TLS-Covid19: A New Annotated Corpus for Timeline Summarization

Arian Pasquali[1](✉), Ricardo Campos[1,2], Alexandre Ribeiro[1], Brenda Santana[1], Alípio Jorge[1,3], and Adam Jatowt[4]

[1] LIAAD – INESCTEC, Porto, Portugal
{arian.r.pasquali,ricardo.campos,alexandre.m.ribeiro, brenda.s.santana}@inesctec.pt
[2] Polytechnic Institute of Tomar, Ci2 - Smart Cities Research Center, Tomar, Portugal
ricardo.campos@ipt.pt
[3] FCUP, University of Porto, Porto, Portugal
amjorge@fc.up.pt
[4] University of Innsbruck, Innsbruck, Austria
adam.jatowt@uibk.ac.at

Abstract. The rise of social media and the explosion of digital news in the web sphere have created new challenges to extract knowledge and make sense of published information. Automated timeline generation appears in this context as a promising answer to help users dealing with this information overload problem. Formally, Timeline Summarization (TLS) can be defined as a subtask of Multi-Document Summarization (MDS) conceived to highlight the most important information during the development of a story over time by summarizing long-lasting events in a timely ordered fashion. As opposed to traditional MDS, TLS has a limited number of publicly available datasets. In this paper, we propose TLS-Covid19 dataset, a novel corpus for the Portuguese and English languages. Our aim is to provide a new, larger and multi-lingual TLS annotated dataset that could foster timeline summarization evaluation research and, at the same time, enable the study of news coverage about the COVID-19 pandemic. TLS-Covid19 consists of 178 curated topics related to the COVID-19 outbreak, with associated news articles covering almost the entire year of 2020 and their respective reference timelines as gold-standard. As a final outcome, we conduct an experimental study on the proposed dataset over two extreme baseline methods. All the resources are publicly available at https://github.com/LIAAD/tls-covid19.

Keywords: Timeline summarization · Datasets · Evaluation

1 Introduction

Following media coverage of long-lasting events like wars, epidemics or economic crises is demanding for readers, journalists, specialists and scholars. How did the S.A.R.S. epidemic crisis evolve in the early 2000s? What are the similarities with modern events? One common solution to this problem that can offer answers to the above-mentioned

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12656, pp. 497–512, 2021.
https://doi.org/10.1007/978-3-030-72113-8_33

example questions is the adoption of timelines to support storytelling as a method to organize the different phases of complex events. For instance, media outlets frequently use timelines to illustrate stories. However, manually building such timelines can be very laborious and time-consuming even with the support of modern search engines. Understanding the evolution and implications of these events often requires a combination of tools and search queries. Timeline summarization systems (TLS) emerge in this context as an alternative to manually digesting huge volumes of data in a short period of time by offering the possibility of creating summaries of multiple documents over time.

The recent surge of the COVID-19 outbreak is a very up-to-date example of this information overload problem exerting tremendous effort and pressure on users who want to keep up with the news. By January 20th 2021, the novel COVID-19 has been reported in 219 countries; resulting in approximately 100M confirmed cases and more than 2M deaths[1] Fighting this pandemic situation requires isolation, social distance measures, research in health and medicine care, but also contributions from the research community. The Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) was one of the firsts to make available a data repository[2] and a visual dashboard that gathers information from multiple sources. Multiple other similar initiatives have also been established worldwide. The Coronavirus Corpus[3], first released in May 2020 and currently 814M of words has also been created to shed light on what people are saying in online newspapers and magazines. Perhaps, the most widely known initiative to date was the release of the COVID-19 Open Research Dataset (CORD-19)[4]. Created by the Allen Institute for AI in partnership with five other institutes, CORD-19 [32] consists of over 158,000 scholarly articles, including over 75,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses and has fostered the emergence of multiple solutions. This is the case of the TREC-COVID challenge [30], which uses the CORD-19 dataset to build a set of Information Retrieval (IR) test collections. Aiming to support the fight against this pandemic Alam et al. [1] has also manually annotated a dataset of COVID-19 related tweets to tackle the problem of disinformation. These datasets were already applied to a variety of NLP tasks such as question answering and abstractive summarization [15]. Similarly, Yang et al. [34] developed a dialog dataset containing conversations between patients and doctors about COVID-19 to support chatbots research. Timelines can also be understood in this context as an essential resource for readers of major news outlets to quickly have access to a concise view of a given topic over time. A good temporal summary of the "*World Health Organization*" topic over the recent months should refer, for instance, to the chronological evolution of the COVID-19 outbreak, possible vaccine solutions, or the Donald Trump's ultimatum to WHO on May 2020, among many other summaries.

While several methods have been proposed to generate condensed news timelines, the problem of timeline generation is yet to be solved. One of the reasons for this is that traditional TLS datasets are restricted to just a limited number of topics [29]. However, deeply understanding long-lasting events, as is the case of COVID-19, requires a

[1] https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/.

