Cross-lingual timeline summarization

Cross-lingual timeline summarization

Luca Cagliero, Moreno La Quatra, Paolo Garza and Elena Baralis
Dipartimento di Automatica e Informatica
Politecnico di Torino
{luca.cagliero,moreno.laquatra,paolo.garza,elena.baralis}@polito.it

***Abstract*—Timeline summarization methods analyze timestamped, topic-specific news article collections to select the key dates representing the event flow and to extract the most relevant per-date content. Existing approaches are all tailored to a single language. Hence, they are unable to combine topic-related content available in different languages. Enriching news timelines with multilingual content is particularly useful for (i) summarizing complex events, whose main facets are covered differently by media sources from different countries, and (ii) generating news timelines in low-resource languages, for which there is a lack of news content in the target language.**

This paper presents three alternative approaches to address cross-lingual timeline summarization. They combine state-of-the-art extractive summarization methods with machine translation steps at different stages of the timeline generation process. The paper also proposes novel Rouge-based evaluation metrics customized for cross-lingual timeline summarization with a twofold aim: (i) quantifying the ability of the cross-lingual process to enhance available content extraction in the target language and (ii) estimating summarizer effectiveness in conveying additional content from other languages. A new multilingual timeline benchmark dataset has been generated to allow a thorough analysis of the factors that mainly influence summarization performance.

***Index Terms*—Cross-lingual summarization, Timeline Summarization, Natural Language Processing**

I. Introduction

Describing complex real-world events is helpful for various purposes, among which content curation, disaster management, and social media analyses.News events are commonly reported in the form of textual news articles. The news content reflects the context of publication (e.g., country, time period), the current level of knowledge on the covered topic, and the authors' viewpoint. Thus, to capture all the relevant event facets, it is necessary to analyze large news collections published in different countries.

News timelines are concise descriptions of specific events or topics. They consist of a selection of key event dates, enriched with the most salient per-date textual content. TimeLine Summarization (TLS, in short) aims at automating the two main steps in the process, date selection and date summarization, by jointly exploiting Natural Language Processing, Graph Mining, and Deep Learning techniques [1]. Current TLS approaches rely either on supervised techniques (e.g., [2]), which learn predictive models from a set of topic-related news articles annotated with a reference timeline, or on unsupervised approaches (e.g., [3]), which analyze news content and date-level relationships to capture the underlying event aspects.

Existing TLS approaches focus on summarizing collections of news articles that are all written in the same target language. This hinders the use of multilingual content, written in non-target languages, in the date selection and date summarization steps. Analyzing a timestamped collection of multilingual news articles can be beneficial for (i) *summary enrichment*, i.e., selecting new dates or text portions that are not present in the collection written in the target language and (ii) *summary focus*, i.e., leveraging the references to specific events/dates that appear in the non-target language to augment the importance of specific content in the timeline generation for the target language.

For example, let us consider the generation of a news timeline relative to the Covid-19 pandemic for a target language. As an example of enrichment, when the department of health of a foreign country releases new scientific evidence on the effectiveness of a specific medical treatment, this can be pointed out in the news timeline even if the local news agencies of the target country do not report it. As an example of focus, the enforcement of mobility restrictions in many different countries may boost the importance of similar content and focus the summary for the target language on this content.

Existing Cross-Lingual Summarization architectures (e.g., [4]–[6]) cannot address the above issues because (1) they are not designed to tackle the TLS problem, i.e., they ignore news timestamps. (2) They operate on a single-language news source and a single (but different) language for the target summary. Hence, they are not able to handle multiple source languages at the same time.

We propose an approach to jointly address date selection and date summarization in a multilingual context, namely Cross-Lingual TimeLine Summarization (CL-TLS). We present three ad hoc CL-TLS methods, conveniently combine graph ranking, machine translation, and sentence-based summarization. To this end, we also tailor three existing summarization methods, based on sentence-level embedding models, to handle multilingual news content. Furthermore, we quantify the ability of the proposed methods to address timeline focus and enrichment by proposing two novel Rouge evaluation metrics. Finally, we extend an English language TLS benchmark [7] and release a new multilingual version tailored to CL-TLS, namely *ML-Crisis*. We evaluate CL-TLS performance on *ML-Crisis* under multiple aspects and gain insights into the effectiveness and usability of the proposed methods.

a) Contributions: To the best of our knowledge, the present work is the first addressing Cross-Lingual TimeLine Summarization. It tackles the following main issues.

- *Interleaving of machine translation and summarization.* We present three alternative methods to address CL-TLS. They combine graph-based date selection, machine translation, and sentence-based news summarization in different manners. They all overcome the limitations of the existing single-language TLS pipeline, which cannot incorporate multilingual news content.
- *Evaluation metrics.* We propose two novel Rouge-based evaluation metrics tailored to the CL-TLS problem. They specialize timeline evaluation to capture the ability of the CL-TLS to (1) include additional content not present in the news articles written in the target language and (2) enhance the selection of the available content in the target language. We denote the first metrics as ECL-Rouge (related to enrichment) and the second as FCL-Rouge (related to focus).
- *Benchmark dataset.* We release a multilingual benchmark dataset for CL-TLS evaluation. Unlike existing benchmarks, it also includes news articles and reference timelines in languages other than English.

II. Related work

Cross-Lingual Summarization (CLS) entails generating an abstractive summary, for a target language, of a news article collection written in a source language. The key differences between CLS and TLS are enumerated below: (1) CLS source and target languages are different, whereas in TLS they are the same. (2) TLS requires timestamped news articles, whereas in CLS the publication dates are not required. (3) TLS generates a news timeline, consisting of both key dates and per-date summaries. Conversely, in CLS date selection and per-date content selection are out of scope.

