Time-Matters: Temporal Unfolding of Texts

Ricardo Campos[1,2] [0000-0002-8767-8126], Jorge Duque[2], Tiago Cândido[2], Jorge Mendes[2], Gaël Dias[3] [0000-0002-5840-1603], Alípio Jorge [1,4] [0000-0002-5475-1382] and Célia Nunes[5] [0000-0003-0167-4851]

[1] LIAAD – INESCTEC, Porto, Portugal
[2] Polytechnic Institute of Tomar, Ci2 - Smart Cities Research Center, Tomar, Portugal
{ricardo.campos, aluno19893, aluno20185, 19263}@ipt.pt
[3] Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, France
gael.dias@unicaen.fr
[4] FCUP, University of Porto, Porto, Portugal
amjorge@fc.up.pt
[5] Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal
celian@ubi.pt

Abstract. Over the past few years, the amount of information generated, consumed and stored on the Web has grown exponentially, making it impossible for users to keep up to date. Temporal data representation can help in this process by giving documents a sense of organization. Timelines are a natural way to showcase this data giving users the chance to get familiar with a topic in a shorter amount of time. Despite their importance, little is known about their use in the context of single documents. In this paper, we present Time-Matters, a novel system to automatically explore arbitrary texts through temporal narratives in an interactive fashion that allows users to get insights into the relevant temporal happenings of a story through multiple components, including temporal annotation, storylines or temporal clustering. In contrast to classical timeline multi-document summarization tasks, we focus on performing text summaries of single documents with a temporal lens. This approach may be of interest to a number of providers such as media outlets, for which automatically building a condensed overview of a text is an important issue.

Keywords: Timeline Generation, Temporal Narratives, Temporal Information.

1 Introduction

Recent times have shown an abundance of textual content creating new challenges for those who want to quickly get insights, without having to read entire documents. Much of this text is in free form. Extracting information from it requires the use of computer resources capable of understanding natural language. Presenting text using temporal structures can help reduce the effort of the reader [1,14]. For example, they can define the time period of events in news articles [16,18], play an important role in communication platforms, such as Twitter [1,2,3] or Wikipedia [12], and help contextualize historical texts [13] or legal documents [11]. Advances on these domains are partially due to the existence of temporal taggers, such as Heideltime [17] or Sutime [8]. Timelines appear in this context as a common approach that leverages the detected temporal signals to summarize the information spread over multiple documents in a temporal order fashion. However, little is known about their use in the scope of single documents. An

optimal summary should cover all the important temporal aspects of a text while disregarding unimportant or irrelevant dates. However, manually building these timelines may be a laborious and time-consuming task, and an impossible effort for average users or professionals interested in making sense of an increasing volume of textual data. This slows down the process of text analytics and data understanding. In this paper, we present Time-Matters, a novel system that can give users an automatic overview of the most important time-periods and associated text stories in a short amount of time without having to read text-heavy documents. This can be very useful in several scenarios and domains, and fits within the recent trend of automatically generating narratives from texts [7]. For instance, it may be of importance for media outlets [15], interested in telling stories and in reaching new audiences with alternative and appealing forms without overburden their workforce.

To accomplish this objective, we adapted a previously introduced version of Time-Matters [5] which worked over queries and multiple documents, to single texts. In particular, we aim to estimate the importance of the temporal expressions detected in a text and hence disregard the non-relevant ones. The goal is to not only provide a temporal annotation of the text with the corresponding scores given by the Time-Matters algorithm, but also to offer users the chance to interact with the system with a temporal storyline component that shows the most important stories of a text. We do this in an interactive fashion that includes a timeline and graphical elements likely related to parts of the story. Further possibilities include exploring the most relevant stories of the text through temporal clustering. Another important key aspect of our approach is that it is **unsupervised**, **domain**, and **corpus-independent** as it does not require any training stage and builds upon local text statistical features extracted from single documents. Hence, it can readily be applied to any text. The core of Time-Matters is also **mostly language-independent**. While it anchors on Heideltime [17] to detect temporal expressions it can also use a rule-based approach, which, while not as effective as Heideltime, may be a good solution when performance and language is an issue. As a contribution to the research community, we make available an **online demo** [http://time-matters.inesctec.pt], a **video demonstration** [http://bit.ly/2HDwqjD], an **API** [http://time-matters.inesctec.pt/api], a **python package** [https://bit.ly/31W93Jg] and a **docker image** [https://dockr.ly/2TzwOC9] of Time-Matters. On the sidelines, we also make public a python package wrapper for Heideltime [https://bit.ly/34Ifvp7] which aims to facilitate the use of this well-known temporal tagger.