[2] https://github.com/CSSEGISandData/COVID-19.

[3] https://www.english-corpora.org/corona/.

[4] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

significantly larger number of topics, news articles, different sources, annotated timelines, and longer time spans. Previous work on TLS also does not make it clear on how proposed methods behave across different languages. This makes it hard to assess how the methods behave under different scenarios, since almost all the datasets, with few exceptions, are for the English language, and none is multi-lingual. Sorting out these questions is crucial for researchers who lack diversified datasets to evaluate their proposed algorithms. Addressing the issues mentioned above, requires significant efforts in terms of: (1) collecting manually edited timelines from credible news sources; (2) collecting timelines and news articles, relevant to Covid-19 both in temporal and textual dimensions; and (3) selecting a representative and diversified number of topics.

TLS-Covid19 corpus emerges in this context to promote the development and the evaluation of new algorithms and applications in the context of the timeline summarization task, and at the same time, to enable the study of news coverage about the COVID-19 pandemic, from the evolution of a topic over time, to the comparison of what is being said about a certain topic by different news outlets. One can also look at keywords, part-of-speech tags, entities or events to see how things have changed over time. It also opens room to look at collocates. A few examples might be: keywords that were common in the same time-period, words that appear near covid-19 in different time-periods, entities, events, nouns or verbs that were more common at the beginning of the pandemics but no longer on December 2020. Finally, as it is common in most of the datasets of this kind, researchers are also offered the chance to create a sub-set of the dataset based on the publication date, the source, the country, etc., and to apply it for different purposes than the one it was initially designed for. Our corpus consists of 178 topics (35 in English and 143 in Portuguese), their associated 100,399 news articles (32,210 in English and 68,508 in Portuguese), and 178 timelines (one for each of the 178 topics). Note, however, that we have considered two news sources per language, each with its timeline, which accounts for 356 timelines. This opens room for researchers two evaluate their systems under two different scenarios. One that considers an evaluation over the news sources, based on the fact that each one has its ground-truth timeline. The other one which considers an evaluation solely over the languages, which could be made possible by a slight modification that involves merging, for each topic, the timelines of the two different news outlets. Our main contributions are as follows:

1. We develop a new TLS corpus - TLS-Covid19 - covering two languages (English and Portuguese) from two different trustworthy news sources per language (CNN and The Guardian for English, and Público and Observador for Portuguese).
2. We open room for researchers to explore language-independent summarization methods as 30 English topics (out of 35) can also be found as topics in the Portuguese variant;
3. TLS-Covid19 is made available to the research community through a Python script that enables to reconstruct the dataset and to keep collecting further news articles and ground-truth timelines;
4. Based on this dataset, we conduct an evaluation process and present experimental results by comparing two different baselines (random; oracle upper bounds) to understand the effectiveness of TLS methods under the proposed dataset.

The remainder of this paper is organized as follows. Section 2 offers an overview of the related work in timeline summarization. Section 3 presents the current available TLS datasets. Section 4 describes the construction of the TLS-Covid19 corpus. Section 5 introduces the experimental setup. Section 6 discusses the results obtained from our comparative experiments. Finally, Sect. 7 concludes this paper by summing up the most important contributions of our research and by pointing out possible future research directions.

2 Related Work on Timeline Summarization Systems (TLS)

Summarization is an active topic that has been discussed since the'50s [21]. According to McCreadie et al. [25] it can be framed within four categories: (1) Multi-Document Summarization; (2) Timeline Generation (aka timeline summarization); (3) Update Summarization; and (4) Temporal Summarization. Most researchers [9, 10, 17] focused on single and multi-document summarization (MDS) where extractive methodologies are usually employed by selecting the most relevant sentences to produce a new single document. More recently, timeline summarization (TLS) appears as a particular case of MDS aiming to summarize events across time and to put them in an automatically generated timeline. The general idea is to extract textual units from related batch documents over time through a retrospective perspective [2–5, 14, 23, 27–29]. In this case, the temporal dimension plays an important role, and documents are assumed to be time-tagged or to have at least some inherent (possibly ambiguous) temporal information in a way that texts can be anchored in a timeline. While automatically generated summaries have proved to be a valuable instrument to digest large volumes of textual data, they are hard to evaluate. The most popular, among the available evaluation methods, focus on comparative textual evaluation, where a summary produced by an automatic system is compared against one or more gold-standard summaries manually constructed by humans. Unlike MDS, which only needs to consider the compression rate between the input documents and the reference summaries, in TLS, one is required to find not only relevant information but also relevant dates to be placed in a timeline. Catizone et al. [12] formalizes this process as follows: 1) relevant documents should be included in the appropriate timeframe; 2) each timeline unit should contain accurate text labels, and 3) the timeline should include the most significant events of the document collection. Manually generating annotated summaries, however, is a laborious and time-consuming task. In the following section, we provide a discussion about the currently available datasets. Despite a few releases over the last few years, none, to the best of our knowledge, has considered making available a multi-lingual dataset across a number of topics, likely slowing down the emergence of novel methods in the context of timeline summarization. TLS-Covid19 dataset allows to fill this gap. Its description will be given in Sect. 4.