The CL-TLS problem presented in this work differs from both CLS and TLS: (1) It takes as input a multilingual set of news articles, including articles written in the target language and not. (2) Similar to TLS (and unlike CLS), it requires timestamped articles. (3) Unlike both TLS and CLS, it addresses date selection and date summarization by also considering the influence of news content written in non-target languages. In a nutshell, it considers summary enrichment and focus due to multilingual content. Hereafter we briefly survey related work in CLS and TLS.

a) Cross-Lingual Summarization: Existing CLS approaches incorporate a machine translation step to translate the news content from the source to the target language. Machine translation can be applied either prior to text summarization (early approach) or after (late approach). For example, in [8] and [9] the authors respectively propose early and late approaches. They integrate a regression-based model to maximize the sentence translation quality. Since the present work relies on unsupervised methods, the aforesaid works are radically different. In [10] the authors address CLS on a bilingual English-Chinese corpus. They exploit graph-based methods that incorporate both language-specific and bilingual sentence-level similarity scores. Unlike [10], the *Late translation* method presented in this paper relies on multilingual contextualized embeddings [11] and integrates also an ad hoc submodular optimization function. More recently, the study presented in [5] optimizes the choice of the summary content by incorporating also the translation quality. To this end, they evaluate text similarity at multiple levels, i.e., single words, sentences, and entire summary.

Deep Learning-based approaches to CLS have recently been proposed. For example, in [12] the authors focus on training a noisy abstractive summarizer for low-resource languages, whereas the work in [13] presents an end-to-end CLS architecture performing summarization and machine translation at the same time. Further research efforts have been devoted to applying multi-task learning in source-to-target summarization [6] and to determining which source words should be translated using attention [4]. To the best of our knowledge, none of the above-mentioned architectures are designed for CL-TLS.

b) TimeLine Summarization: A relevant body of work has been devoted to tackling the TLS problem. The proposed methods can be classified as (1) *Date selection methods*, which specifically address the identification of the most salient dates (e.g., [14], [15]) (2) *Date summarization methods*, whose main goal is to extract the per-date summaries (e.g., [16], [17]), and (3) *Full pipeline methods*, which address both the tasks mentioned above (e.g., [18], [19]). The main contribution of the present work is the integration of multilingual news content to standard TLS. To the best of our knowledge, multilingual news are not addressed in previous TLS approaches.

TABLE I
Summary of notations and their meanings.

$\mathcal{L}$	Set of languages
$\mathcal{Q}$	Set of topics
l^T	Target language in $\mathcal{L}$
$\mathcal{T}$	Reference time period in which the news story happened
$\mathcal{N}_l$	News articles written in language $l \in \mathcal{L}$, published in $\mathcal{T}$ and pertinent to the news story
$\mathbf{D}(\mathcal{N}_l)$	Set of publication dates of the news articles in $\mathcal{N}_l$
$\mathbf{S}(\mathcal{N}_l, d)$	Subset of sentences in $\mathcal{N}_l$ assigned to date $d \in \mathbf{D}(\mathcal{N}_l)$
$\mathcal{R}_l$	Reference timeline for language $l \in \mathcal{L}$
$\mathbf{D}(\mathcal{R}_l)$	Set of dates in the reference timeline $\mathcal{R}_l$
$\mathbf{S}(\mathcal{R}_l, d)$	Subset of sentences in $\mathcal{R}_l$ associated with any article published on date $d \in \mathbf{D}(\mathcal{R}_l)$

III. Problem statement

We formally state the newly proposed cross-lingual timeline summarization problem. For the sake of clarity, the used notation is summarized in Table I.

A news story describes the complex events happened in the reference time period $\mathcal{T}$. A large set of news articles $\mathcal{N}_l$ pertinent to the news story is available for each language $l \in \mathcal{L}$. Our goal is to generate an extractive summary of the main news story events written in a target language $l^T \in \mathcal{L}$.

The traditional TimeLine Summarization (TLS) problem [1] exclusively considers the news content written in the target language, i.e., it disregards all articles in $\mathcal{N}_l, l \neq l_T$. The aim of TLS is twofold: (i) Select the publication dates on which the main events happened, i.e., pick the most representative publication dates from the candidate date set $\mathbf{D}(\mathcal{N}_{l_T})$ associated

with the news articles in $\mathcal{N}_{l_T}$. (ii) Generate a news timeline $\mathbf{TL}(l_T)$ for the target language l^T. It consists of a selection of sentences relative to the news articles associated with the previously selected dates.

This paper presents an extension of the traditional TLS problem, namely *Cross-Lingual TimeLine Summarization* (CL-TLS, in short). CL-TLS aims at enriching the news story summarization process with multilingual content. More specifically, it first selects the most relevant dates from the entire multilingual news article set, i.e., it picks the best representative dates from the extended candidate sets $\mathbf{D}^* = \cup_{l \in \mathcal{L}} \mathbf{D}(\mathcal{N}_l)$. Notice that the candidate dates include also those associated with news articles written in non-target languages. Next, for each selected date $d \in \mathbf{D}^*$ it generates a summary of the date-specific news content by conveying either extractive information for the target language or abstractive content from non-target ones (e.g., translations of sentences extracted from the non-target language corpora).

Adding multilingual content to the TLS pipeline is potentially beneficial for two main reasons. (1) *Improve focus*. It improves the selection quality of the already available content on the target language, especially when the event descriptions are too general and miss relevant event aspects (low-resource news flows). (2) *Enrich content*. it enriches the available content with new dates (and corresponding content) that are missing in the target language.