2 Time-Matters Algorithm

Our assumption is that the relevance of a candidate date d_j may be determined with regards to the relevant terms W_j^* that it co-occurs with in a given context (defined as a window of *n* terms in a sentence or the sentence itself). That is, the more a given candidate date is correlated with the most relevant keywords of a text t_i, the more relevant the candidate date is for the text at hand. To model this temporal relevance, we rely on the Generic Temporal Similarity measure (GTE) [5], which makes use of co-occurrences of keywords and temporal expressions as a means to identify relevant dates within a text. In this work, relevant keyphrases and temporal expressions are respectively detected by YAKE! keyword extractor [6], and Heideltime temporal tagger [8].

GTE is formalized in Equation 1 and ranges between 0 (irrelevant) and 1 (relevant), where *IS* is the InfoSimba similarity measure [9].

$$\mathrm{GTE}(t_i, d_j) = \mathrm{median}\left(\mathrm{IS}(\mathrm{w}_{\ell,j}, d_j)\right), \mathrm{w}_{\ell,j} \in W_j^* \quad (1)$$

A fully detailed description of the underlying scientific approach and the evaluation methodology for the study of queries and multiple documents can be found in Campos et al. [5]. Readers are also recommended to refer to our wiki documentation [https://github.com/LIAAD/Time-Matters/wiki] for an in-depth understanding of the single document version explored in this demo.

3 Time-Matters Demonstration

We demonstrate our approach using an arbitrary text related to the 1st anniversary of the Haiti earthquake held on January 12, 2011. Texts can be given as input in the homepage or as an URL, in which case, we make use of the well-known Newspaper 3k library [https://newspaper.readthedocs.io] to extract contents. The resulting interface is divided into five major components: "Annotated Text"; "Storyline", "Temporal Clustering"; "Timeline"; and "Scores". In this paper, we put an emphasis on the first two, "Annotated Text" and "Storyline", due to space reasons.

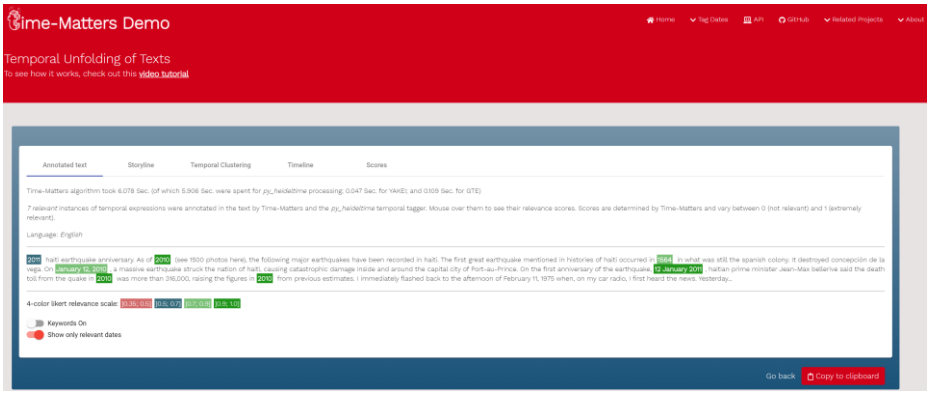

Fig. 1. "Annotated Text" interface.