3 Shared Tasks and TLS Datasets

With the growing maturity and understanding of TLS task, the attention of researchers has progressively shifted to include formal and standard ways of evaluating their algorithms. In this section, we begin by describing two related shared tasks, before presenting five state-of-the-art TLS datasets.

3.1 Shared Tasks

The problem of evaluating timeline summarization systems is long-standing. Within this research area, there are two shared tasks, *TREC-TS* and *SemEval 2015 Task 4*, which are worth mentioning as an alternative to datasets dedicated to TLS.

TREC-TS: From 2013 to 2015 the Text Retrieval Conference (TREC) promoted the Temporal Summarization track (TREC-TS) to formalize the process of real-time temporal summarization [6–8]. This task is similar to update summarization, where a stream of documents is processed, and each sentence is evaluated in terms of its novelty and information gain. Relevant sentences are then selected to illustrate the event in summary. Although relevant, the task definition and assumptions at TREC-TS are not explicitly designed for TLS due to its streaming nature. The robustness of these datasets has also been discussed by McCreadie et al. [24].

SemEval 2015 Task 4: Another example of a related shared task is the SemEval 2015 Task 4 [26] which focusses on cross-document event coreference resolution and cross-document temporal relation extraction to identify temporal expressions. The challenge is to use a set of full-text documents as input to extract temporal relations related to a given target entity and to present a timeline with ordered events. Although related, this shared task differs from the usual timeline summarization as its purpose is to order events instead of sentences.

3.2 TLS Tasks

While several approaches have been proposed over the years, including the above-cited shared tasks, the lack of specifically annotated corpora has limited the evaluation of the initial attempts, thus demanding researchers to create their own evaluation datasets. In this section, we describe five state-of-the-art datasets (the *Timeline17*, the *crisis dataset*, the *social timeline*, the *Chen2019* dataset, and the *entities* dataset) which have been used in the process of evaluating TLS algorithms.

Timeline17: Tran et al. [28] proposed a method that links news articles with already existing timelines edited by journalists as reference summaries. The authors selected 17 of such timelines from 9 different topics published by six different news agencies, including CNN and BBC. Considering these topics as queries, they used Google search engine to retrieve the top 400 articles published in the same timespan as the original timeline. Their final dataset consists of 4,650 articles and was made publicly available[5] to the community.

Crisis Dataset: Tran et al. [29] follows Timeline17 with a similar methodology. Authors built a new and larger dataset focused on long-timespan stories on armed conflicts, such as the Egypt Revolution, Syria War, Yemen Crisis, and Libya War. The dataset comprises 15,534 news articles and 25 manually constructed timelines extracted from 24 news agencies, obtained from January 2011 to July 2013.

[5] https://l3s.de/~gtran/timeline/.

Social Timeline: Wang et al. [31] proposed the TIMELINE2014[6], which includes news articles and their respective user comments. Similar to other works, the authors crawled articles from news providers. The timeline dataset comprises 5,788 articles and 1,436,332 comments collected from the CNN, BBC, and the NYTimes on four topics, the missing Malaysia Airlines Flight MH370, the political crisis in Ukraine, the Israel-Gaza conflict and the NSA surveillance leaks. Authors provide six timelines as ground-truth based on respective Wikipedia entries for each topic.

Chen2019: Chen et al. [13] built a Chinese language dataset based on a Chinese encyclopedia[7] specially designed for abstractive timeline summarization. The dataset consists of timelines about celebrities from different countries. Each celebrity's entry in the encyclopedia contains a biographical timeline summary and a larger section detailing their experiences. In the experiences section, each event is a paragraph with an explanation and details, which is selected as an input article.

Entities: More recently, Ghalandari and Ifrim [16] have developed a dataset with 47 timelines extracted from CNN Fast[8], a CNN directory containing a large list of curated timeline articles. Authors selected mainly timeline articles about personalities as ground-truth. For each timeline, the authors defined a set of keyphrases as queries. They collected the input articles using The Guardian's API.

A summary of the datasets' statistics (including the proposed TLS-Covid19) is given in Table 1. Next, we describe the construction of our dataset.

Table 1. Available datasets for TLS.