IV. Cross-Lingual TimeLine Summarization

We present three alternative CL-TLS methods. The key differences between them are in the ways we combine the following three core elements: *the graph ranker, the summarizer, and the machine translator.*

a) Graph ranker: This module takes as input a directed weighted graph $\mathcal{G}$=(V,E), whose vertices V are dates $\mathbf{D}(\mathcal{N}_l)$, $l \in \mathcal{L}$, whereas E is a set of paired vertices weighted by a relevance score. A directed edge $e : d_1 \rightarrow d_2$, $e \in E$, $d_1, d_2 \in V$ connects vertex d_1 to vertex d_2 if a news article published in d_1 contains at least one sentence referencing d_2. For the identification of in-text date references we use HeidelTime temporal tagger [20] whereas the relevance score is computed as the number of date-granularity references from $d_1 \rightarrow d_2$. This module returns a ranking of vertices in V. Previous approaches to traditional TLS consider as relevance scores either the number of explicit date-level references or the date-level content similarity [14].

The date ranking function reflects the authoritativeness of the vertex in the graph-based model. A variety of different functions can be integrated, e.g., vertex in-degree, out-degree, PageRank [21], HITS [22]. Following the guidelines provided in [14], we currently use the date-level references as relevance score and the vertex in-degree as ranking function.

b) Sentence-based summarizer: Summarization algorithms can be either single-language, if they can handle news articles all written in the same language thus can be denoted as the following function Sum_{SL}

$$\mathbf{S}(\mathcal{N}_{l^T}, d) = Sum_{SL}(\mathbf{S}(\mathcal{N}_{l^T}, d))$$

or inherently multilingual, if they are capable of processing article sentences written in different languages and can be denoted as follows

$$\mathbf{S}(\mathcal{N}_{l^T}, d) = Sum_{ML}(\mathbf{S}(\mathcal{N}_{l^T}, d), \mathbf{S}(\mathcal{N}_{l^*}, d), \ldots, \mathbf{S}(\mathcal{N}_{l^{**}}, d))$$

where $l^* \neq l^{**} \neq \ldots \neq l^T$.

To allow sentence-based summarization methods to handle multilingual news content we apply ad hoc modifications to a selection of existing single-language summarizers. Specifically, we propose to tailor the following three existing summarization methods all relying on sentence-level vector representations: (1) SubModular [23], which is based on submodular optimization, (2) Centroid-Opt [19] and (3) EmbeddingRank [24], which perform graph ranking. The key idea is to integrate a recently proposed multilingual sentence embedding representation, namely Sentence-BERT [25]. Each sentence is modelled as a vector in a common latent space, which is characterized by the following properties: (1) semantically related sentences are represented by similar vectors and (2) sentence translations in different languages are represented by aligned vectors. Multilingual vector alignment enables the computation of the pairwise similarity between sentences written in different languages. Therefore, the integration of Sentence-BERT-based vector representations in the aforesaid summarizers enables the handling of news multilingual content.

c) Machine translator: This module takes an arbitrary sentence written in a non-target language and translates it in the target language. Currently, in our implementation we have used the multilingual *mBART* translation system [26]. However, alternative translators can be trivially integrated.

A sketch of the proposed CL-TLS pipelines is given in Figure 1, where we consider the Italian as target language, whereas the news timelines written in French, Spanish, and English are all together used to enrich the news content in the target language. A separate description of each method is given below.

A. Single-language method

The *single language method* (namely *Single*) performs a cascade of the graph ranking and summarization steps to the news articles in the target language. The first step aims at selecting the key publication dates, whereas the last extracts the most salient per-date sentences. Since it ignores the news content written in languages other than the target one, it corresponds to the traditional TLS pipeline.

B. Early translation method

The *early translation method* (*Early*, in short) focuses on translating all the news content in the target language first. To this end, it performs a cascade of machine translation, graph ranking, and summarization. The latter step can be

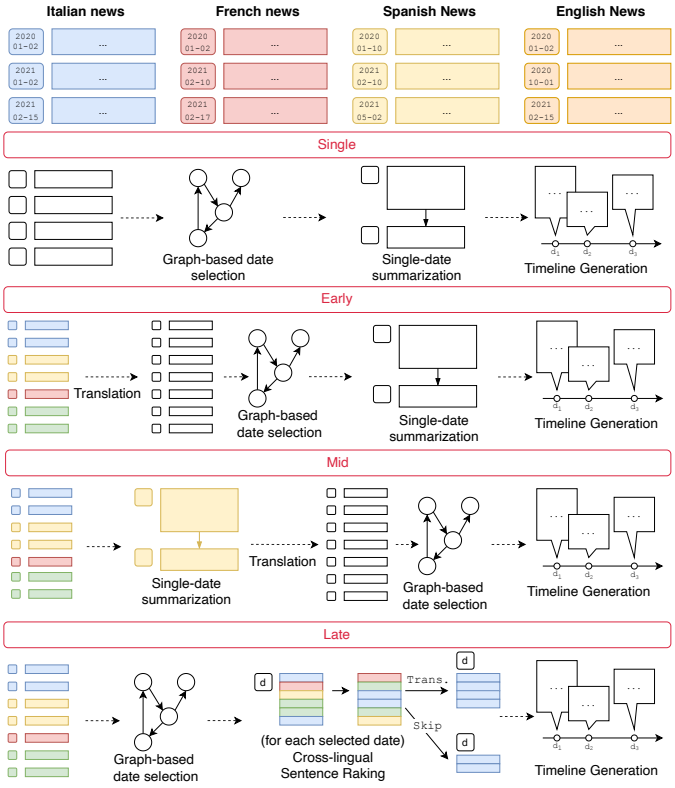

Fig. 1. Sketch of the proposed pipelines.

accomplished by any summarization method as the provided input sentences are all written in the target language.

This method is recommended when the machine translation step is particularly effective in handling the languages and topics covered by the non-target news content. Conversely, if the quality of the translation is not guaranteed, then the summarization step could be biased.