Annotated Text. Fig. 1 shows the "Annotated Text" component. At the top, we can observe the time spent to obtain the results, the number of relevant annotated temporal expressions instances and the text language. Time performance is highly dependent on the Heideltime component as computing GTE scores is a quick process. Each date is tagged with a 5-color likert relevance scale, from highly irrelevant dates (bold red) to highly relevant ones (bold green). To get a sense of the relevance of the dates, users can also mouse over a given temporal expression. By default, only relevant temporal expressions, those with GTE scores equal or above 0.35 (according to the experiments conducted in [5]) are shown to the user. Scores close to 1 are considered highly relevant in the particular part of the text being analyzed. Equal date instances in different sentences can also result in different scores (one such approach can be explored in the advanced options section in the homepage). In addition to relevant dates, users can also

ask for irrelevant ones (scores<0.35) as exemplified in Fig. 1 for the temporal expression "the afternoon of February 11, 1975" (marked in bold red), which is shown a score of 0. By doing this, we give users the opportunity to understand the effectiveness of the Time-Matters algorithm in filtering out irrelevant dates initially marked by Heideltime. A more formal evaluation, in line with what has been done for the multiple-documents approach, should, however, be conducted in the future. One can also observe, marked as bold, the relevant keyphrases co-occurring next to the date and that most contribute to the results of Time-Matters. By default, *n*-grams are set to 1, meaning that keywords will be formed by 1 single token only, though other options can be defined in the advanced options setting.

Storyline visualization. The storyline interface (see Fig. 2) explores the different stories of a text through a temporal lens. The component at the top, highlights the relevant dates ("1564"), its score ("0.799"), the sentence where the date occurs and a summary of that particular part of the story ("great earthquake mentioned") given by YAKE! [6]. The story is also illustrated automatically with images. We leverage on the Portuguese web archive Arquivo.pt [10] images search API v1 [https://github.com/arquivo/pwa-technologies/wiki]. While this API can obtain results for any language it naturally works better for its native language, Portuguese. Users can then navigate between the different time-periods by either clicking at the right row (labelled in this figure example as "Recorded in Haiti, 2010") or at the bottom timeline component which gives, per se, a temporal overview of the story.

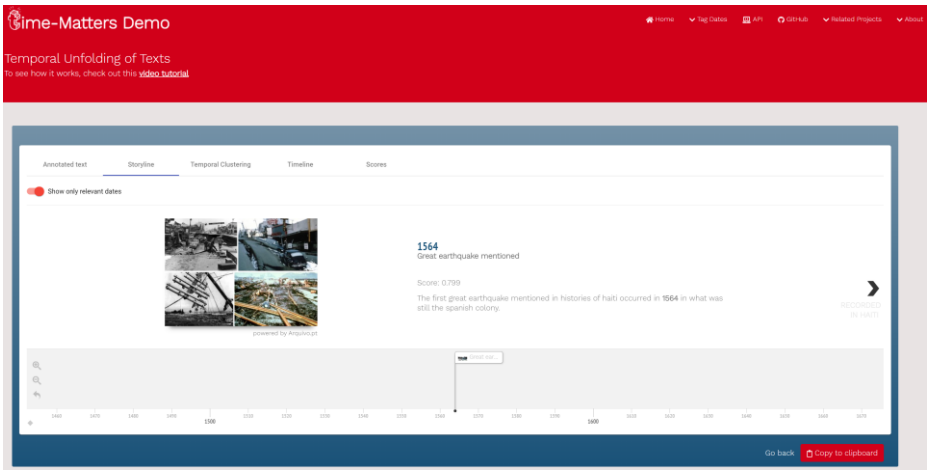

Fig. 2. "Storyline" interface.

In this paper, we suggest a simple yet effective approach for summarizing a text through a temporal perspective, highlighting the most important temporal aspects of the text. As future research, we plan to investigate further elaborated solutions that study the correlation between the detected relevant dates and the relevant events found in the surroundings of the date. This can be used to improve not only the story description but also the retrieval of images.