Dataset	Language	Domain	Timespan	#Topics	#Docs	#Timelines
Timeline17	English	News	3 years	9	4, 650	17
Crisis	English	News	4 years	4	15,534	25
Social Timeline	English	News, Comments	1 year	4	5,788	6
Chen2019	Chinese	Biographies	Decades	NA	179,423	NA
Entities	English	News	Decades	47	~ = 45,075	47
TLS-Covid19	English, Portuguese	News	11 months	178	100,399	356

4 TLS-Covid19 Dataset

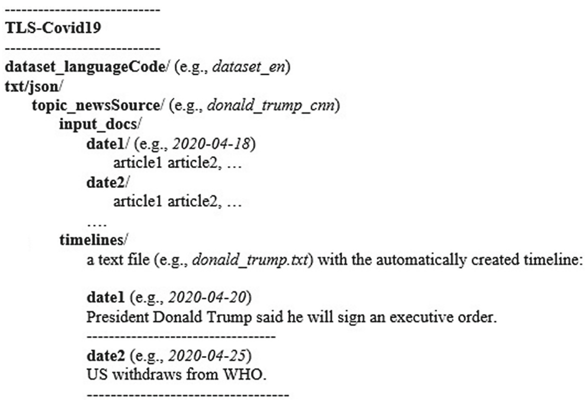
While several COVID-19 related datasets have been made available over the last few months [1], none to the best of our knowledge, is related to the timeline summarization

[6] https://web.eecs.umich.edu/~wangluxy/data.html.

[7] https://baike.baidu.com/.

[8] https://edition.cnn.com/specials/world/fast-facts.

task. In addition to this, existing datasets, as shown in the previous section, are mostly limited to a single language, thus hampering the evaluation of the proposed solutions across different scenarios. In this paper, we propose a dataset on a timely subject and relevant task that does not only address English, but also low resource languages such as Portuguese. Our future plans involve keeping collecting news articles and possibly expanding it for other languages as a means to improve its multi-lingual aspects. We invite the interested researchers on this task to join us in this effort. The current version of TLS-Covid19, consists of 178 topics (35 in English and 143 in Portuguese), their associated 100,399 news articles (31,891 in English and 68,508 in Portuguese) and timelines corresponding to the topics that cover the time period of January 2020 until December 2020. For each topic there is a number of related news articles and the corresponding ground-truth timeline. Both the news articles, as well as the timelines, are provided in two different formats (json and txt) and structured to be easily read by the tilse[9] timeline evaluation framework proposed by Martschat and Markert [23]. Figure 1 shows the format, the structure and the organization of the dataset. Details about its construction and corresponding statistics will be given in the next sections.



```
----------------------------
TLS-Covid19
----------------------------
dataset_languageCode/ (e.g., dataset_en)
txt/json/
     topic_newsSource/ (e.g., donald_trump_cnn)
          input_docs/
               date1/ (e.g., 2020-04-18)
                    article1 article2, …
               date2/
                    article1 article2, …
               ….
          timelines/
               a text file (e.g., donald_trump.txt) with the automatically created timeline:

               date1 (e.g., 2020-04-20)
               President Donald Trump said he will sign an executive order.
               ---------------------------------
               date2 (e.g., 2020-04-25)
               US withdraws from WHO.
               ----------------------------------
```

Fig. 1. Organization and structure of the dataset.

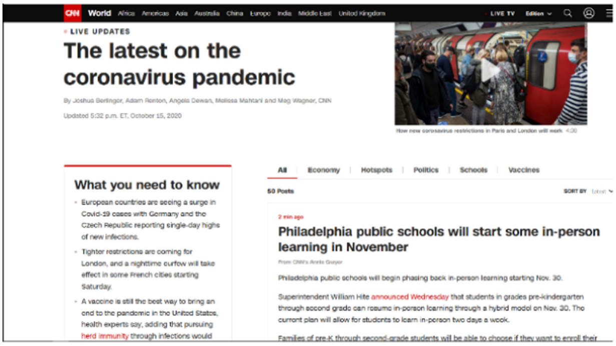
4.1 Data Collection (Input Documents and Ground-Truth)

To build this dataset, we considered two credible news sources for each language, CNN and The Guardian as the English news sources, Público and Observador as the Portuguese ones. All of them provide an everyday live coverage of the COVID-19 outbreak. The referred live coverage is provided by what is commonly known as liveblogs (CNN[10],

[9] https://github.com/smartschat/tilse.

[10] https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-12-31-20/index.html.

The Guardian[11], Público[12] and Observador[13]), a webpage (which usually has a different URL everyday) where media outlets provide news about an ongoing event, typically in the form of frequent short updates and links to news articles. In addition to the published news articles, liveblogs contain a section that highlights in a sentence-based manner the most important events during the day. These highlights are defined by journalists, thus guaranteeing their quality and credibility, and form our ground-truth timeline for that particular date. Figure 2 depicts an example of the CNN liveblog. In the figure one can observe the highlights in the left box named "What we need to know". Articles are shown on the right-hand side.