C. Mid translation method

The *mid translation method* (namely *Mid*) first summarizes the news articles published on the same date separately for each language. Next, it performs sentence-level machine translation to all summaries written in a non-target language. Finally, it selects the key timeline dates on top of the translated summaries. To this end, the graph ranking process considers only the date-level references that appear in the summarized content. The motivation behind it is to early discard the references coming from lowly informative news content so that the date selection process relies on a smaller subset of higher-quality references. Notice that the output of the graph ranking step may consist of an arbitrary number of per-summaries sentences. Therefore, in few cases, the (user-specified) maximal summary length can be exceeded. To handle these exceptions, a further stage of date-level content summarization is applied whenever strictly necessary.

D. Late translation methods

To alleviate the effect of machine translation on the summarization step the *late translation method* (namely *Late*) performs machine translation at the latest possible time point. It performs graph ranking followed by per-date multilingual summarization. At the latter stage, the Sentence-BERT-based summarizers are able to compare and rank sentences written in different languages. Finally, the ranked list of news article sentences is explored to produce the output news timeline. To this end, two alternative strategies can be applied: (1) Pick the top ranked sentences written in the target language thus skipping those written in any other language (namely *Late-Skip*) or (2) Pick the top ranked sentences and apply machine translation whenever necessary, i.e. to translate a sentence written in a non-target language but placed at the top of the rank (namely *Late-Translate*).

V. Evaluation Metrics

TLS outcomes are commonly evaluated by comparing them with a hand-written reference timeline provided by a pool of domain experts (i.e., the ground truth). Rouge [27] is the most established toolkit used to quantify the syntactic overlap between the generated timeline and the reference one. It counts the percentage of overlapped textual units. According to the end-user preferences, it supports different metrics (e.g., Rouge-1 for unigrams, Rouge-2 for bigrams, Rouge-L for the longest matching sub-sequence). Separately for each Rouge metric, TLS system performance is quantified by the corresponding precision, recall, and F1-score values. Since in the TLS context both the input news content and the reference summary are timestamped, the timeliness of the selection is usually evaluated by considering the following Rouge variants: *concatenation*, *agreement*, and *alignment* [28]. *Concatenation-based scores* (*concat*, in short) replicate the standard ROUGE evaluation concatenating all the per-date summaries into a unique summary, regardless of the associated timestamp. *Agreement-based scores* (*agreement*) tailor the summary comparisons to the dates that actually occur in the reference timeline. *Alignment-based scores* (namely *align*) rely on an approximated date matching between the selected dates and the reference ones. Whenever a date is missing in the reference timeline, an approximated match is found by considering the closest date in the reference timeline. Then, a penalty is applied to take the inaccuracy in date selection into account during per-date summary evaluation. Hereafter, we will consider the *align+m:1* F1-score proposed in [28].

Similar to TLS, in the CL-TLS scenario the output timeline $\mathbf{TL}(l_T)$ for the target language can be evaluated by comparing it with the reference timeline $\mathcal{R}_l$. However, the benefits from multilingual data analyses in terms of focus and enrichment objectives are unclear.

With the aim at investigating the separate contribution of multilingual data to focus improvement and target enrichment we propose two ad hoc Rouge metrics, namely *Focused Cross-Lingual Rouge* (FCL-Rouge) and *Enriched Cross-Lingual Rouge* (ECL-Rouge).

a) Enriched Cross-Lingual Rouge: ECL-Rouge aims at comparing the per-date summaries in the output and reference timeline by focusing on those publication dates that do not appear in the target language, i.e., in the Rouge score computation it considers only the publication dates s.t.

$$\mathbf{D}_{ECL} = \Big(\mathbf{D}(\mathcal{R}_l) \cap \bigcup_{\{l \in \mathcal{T},\ l \neq l^T\}} \mathbf{D}(\mathcal{N}_l)\Big) / \mathbf{D}(\mathcal{N}_{l^T})$$

The key idea is to disregard the timeline portion that refers to any content that is potentially retrievable from the news articles written in the target language. By removing all the dates appearing in the news articles of the target language, the possible presence of additional dates in the output timeline is exclusively due to their presence in the non-target news content. Therefore, traditional TLS methods are, by construction, unable to get non-zero ECL-Rouge scores whereas CL-TLS ones are capable of exploiting multilingual content to enrich the news timeline.

b) Focused Cross-Lingual Rouge: FCL-Rouge aims at comparing the per-date summaries in the output and reference timeline by focusing exclusively on those dates that appear in the target language, i.e., in the Rouge scores computation it considers the dates s.t.

$$\mathbf{D}_{FCL} = \mathbf{D}(\mathcal{R}_l) \cap \mathbf{D}(\mathcal{N}_{l^T})$$

The idea behind is to specifically study the effect of content enrichment on the target language. Notice that the presence of additional content written in languages other than the target indirectly influences the selection of dates or sentences from the target news content (either positively or negatively).

VI. Benchmark dataset

The performance of TLS methods are commonly evaluated on English-written benchmark datasets. Each dataset consists of a set of source news articles and a set of reference timelines. More in detail

- Timeline 17 [29]: 19 timelines extracted from various news agencies. 9 topics per timeline. Different event types (e.g., catastrophic events or civil wars).
- Crisis [7]: 22 timelines and 4 topics, all related to the long-term armed conflicts happened in North Africa.
- Entities [19]: 47 timelines, each one on a single topic. Most of the covered topics are related to life-spanning events of famous people. The other ones are related to business companies and no-profit organizations.

