter version. This threshold is a parameter and can be fine tuned for different cases. For each time interval, we then select the top 20 key-phrases ordered by their relevance. It also is possible to experiment with different methods like Jaccard, Monge-Elkan and Jaro-Winkler. A formal evaluation on this and other parameters are left for future work.

[3] https://docs.scipy.org/doc/scipy/reference/signal.html.

3 Demonstration

Users can interact with our demo either through Conta-me Histórias[4] or Tell me Stories[5]. The former allows users to explore the Portuguese Web Archive [5] through their free-text search API which enable users to explore their archived results from a period that is mostly concentrated on 2010 to the present. To guarantee the plurality and the diversity of the information, we consider news from 24 popular Portuguese media outlets. The latter is built on top of the Signal Media Dataset [4], a one-million news articles collection (mainly English, but also non-English and multi-lingual articles) which were originally collected rom a variety of news sources (such as Reuters) for a period of 1 month (1–30 September 2015).

Our proposed solution can be easily adapted to other scenarios including different kinds of data sources (e.g. social media posts, academic papers, proprietary repository, etc) and languages since it is mostly language independent. This may be understood as an important contribution for anyone interested in having access to a summarized temporal view of their data. In its current form, users may interact with the system through an interface where several options, including the specification of a query (a free text field) and a time interval (last five, ten, twenty or thirty years) are offered to the users. Below we present the results for the query War at Syria obtained from the Portuguese Web Archive project. The results (which were translated from Portuguese to English) helps us to understand this long-lasting long-lasting conflict, highlighting the most important headlines since 2010, when the first popular protests reached the streets and the news. We can see the brutal repression and the escalation of violence against civilians. Subsequently, that year, a famous journalist died in an attack, other countries got involved in the conflict and news about humanitarian crises appeared. We also confirm events where civilians suffered attacks like the infamous air strike and bombing on a maternity unit. This piece of example illustrates how this kind of tool can help to understand the escalation of violence in Syria, helping users to explore relevant dates, headlines and actors in the story.

From	To	Top headlines
5/2010	7/2011	Syrian officials launch tear gas against protesters Security forces shoot at protesters New York Times journalist with the Pulitzer died of an Asthma attack in Syria
8/2011	3/2012	Assad promises elections in February in Syria US withdraws ambassador from Syria for security reasons NATO says goodbye to Libya and the world turns to Syria
7/2012	12/2012	Meeting of senior officials in Geneva failed agreement to end violence in Syria Russia delivers three war helicopters to Syria Red Cross says Syria is in civil war
7/2016	11/2016	Maternity unit among hospitals bombed in Idlib air strikes Russian helicopter shot down in Syria. Turkish army enters Syria

[4] http://contamehistorias.pt/arquivopt.
[5] http://signal.tellmestories.pt.

An additional available experimental interface[6] was made available such that interested researchers may experimentally test our demo with several different options, such as deduplication and event detection methods.

Acknowledgements. This work is partially funded by the ERDF through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT as part of project UID/EEA/50014/2013.

References

1. Aslam, J.A., Ekstrand-Abueg, M., Pavlu, V., Diaz, F., Sakai, T.: TREC 2013 temporal summarization. In: TREC (2013)
2. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A.M., Nunes, C., Jatowt, A.: YAKE! Collection-independent automatic keyword extractor. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 806–810. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_80
3. Campos, R., et al.: A text feature based automatic keyword extraction method for single documents. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 684–691. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_63
4. Corney, D., Albakour, D., Martinez, M., Moussa, S.: What do a million news articles look like? In: Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, 20 March 2016, pp. 42–47 (2016)
5. Gomes, D., Cruz, D., Miranda, J., Costa, M., Fontes, S.: Search the past with the Portuguese web archive. In: 22nd International World Wide Web Conference, Rio de Janeiro, Brasil (2013)
6. Jorge, A.M., et al.: Report on the first international workshop on narrative extraction from texts (Text2Story 2018). In: SIGIR Forum, vol. 52, no. 1, pp. 150–152. ACM Press (2018)
7. Schubotz, T., Krestel, R.: Online temporal summarization of news events. In: 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) vol. 1, pp. 409–412 (2015)
8. Tran, G., Alrifai, M., Herder, E.: Timeline summarization from relevant headlines. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 245–256. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16354-3_26

[6] http://labs.tellmestories.pt.