LIVE UPDATES

The latest on the coronavirus pandemic

By Joshua Berlinger, Adam Renton, Angela Dewan, Melissa Mahtani and Meg Wagner, CNN

Updated 5:32 p.m. ET, October 15, 2020

How new coronavirus restrictions in Paris and London will work 4:30

What you need to know

- European countries are seeing a surge in Covid-19 cases with Germany and the Czech Republic reporting single-day highs of new infections.
- Tighter restrictions are coming for London, and a nighttime curfew will take effect in some French cities starting Saturday.
- A vaccine is still the best way to bring an end to the pandemic in the United States, health experts say, adding that pursuing herd immunity through infections would

2 min ago

Philadelphia public schools will start some in-person learning in November

From CNN's Annie Grayer

Philadelphia public schools will begin phasing back in-person learning starting Nov. 30.

Superintendent William Hite announced Wednesday that students in grades pre-kindergarten through second grade can resume in-person learning through a hybrid model on Nov. 30. The current plan will allow for students to learn in-person two days a week.

Families of pre-K through second-grade students will be able to choose if they want to enroll their

Fig. 2. Liveblog of CNN (snapshot taken at 15/10/2020).

As a rule-of-thumb, we consider the beginning of the liveblog coverage as the start time period of collecting the articles, and December 31st, 2020 as the end time period. For instance, CNN is tracked since January 22nd, 2020; The Guardian since January 24th, 2020; Público since March 16th, 2020; and Observador since January 30th, 2020. The acquisition of the data is entirely automatic. Instructions on how to collect this data are available on a public repository[14] under which a Python script that enables the reconstruction of the dataset is provided along with all the statistics and documentation about the dataset. Our aim is to continue expanding the dataset with further articles and possibly new topics until the end of the outbreak and/or the end of the liveblogs' coverage. We anticipate that as the pandemic evolves, the amount of data collected will grow significantly.

4.2 Selecting Candidate Topics

Next step in this process is to select a list of relevant topics. Instead of conducting a topic analysis which does not fit the purposes of our study, we consider selecting topics as

[11] https://www.theguardian.com/world/live/2020/dec/30/coronavirus-live-news-uk-approves-oxford-astrazeneca-vaccine-updates.

[12] https://www.publico.pt/2020/12/31/sociedade/noticia/covid19-portugal-1944703.

[13] https://observador.pt/liveblogs/passagem-de-ano-com-restricoes-arranca-com-proibicao-de-circulacao-entre-concelhos.

[14] https://github.com/LIAAD/tls-covid19.

named entities (persons, organizations and locations), as broad concepts tend to be often used by ordinary users accessing timeline summarization systems [27]. To accomplish this objective, we apply the well-known spaCy's NLP framework [19]. Further to this, we consider selecting relevant keyphrases from our collection of highlighted texts and news articles as popular keywords are often issued by users when interacting with search engines. To this regard, we apply YAKE! [11] keyphrase extraction tool which has shown to be effective in capturing relevant keywords (e.g., "vaccine", "easter", "coronavirus", etc.).

As the first preliminary step, we begin by selecting candidate topics within the highlighted data. Our assumption is that topics appearing within text editorially defined by journalists as daily representative are likely to be relevant topics. Next, we conduct a search and match process to find the occurrences of each candidate topic in the news articles, thus collecting the corresponding input documents. Afterwards, we remove all topics from the dataset that have low temporal coverage or that appear too often. To this regard, we set the following criteria:

1. To remove candidate topics with low temporal coverage, a candidate topic must be present, similarly to Ghalandari and Ifrim [16], in at least 5 highlighted events, in both news sources;
2. To ignore candidates that appear too often (thus moving away from the summarization task), the number of occurrences for a candidate topic in the highlights should not exceed 50% of its number of occurrences in the news articles, in both news sources.

Finally, we manually curated the list of topics to consider, merging overlapping topics (e.g., "donald trump" with "trump"), and removing noise data and typos. Figure 3 shows the word cloud of the topics for both languages. The larger the font size of the text, the higher the topic frequency. As can be observed, most of the topics, regardless the language, are related to the pandemic situation in countries/locations ("*France*", "*China*", "*Italy*"), but other entities such as persons ("*Boris Johnson*") and organizations ("*Johns Hopkins University*") can also be found. Overall, we have 143 PT topics (*PER*:

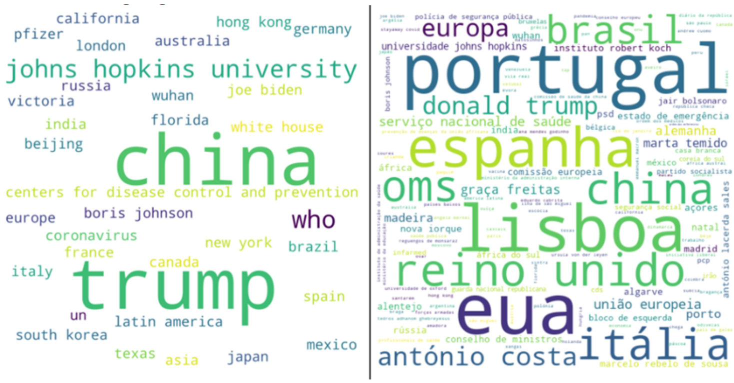


Fig. 3. English liveblog topics (left-hand side) and Portuguese liveblog topics (right-hand side).