To the best of our knowledge, no multilingual benchmark for TLS has been released. Hence we present *ML-Crisis*, a multilingual version of the (English-written) Crisis dataset. The dataset is available, for research purposes, upon request to the authors. It consists of 16 timelines, each one written in a different language (Italian, French, Spanish and English). The covered topics and English-written news content correspond to those present in the original *Crisis* dataset[1]. The main dataset properties are summarized in Table II.

TABLE II
ML-Crisis dataset characteristics.

Language	Avg. # articles	Avg. # sentences	Avg. # timeline dates
English	5114.25	33801.75	25.75
Spanish	197.75	7135.0	45.25
French	117.25	6369.5	65.75
Italian	102.75	5448.75	96.25

The key steps of the procedure used to crawl the *ML-Crisis* news data and reference timelines are enumerated below.

1) For each pair of language and topic $(l_i, q_i), l_i \in \mathcal{L}, q_i \in \mathcal{Q}$ we define a seed of manually annotated keywords.
2) For each pair of language and topic we set a reference time period $\mathcal{T}$.
3) We query the GlobalVoices news collection [30] to retrieve the reference timelines separately for each language and topic within the reference period. The articles that include less than two keywords per topic are discarded.
4) We query the Google News crawler[2] by properly setting the date range filter to retrieve the multilingual news content within the selected time range. To cover the news story at best, in the Google News query the starting and ending dates of the reference period have been conveniently extended by ten days.

VII. Experiments

We ran a set of experiments on the *ML-Crisis* benchmark with the goal of (i) empirically exploring the effect of integrating multilingual news content into a target collection, (ii) evaluating timeline summarization performance using classical Rouge evaluation metrics, and (iii) quantifying summary enrichment and focus using ad hoc Rouge metrics.

A. Experimental setup

a) Hardware: Experiments were run on a machine equipped with AMD® Ryzen 9® 3950X CPU, Nvidia® RTX 3090 GPU, 128 GB of RAM and running Ubuntu 20.04 LTS.

b) Evaluation metrics: In compliance with previous studies on TLS (e.g., [3]), we collect the traditional Rouge-1 and Rouge-2 precision, recall, and F1-score for the following metrics: *concatenation*, *agreement*, and *alignment*. Furthermore, to quantify the level of enrichment and focus due to CL-TLS we collect the metrics for ECL-Rouge and FCL-Rouge as well (see Section V).

c) Summarizers: We test the following state-of-the-art summarizers using the configuration settings recommended by the respective authors.

Single-language models: (1) TextRank-BM25 [31]: Established graph-based summarization methods that relies on the the Okapi-BM25 text similarity score [32]. (2) CoreRank [33]:

[1]To ensure consistency among languages, we considered a single reference timeline per topic for the English language

[2]https://www.news.google.com (latest access: May 2021)

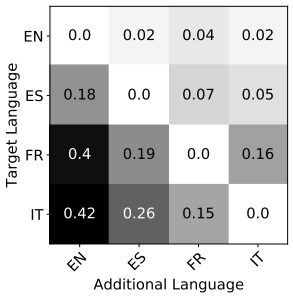

Fig. 2. Contribution of additional languages: date enrichment level (%).

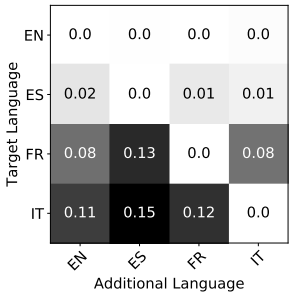

Fig. 3. Contribution of additional languages: date enrichment level (weighted %).

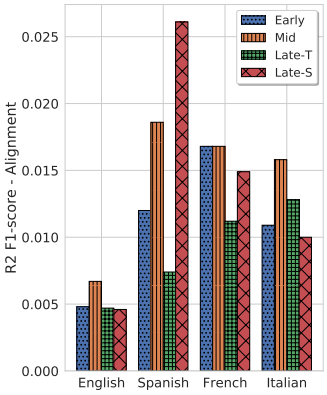

Fig. 4. FCL-Rouge scores.

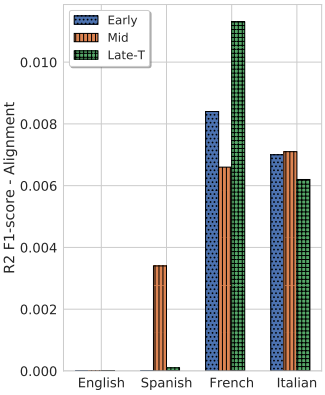

Fig. 5. ECL-Rouge scores.

a recently proposed summarizer that combines submodular optimization with graph-based text modelling. (3) ELSA [34]: a recently proposed itemset- and LSA-based summarization system. (4) SubModular [23], which relies on submodular optimization, (5) Centroid-Opt [19] and (6) EmbeddingRank [24].

Models tailored to the cross-lingual context: (1) SubModular [23], (2) Centroid-Opt [19], and (3) EmbeddingRank [24].

d) Execution time: We report here the average execution time taken by the core CL-TLS modules: (i) Graph ranking: 0.2-1.38s. (ii) Summarization (per single date): 2.30-5.69s. (iii) Machine translation (per sentence): 0.07s.

B. Dataset exploration

To our purposes, we explore the date- and content-level relationships between the news articles written in different languages. We quantify the percentage of dates in the reference timeline added to the target collection separately by each additional language. Specifically, the heatmap in Figure 2 reports, for each target language, the fraction of new reference timeline dates added by separately considering each language-specific collection. It answers the following question: *how many new reference dates could be potentially revealed when adding an additional language to the target one?*. The contribution is maximal for the Italian language (e.g., adding English-written news content yields a 42% date enrichment), fair for French, low for Spanish, and very low for English. The topic-related news flows in Italian and French can be deemed to be *low-resource*. This is confirmed by the significantly lower number of available articles and sentences compared to the English-written collection (see Table II).