References

1. Alonso, O., Shiells, K. (2013). Timelines as Summaries of Popular Scheduled Events. In Proceedings of the 22nd International Conference on World Wide Web (WWW'13). pp. 1037-1044. Rio de Janeiro, Brazil. May 13-17.
2. Alonso, O., Tremblay, S-E., & Diaz, F. (2017). Automatic Generation of Event Timelines from Social Data. In Proceedings of the 2017 ACM on Web Science Conference (WebSci'17). pp. 207-211. New York, USA. June 25-28.
3. Alonso, O., Kandylas, V., & Tremblay, S-E. (2018). How it Happened: Discovering and Archiving the Evolution of a Story Using Social Signals. In Proceedings of the ACM/IESS Joint Conference on Digital Libraries (JCDL'18). pp. 193-202. Texas, USA. June 3-7.
4. Campos, R., Dias, G., Jorge, A., & Jatowt, A. (2014). Survey of Temporal Information Retrieval and Related Applications. ACM Computing Surveys, 47(2), Article 15
5. Campos, R., Dias, G., Jorge, A. and Nunes, C. (2017). Identifying Top Relevant Dates for Implicit Time Sensitive Queries. In Information Retrieval Journal. Springer, Vol 20(4), pp 363-398.
6. Campos R., Mangaravite V., Pasquali A., Jorge A.M., Nunes C., and Jatowt A. (2018). A Text Feature Based Automatic Keyword Extraction Method for Single Documents. In Proceedings of the 40th European Conference on Information Retrieval (ECIR'18). pp. 684 - 691. Grenoble, France. March 26 – 29. Springer.
7. Campos, R., Jorge, A., Jatowt, A., and Sumit, B. (2020). Third International Workshop on Narrative Extraction from Texts (Text2Story'20). In Proceedings of the 42nd European Conference on Information Retrieval (ECIR'20). pp. 648 - 653. Lisbon, Portugal. April 14 – 17. Springer.
8. Chang, A. X., & Manning, C. D. (2012). SUTIME: A Library for Recognizing and Normalizing Time Expressions. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12). pp. 3735–3740. Istambul, Turkey. May 23 - 25.
9. Dias, G., Alves, E., & Lopes, J. (2007). Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. In Proceedings of the 22nd Conference on Artificial Intelligence (AAAI'07). pp. 1334 - 1340. Vancouver, Canada. July 22 – 26.: AAAI Press.
10. Gomes, D., Cruz, D., Miranda, J., Costa, M., & Fontes, S. (2013). Search the Past with the Portuguese Web Archive. In Proceedings of the 22nd International Conference on World Wide Web (WWW'13). pp. 321-324. Rio de Janeiro, Brazil. May 13-17.
11. Hausner, P., Aumiller, D., Gertz, M. (2020). Time-Centric Exploration of Court Documents. In Proceedings of the 3rd International Workshop on Narrative Extraction from Texts (Text2Story20@ECIR'20). pp. 31- 37. Lisbon, Portugal. April 14.
12. Hausner, P., Aumiller, D., Gertz, M. (2020). TiCCo: Time-Centric Content Exploration. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM'20). pp. 3413 – 3416. Virtual Event, Ireland. October 19-23, ACM Press.
13. Jatowt, A., Campos, R., Bhowmick, S., & Doucet, A. (2019). Document in Context of Time (DICT): System that Provides Temporal Context for Analyzing Old Documents. In Proceedings of the 28th ACM International Conference on Knowledge Management (CIKM'19). Beijing, China. pp. 2869 - 2872. November 03 – 07, ACM Press.

14. Kanhabua, N., Blanco, R., & Nørvåg, K. (2015). Temporal Information Retrieval. In Foundations and Trends in Information Retrieval, Vol 9(2), pp. 91-208.
15. Martinez-Alvarez, M., Kruschwitz, U., Kazai, G., Hopfgartner, F., Corney, D., Campos, R. and Albakour, D. (2016). First International Workshop on Recent Trends in News Information Retrieval (NewsIR'16). In Proceedings of the 38th European Conference on Information Retrieval (ECIR'18). pp. 878 - 882. Padova, Italy. March 20 – 23. Springer.
16. Pasquali, A., Mangaravite, V., Campos, R., Jorge, A., & Jatowt, A. (2019). Interactive System for Automatically Generating Temporal Narratives. In Proceedings of the 41st European Conference on Information Retrieval (ECIR'19). Cologne, Germany. April 14-18: Springer.
17. Strötgen, J., & Gertz, M. (2013). Multilingual and cross-domain temporal tagging. Language Resources and Evaluation, 47(2), 269–298.
18. Tran, G., Alrifai, M., & Herder, E. (2015). Timeline Summarization from Relevant Headlines. In Proceedings of the 37th European Conference on Information Retrieval (pp. 245-256). Vienna, Austria. March 29 - April 2: Springer.