Table 2. Overall statistics of the corpus with averages by language.

		Input Docs				Ground-Truth			Compression	
Lang	#Topics	#docs	Avg #sents	Avg #dates	Avg sents/dates	Avg #sents	Avg #dates	Avg sents/dates	Sent	Date
EN	35	31,891	3648.70	135.20	26.99	27.69	21.47	1.29	0.76	15.89
PT	143	68,508	1372.69	110.23	12.45	88.86	48.91	1.82	6.47	44.37

17; *ORG*: 33; *LOC*: 82; *Keyphrases*: 11) and 35 EN topics (*PER*: 3; *ORG*: 6; *LOC*: 25; *Keyphrases*: 1) thus representing a number of diverse topics related to the COVID-19 situation. It is also important to note that, the majority of the topics (30 out of 35) in the English dataset are represented in the Portuguese one too, thus opening room for multi-language timeline summarization research.

4.3 Dataset Statistics

Table 2 displays the main statistics of the corpus. In the table, we can observe information related to the input documents (collected news articles), ground-truth (timelines) and the compression rate, that is, the ratio between the number of sentences (or dates) in the input documents and the sentences (or dates) in the ground-truth. One can also observe that the compression rate for sentences in the English dataset is just 0.76%. Such compression rate indicates how difficult it may be to achieve high effectiveness. The lower the value, the higher the difficulty (Table. 3).

Table 3. Overall statistics by news source.

		Input Docs				Ground-Truth		
Source	#Topics	#docs	Avg #sents	Avg #dates	Avg sents/dates	Avg #sents	Avg #dates	Avg sents/dates
CNN	35	26,043	6178.54	189.71	32.57	30.11	20.97	1.44
The Guardian	35	5,848	1118.86	80.69	13.87	25.26	21.97	1.15
Público	143	28,327	1092.15	99.93	10.93	62.82	40.05	1.57
Observador	143	40,181	1653.22	120.52	13.72	114.90	57.77	1.99

5 Experimental Setup

To provide a demonstration of the validity of the proposed dataset, we conduct a set of experiments on available methods for the TLS task. The experiments conducted here serve as a guiding example. It is out of the scope of this work to make comparative experiments on top of different datasets. Although the immediate use of the dataset is tuned for unsupervised approaches, its future use is not limited to this particular setting as researchers may easily adapt it to their own needs.

5.1 Evaluation Metrics

To conduct the evaluation, we apply the tilse framework [23], a reference evaluation framework specifically designed to evaluate timeline summarization methods. In this research, we make use of the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) extension metric provided by Martschat and Markert [22] to evaluate the effectiveness of the different state-of-the-art methods. Rouge extension is particularly suited to evaluate n-grams overlaps by also taking into account the temporal information embedded in the timelines. In this work, we report the F1 scores of ROUGE-1 and ROUGE2 for the concatenation, agreement, date alignment, and date selection metrics that can be found in the tilse evaluation framework. ROUGE-1 stands for the overlap of unigrams between the automatically generated timeline and the ground-truth reference timeline, and ROUGE-2 refers to the overlap of bi-grams between the generated timeline and the ground-truth timeline summary. Naturally, dates in both the generated timeline and the ground-truth timeline may consist of one or more sentences depending solely on the number of topic references found throughout the day. Overlaps of n-grams are naturally measured within the available summary, be it a single sentence or multiple sentences. In the following, we briefly introduce each of the evaluation metrics considered in our experiments.

Concatenation: In this metric, temporal information is not considered, that is, we only look at the overlap (unigram or bigram) between the generated timeline textual summary and the corresponding ground-truth.

Agreement: In this metric, both textual, as well as temporal overlap, are taken into account. This means that, while the textual overlap between the generated timeline and the ground-truth is important, it only matters if their dates match. Otherwise, a score of 0 is assigned.

Date Selection: Finally, we consider date selection to assess how well the model behaves in exactly selecting the same dates (regardless of the textual content) between the generated timeline and the reference timelines.

5.2 Methods

In this section, we present the experimental results for the baselines *random* and *Oracle Upper Bounds*. All baselines are available in the evaluation framework tilse framework [23]. A succinct description of each one of them is presented below.

Random: is a naive baseline model that selects sentences randomly. Its results represent the worst-case scenario for a TLS constraints model.

Oracle Upper Bound (TLS Oracle): aims to calculate the best possible ROUGE scores under the input documents and the available ground-truth [18]. Such a baseline aims to estimate the best-case scenario and the level which extractive summarization algorithms can reach.