The heatmap in Figure 3 shows a variant of the aforesaid statistics. Each new date is weighted by the number of added sentences. The rationale is that the additional dates on which many articles are published are deemed as the most relevant ones for summarization purposes. The comparison between the weighted and unweighted statistic confirms the knowledge gap between Italian/French and English/Spanish languages in the *ML-Crisis* dataset.

C. Summary evaluation using standard Rouge

Tables III-VI compare the standard Rouge results on *ML-Crisis* separately for each target language. As expected, the benefits for adding multilingual content to the Spanish and English collections are limited. Conversely, in Italian and French the early and mid translation methods outperform the single-language method.

We also run a statistical test to verify the significance of the difference in performance between the CL-TLS methods using the two-sided paired approximate randomization test [35]. To this end, we compare the outcomes achieved by the summarizers available in all methods, i.e., SubModular [23], Centroid-Opt [19], and EmbeddingRank [24]. The results confirm that *Mid* and *Early* methods perform significantly better than all the other methods in terms of Rouge-1 Alignment (with significance level 0.05) for the Italian and French languages, respectively.

D. Summary evaluation using CL-Rouge metrics

In Table VII we compare the ECL- and FCL-Rouge outcomes achieved by a representative method (*Early*) and summarizer (TextRank-BM25) pair. The obtained levels of enrichment and focus are practically zero for the English language, whereas are fairly relevant for Spanish, French, and Italian. Notably, for the Spanish target incorporating the multilingual content is beneficial even if the percentage of newly added dates is rather low. Note that such a positive effect cannot be highlighted by using the Standard Rouge metrics.

Figures 4 and 5 compare similar results obtained by different methods (the *Late-Skip* method is missing for ECL-Rouge because, by construction, it cannot achieve non-zero values). Unlike with traditional Rouge metrics, we achieve relevant FCL- and ECL-Rouge scores' improvements with the *Late-Translate* method for the Spanish and French targets, respectively.

VIII. Conclusions

We address the novel problem of cross language timeline summarization (CL-TLS). It focuses on extracting a timeline for a topic-specific news article collection by also considering related news content written in languages other than the target language. Our approach allows enriching the generated timeline summary with novel information derived from

TABLE III
CL-TLS ROUGE RESULTS ON *ML-Crisis*. TARGET LANGUAGE: ITALIAN.

Summarizer	Date F1	Concat R1	Concat R2	Agreement R1	Agreement R2	Alignment R1	Alignment R2
Single							
ELSA	0.2535	0.3517	0.0779	0.0296	0.0041	0.0441	0.0051
TextRank-BM25	0.2535	0.3496	0.0794	0.0276	0.0041	0.0420	0.0051
CoreRank	0.2535	0.3253	0.0742	0.0282	0.0042	0.0414	0.0052
EmbeddingRank	0.2535	0.3587	0.0810	0.0298	0.0045	0.0444	0.0055
Centroid-opt	0.2535	0.3502	0.0771	0.0272	0.0037	0.0417	0.0049
Submodular	0.2535	0.3511	0.0781	0.0269	0.0037	0.0410	0.0048
Early							
ELSA	0.2688	0.4855	0.1567	0.0591	0.0076	0.0766	0.0091
TextRank-BM25	0.2688	0.4816	**0.1659**	0.0651	0.0108	0.0849	0.0131
CoreRank	0.2688	0.5043	0.1652	0.0604	0.0090	0.0806	0.0113
EmbeddingRank	0.2688	0.4550	0.1545	0.0586	0.0078	0.0763	0.0095
Centroid-opt	0.2688	0.4783	0.1615	0.0595	0.0078	0.0778	0.0095
Submodular	0.2688	0.4657	0.1584	0.0594	0.0080	0.0767	0.0097
Mid							
ELSA	0.2929	0.4851	0.1571	0.0609	0.0073	0.0802	0.0092
TextRank-BM25	0.3113	0.4818	0.1651	0.0706	**0.0122**	0.0898	**0.0143**
CoreRank	0.3133	**0.5211**	0.1636	0.0628	0.0095	0.0857	0.0116
EmbeddingRank	**0.3198**	0.4586	0.1585	0.0713	0.0109	0.0882	0.0130
Centroid-opt	0.3114	0.4856	0.1632	**0.0726**	0.0110	**0.0919**	0.0129
Submodular	0.3002	0.4745	0.1608	0.0660	0.0104	0.0855	0.0125
Late Translate							
EmbeddingRank	0.2688	0.5062	0.1597	0.0590	0.0083	0.0760	0.0100
Centroid-opt	0.2688	0.5056	0.1601	0.0591	0.0083	0.0760	0.0100
Submodular	0.2688	0.5065	0.1602	0.0591	0.0083	0.0761	0.0100
Late Skip							
EmbeddingRank	0.2688	0.3093	0.0757	0.0338	0.0058	0.0427	0.0064
Centroid-opt	0.2688	0.3078	0.0755	0.0336	0.0058	0.0427	0.0064
Submodular	0.2688	0.3104	0.0759	0.0337	0.0058	0.0427	0.0064

multilingual content. We propose different CL-TLS methods and two novel Rouge-based metrics used to evaluate CL-TLS performance under various perspectives. The most important takeaways from the empirical evidence achieved on a newly released benchmark dataset are as follows. (1) The scarser the topic-related news collection in the target language (e.g., Italian), the clearer the benefits from the integration of multilingual news content. (2) Traditional Rouge scores are unable to quantify the value added by multilingual data integration, whereas the ECL- and FCL-Rouge metrics provide deeper insights into this specific aspect. (3) The summarization methods that rely on sentence embedding representation can be adapted to postpone machine translation after text summarization. As highlighted by the ECL-Rouge score, this change is beneficial for content enrichment. Future extensions of the current work will address the integration of abstractive summarization models in the CL-TLS pipeline.