6 Results and Discussion

The results obtained from our comparative experiments are displayed in Tables 4 and 5 averaged over all generated timelines for all topics from each language in the corpus. Table 4 begins by showcasing the scores for date selection. The random baseline shows the lower bound scores that are acceptable for this task while the TLS Oracle shows the best possible results considering an extractive summarization approach. One can observe that selecting dates that match exactly with the ground-truth is easier for the Portuguese dataset because it contains a higher date coverage in the ground-truth. It is also visible how difficult it is to select the right content for the right date once we compare the ROUGE scores for the simple concatenation metric against the ROUGE in the date agreement. The difference between these two baselines represents the room for improvement that researchers can focus on. The reported results also show that scores decrease to a great extent when applying Rouge-2, thus indicating the difficulty of this task. One can conclude that, regardless of the case, there is still a long way to reach the upper bounds established by the Oracle baseline, thus opening room for further improvements within the research community. More extensive results with additional baselines are available at https://github.com/LIAAD/tls-covid19.

Table 4. Date selection scores.

	English dataset			Portuguese dataset		
Methods	Precision	Recall	F1	Precision	Recall	F1
Random	0.252	0.252	0.252	0.484	0.484	0.484
TLS Oracle	0.968	0.968	0.968	0.999	0.999	0.999

Table 5. Content selection scores using ROUGE.

			Rouge 1			Rouge 2		
Lang	Method	Metric	Prec	Recall	F1	Prec	Recall	F1
English	Random	Concat	0.183	0.190	0.187	0.022	0.023	0.023
		Agreement	0.018	0.020	0.019	0.003	0.004	0.004
	TLS Oracle	Concat	0.423	0.531	0.471	0.185	0.216	0.199
		Agreement	0.347	0.438	0.388	0.177	0.211	0.192
Portuguese	Random	Concat	0.281	0.466	0.351	0.065	0.106	0.080
		Agreement	0.059	0.097	0.073	0.013	0.023	0.017
	TLS Oracle	Concat	0.373	0.675	0.480	0.168	0.304	0.216
		Agreement	0.280	0.517	0.363	0.139	0.265	0.183

7 Conclusions

In this paper, we present the TLS-Covid19 dataset, an important resource for the TLS task. Compared to existing datasets, we provide a larger number of topics and multilingual resources on a timely subject. TLS-Covid19 consists of 178 COVID-19 related topics, 100,399 news articles and 356 reference timelines extracted from 4 news sources.

Our plan is to keep expanding this dataset until COVID-19 pandemics is over. To foster reproducibility, we provide scripts for that. To test the validity of our dataset, we performed baseline evaluations using tilse framework, a specially designed framework for TLS evaluation. The experimental results show that there is still room for improvements in this area. We believe that by providing a new dataset in this domain, we will contribute to promote the "development" of new algorithms.

Acknowledgements. The first five authors of this paper were financed by the ERDF – European Regional Development Fund through the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/CCI-COM/31857/2017 (NORTE-01–0145-FEDER-03185). This funding fits under the research line of the Text2Story project. The first author of this paper was employed by Signal Media Ltda. When part of this work was developed. The last author was employed by Kyoto University when the first version of this paper was completed.

References

1. Alam, F., et al.: Fighting the COVID-19 infodemic: modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. arXiv preprint arXiv:2005.00033 (2020)
2. Allan, J., Gupta, R., Khandelwal, V.: Temporal Summaries of New topics. SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, Louisiana, USA. September 9 – 13, pp. 1018. ACM (2001)
3. Alonso, O., Baeza-Yates, R., Gertz, M.: Exploratory search using timelines. In: ESCHI 2007: Proceedings of the Workshop on Exploratory Search and Computer Human Interaction associated to CHI2007: SIGCHI Conference on Human Factors in Computing Systems. San Jose, CA, USA. April 29, pp. 2326. ACM (2007)
4. Alonso, O., Berberich, K., Bedathur, S., Weikum, G.: Time-based exploration of News archives. In: Proceedings of the fourth Workshop on Human-Computer Interaction and Information Retrieval (HCIR), New Brunswick, USA, pp. 12–15 (2010)
5. Ansah, J., Liu, L., Kang, W., Kwashie, S., Li, J., Li, J.: A Graph is worth a thousand words: telling event stories using timeline summarization graphs. In: Proceedings of the World Wide Web Conference (WWW 2019). San Francisco, USA. May 13 – 17, pp. 25652571. ACM (2019)
6. Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., Sakai, T.: TREC 2014 Temporal Summarization Track Overview. In: Proceedings of the Twenty-Third Text Retrieval Conference (TREC 2014). Gaithersburg, USA, MIT Press (2015)
7. Aslam, J., Diaz, F., Ekstrand-Abueg, M., McCreadie, R., Pavlu, V., Sakai, T.: TREC 2015 Temporal Summarization TrackOverview. In: Proceedings of the Twenty-fourth Text REtrieval Conference (TREC 2014). Gaithersburg, USA. November 17 - 20: MIT Press (2016)
8. Aslam, J., Diaz, F., Ekstrand-Abueg, M., Pavlu, V., Sakai, T.: TREC 2013 Temporal Summarization. In: Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013). Gaithersburg, USA. November 19 - 22: MIT Press (2014)
9. Barzilay, R., Elhadad, N., McKeown, K.R.: Inferring strategies for sentence ordering in multidocument News summarization. J. Artif. Intell. Res. **17**(1), 35–55 (2002)