REFERENCES

[1] R. Swan and J. Allan, "Automatic generation of overview timelines," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '00. New York, NY, USA: ACM, 2000, p. 49–56.

[2] D. G. Ghalandari and G. Ifrim, "Examining the state-of-the-art in news timeline summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*. ACL, 2020, pp. 1322–1334.

[3] S. Martschat and K. Markert, "A temporally sensitive submodularity framework for timeline summarization," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: ACL, 2018, pp. 230–240.

[4] J. Zhu, Y. Zhou, J. Zhang, and C. Zong, "Attend, translate and summarize: An efficient method for neural cross-lingual summarization," in *Proceedings of the 58th Annual Meeting of the ACL*. ACL, 2020, pp. 1309–1321.

[5] X. Wan, F. Luo, X. Sun, S. Huang, and J.-g. Yao, "Cross-language document summarization via extraction and ranking of multiple summaries," *Knowledge and Information Systems*, vol. 58, no. 2, pp. 481–499, 2019.

[6] R. Xu, C. Zhu, Y. Shi, M. Zeng, and X. Huang, "Mixed-lingual pretraining for cross-lingual summarization," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the ACL and the 10th International Joint Conference on Natural Language Processing*. Suzhou, China: ACL, 2020, pp. 536–541.

[7] G. B. Tran, T. A. Tran, N.-K. Tran, M. Alrifai, and N. Kanhabua, "Leveraging learning to rank in an optimization framework for timeline summarization," in *Workshop on Time-aware Information Access*, 2013.

[8] F. Boudin, S. Huet, and J.-M. Torres-Moreno, "A graph-based approach to cross-language multi-document summarization," *Polibits*, no. 43, pp. 113–118, 2011.

[9] X. Wan, H. Li, and J. Xiao, "Cross-language document summarization based on machine translation quality prediction," in *Proceedings of the 48th Annual Meeting of the ACL*, 2010, pp. 917–926.

[10] X. Wan, "Using bilingual information for cross-language document summarization," in *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, 2011, pp. 1546–1555.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: ACL, 2019, pp. 3982–3992.

[12] J. Ouyang, B. Song, and K. McKeown, "A robust abstractive system for cross-lingual summarization," in *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL, 2019, pp. 2025–2031.

[13] J. Zhu, Q. Wang, Y. Wang, Y. Zhou, J. Zhang, S. Wang, and C. Zong, "NCLS: Neural cross-lingual summarization," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL, 2019, pp. 3054–3064.

[14] R. Kessler, X. Tannier, C. Hagège, V. Moriceau, and A. Bittar, "Finding salient dates for building thematic timelines," in *Proceedings of the 50th Annual Meeting of the ACL)*. ACL, 2012, pp. 730–739.

[15] G. Tran, E. Herder, and K. Markert, "Joint graphical models for date selection in timeline summarization," in *Proceedings of the 53rd Annual*

TABLE IV
CL-TLS ROUGE RESULTS ON *ML-Crisis*. TARGET LANGUAGE: FRENCH.

Summarizer	Date F1	Concat R1	Concat R2	Agreement R1	Agreement R2	Alignment R1	Alignment R2
Single							
ELSA	0.2999	0.4235	0.1222	0.0456	0.0086	0.0654	0.0114
TextRank-BM25	0.2999	0.4315	0.1299	0.0505	0.0112	0.0714	0.0144
CoreRank	0.2999	0.4259	0.1233	0.0440	0.0077	0.0636	0.0102
EmbeddingRank	0.2999	0.4108	0.1224	0.0432	0.0082	0.0621	0.0106
Centroid-opt	0.2999	0.4242	0.1233	0.0444	0.0083	0.0640	0.0107
Submodular	0.2999	0.4186	0.1228	0.0453	0.0096	0.0648	0.0122
Early							
ELSA	**0.3446**	0.4317	0.1423	0.0709	0.0089	0.0989	0.0117
TextRank-BM25	**0.3446**	0.4070	0.1515	**0.0823**	**0.0172**	**0.1143**	**0.0219**
CoreRank	**0.3446**	0.4468	0.1500	0.0773	0.0134	0.1075	0.0172
EmbeddingRank	**0.3446**	0.3767	0.1366	0.0708	0.0092	0.0999	0.0132
Centroid-opt	**0.3446**	0.3945	0.1406	0.0721	0.0102	0.1011	0.0137
Submodular	**0.3446**	0.3811	0.1353	0.0713	0.0095	0.1003	0.0131
Mid							
ELSA	0.2651	0.4226	0.1399	0.0540	0.0077	0.0788	0.0101
TextRank-BM25	0.2694	0.4229	**0.1556**	0.0674	0.0156	0.0971	0.0199
CoreRank	0.2837	**0.4479**	0.1452	0.0657	0.0127	0.0927	0.0163
EmbeddingRank	0.2430	0.3833	0.1415	0.0540	0.0113	0.0829	0.0153
Centroid-opt	0.2344	0.4057	0.1419	0.0528	0.0096	0.0816	0.0131
Submodular	0.2535	0.3945	0.1423	0.0558	0.0107	0.0829	0.0142
Late Translate							
EmbeddingRank	**0.3446**	0.4346	0.1333	0.0715	0.0093	0.0991	0.0122
Centroid-opt	**0.3446**	0.4365	0.1337	0.0714	0.0091	0.0993	0.0120
Submodular	**0.3446**	0.4350	0.1334	0.0715	0.0092	0.0991	0.0121
Late Skip							
EmbeddingRank	**0.3446**	0.3896	0.0990	0.0540	0.0080	0.0721	0.0099
Centroid-opt	**0.3446**	0.3917	0.0996	0.0540	0.0080	0.0720	0.0099
Submodular	**0.3446**	0.3915	0.0997	0.0540	0.0080	0.0720	0.0099

TABLE V
CL-TLS ROUGE RESULTS ON *ML-Crisis*. TARGET LANGUAGE: SPANISH.