10. Berger, A., Mittal, V.O.: Query-relevant Summarization using FAQs. In: Proceedings of the 38th annual meeting on association for computational linguistics (ACL 2000), Hong Kong, China. October 03 – 06, pp. 294–301 (2000)
11. Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C.: YAKE! keyword extraction from single documents using multiple local features. Inf. Sci. J. **509**, 257–289 (2020)
12. Catizone, R., Dalli, A., Wilks, Y.: Evaluating automatically generated timelines from the web. In: LREC 2006: Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy. May 24 - 26: ELDA, pp. 885888 (2006)
13. Chen, X., Chan, Z., Gao, S., Yu, M.-H., Zhao, D., Yan, R.: Learning towards Abstractive Timeline Summarization. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), pp. 4939–4945 (2019)
14. Chieu, H.L., Lee, Y.K.: Query based event extraction along a timeline. In: Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR2004), Sheffield, UK. July 25–29, pp. 425–432. ACM (2004)
15. Esteva, A., et al.: Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv preprint arXiv:2006.09595 (2020)
16. Ghalandari, D.G., Ifrim, G.: Examining the state-of-the-art in News timeline summarization. arXiv preprint arXiv:2005.10107 (2020)
17. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document Summarization by Sentence Extraction. In: Proceedings of the Workshop on Automatic summarization (ANLP@NAACL2000), Seattle, Washington. April 30, pp. 40–48 (2000)
18. Hirao, T., Nishino, M., Suziki, J., Nagata, M.: Enumeration of extractive oracle summaries. arXiv preprint arXiv:1701.01614 (2017)
19. Honnibal, M., Montani, I.: spaCy 2: natural language understanding with bloom embeddings. Convolutional Neural Netw. Incremental Parsing **7**(1) (2017)
20. Lin, H., Bilmes, J.: Multi-document summarization via budget maximization of submodular functions. In: Proceedings of Human Language Technologies 2010: The Conference of the North American Chapter of the Association for Computational Linguistc, Los Angeles, pp. 912–920 (2010)
21. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
22. Martschat, S., Markert, K.: Improving {ROUGE} for timeline summarization. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain. April 3–7, pp. 285–290 (2017)
23. Martschat, S., Markert, K.: A temporally sensitive submodularity framework for timeline summarization. In: Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018). Brussels, Belgium. October 31 - November 1: Association for Computational Linguistic, p. 230 (2018)
24. McCreadie, R., Rajput, S., Soboroff, I., Macdonald, C., Ounis, I.: On enhancing the robustness of time-line summarization test collections. Inf. Process. Manage. **56**(5), 18151836 (2019)
25. McCreadie, R., Santos, R.L.T., Macdonald, C., Ounis, I.: Explicit diversification of event aspects for temporal summarization. ACM Trans. Inf. Syst. **36**(3), 1–31 (2018). https://doi.org/10.1145/3158671
26. Minard, A.-L., et al.: SemEval-2015 Task 4: Timeline: cross-document event ordering. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval2015). Denver, USA, June 4–5: Association for Computational Linguistic, pp. 778–786 (2015)
27. Pasquali, A., Mangaravite, V., Campos, R., Jorge, A.M., Jatowt, A.: Interactive system for automatically generating temporal narratives. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11438, pp. 251–255. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_34

28. Tran, G.B., Alrifai, M., Nguyen, D.Q.: Predicting relevant news events for timeline summaries. In: WWW2013 Proceedings of the Companion Publication of the 22nd International Conference on World Wide Web Companion, Rio de Janeiro, Brazil. May 13 – 17, pp. 91–92 (2013)
29. 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
30. Voorhees, E., et al.: TREC-COVID: constructing a pandemic information retrieval test collection. ArXiv abs/2005.04474 (2020)
31. Wang, L., Cardie, C., Marchetti, G.: Socially-informed timeline generation for complex events. In: Proceedings of the Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. Denver, Colorado. May 31-June 5: Association for Computational Linguistic, p. 1055 (2015)
32. Wang, L., et al.: CORD-19: The Covid-19 open research dataset. arXiv:2004.10706v4 (2020)
33. Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., Zhang, Y.: Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In: Proceedings of the 34th International Conference on Research and Development in Information Retrieval (SIGIR 2011). Beijing, China. July 24–28, pp. 745–754. ACM (2011)
34. Yang, W., et al.: On the generation of medical dialogues for COVID19. arXiv:2005.05442v2 (2020)