Summarizer	Date F1	Concat R1	Concat R2	Agreement R1	Agreement R2	Alignment R1	Alignment R2
Single							
ELSA	**0.3478**	0.4104	0.1394	0.0849	0.0302	0.1010	0.0325
TextRank-BM25	**0.3478**	0.4114	**0.1441**	**0.0935**	**0.0338**	**0.1118**	**0.0374**
CoreRank	**0.3478**	0.4242	0.1383	0.0821	0.0264	0.0998	0.0294
EmbeddingRank	**0.3478**	0.4002	0.1332	0.0873	0.0321	0.1038	0.0345
Centroid-opt	**0.3478**	0.4049	0.1318	0.0875	0.0277	0.1040	0.0300
Submodular	**0.3478**	0.3934	0.1352	0.0845	0.0314	0.1011	0.0339
Early							
ELSA	0.3063	0.3108	0.1030	0.0564	0.0097	0.0708	0.0115
TextRank-BM25	0.3063	0.3345	0.1136	0.0669	0.0149	0.0860	0.0182
CoreRank	0.3063	0.3627	0.1125	0.0596	0.0097	0.0775	0.0125
EmbeddingRank	0.3063	0.3068	0.1072	0.0565	0.0089	0.0726	0.0112
Centroid-opt	0.3063	0.3093	0.1034	0.0580	0.0090	0.0742	0.0115
Submodular	0.3063	0.3089	0.1089	0.0577	0.0095	0.0742	0.0121
Mid							
ELSA	0.2659	0.3356	0.1180	0.0547	0.0196	0.0752	0.0218
TextRank-BM25	0.2467	0.3577	0.1307	0.0579	0.0209	0.0796	0.0239
CoreRank	0.2561	0.3390	0.1130	0.0479	0.0134	0.0711	0.0160
EmbeddingRank	0.2806	0.3125	0.1131	0.0564	0.0142	0.0773	0.0171
Centroid-opt	0.2567	0.3146	0.1112	0.0445	0.0081	0.0657	0.0107
Submodular	0.2640	0.3098	0.1086	0.0529	0.0129	0.0726	0.0157
Late Translate							
EmbeddingRank	0.3063	0.3665	0.1094	0.0581	0.0083	0.0732	0.0099
Centroid-opt	0.3063	0.3670	0.1096	0.0582	0.0084	0.0731	0.0099
Submodular	0.3063	0.3668	0.1095	0.0584	0.0083	0.0735	0.0099
Late Skip							
EmbeddingRank	0.3063	0.4301	0.1292	0.0779	0.0234	0.0936	0.0256
Centroid-opt	0.3063	**0.4319**	0.1310	0.0792	0.0251	0.0950	0.0273
Submodular	0.3063	0.4306	0.1292	0.0779	0.0235	0.0936	0.0256

TABLE VI
CL-TLS ROUGE RESULTS ON *ML-Crisis*. TARGET LANGUAGE: ENGLISH.

Summarizer	Date F1	Concat R1	Concat R2	Agreement R1	Agreement R2	Alignment R1	Alignment R2
Single							
ELSA	0.2254	0.3018	0.0460	0.0323	0.0037	0.0420	0.0044
TextRank-BM25	0.2254	0.3238	0.0568	**0.0413**	0.0075	0.0522	0.0088
CoreRank	0.2254	0.2994	**0.0575**	0.0362	0.0070	0.0490	0.0090
EmbeddingRank	0.2254	0.2869	0.0481	0.0369	0.0051	0.0480	0.0064
Centroid-opt	0.2254	0.3079	0.0521	0.0373	0.0053	0.0491	0.0069
Submodular	0.2254	0.2942	0.0508	0.0369	0.0053	0.0486	0.0066
Early							
ELSA	**0.2298**	0.2807	0.0374	0.0285	0.0049	0.0357	0.0054
TextRank-BM25	**0.2298**	0.3213	0.0516	0.0410	0.0077	**0.0538**	0.0092
CoreRank	**0.2298**	0.2823	0.0549	0.0387	**0.0090**	0.0499	0.0107
EmbeddingRank	**0.2298**	0.2868	0.0485	0.0347	0.0040	0.0453	0.0048
Centroid-opt	**0.2298**	0.2955	0.0481	0.0369	0.0050	0.0470	0.0063
Submodular	**0.2298**	0.2889	0.0496	0.0344	0.0054	0.0454	0.0068
Mid							
ELSA	0.1829	0.3003	0.0462	0.0205	0.0026	0.0320	0.0048
TextRank-BM25	0.2082	**0.3246**	0.0570	0.0357	0.0083	0.0478	**0.0112**
CoreRank	0.1568	0.2926	0.0483	0.0227	0.0047	0.0360	0.0081
EmbeddingRank	0.1707	0.2960	0.0451	0.0253	0.0050	0.0406	0.0073
Centroid-opt	0.1445	0.3063	0.0440	0.0201	0.0041	0.0314	0.0059
Submodular	0.1637	0.2985	0.0456	0.0204	0.0038	0.0348	0.0058
Late Translate							
EmbeddingRank	**0.2298**	0.2915	0.0427	0.0222	0.0037	0.0317	0.0047
Centroid-opt	**0.2298**	0.2931	0.0423	0.0221	0.0037	0.0317	0.0047
Submodular	**0.2298**	0.2920	0.0428	0.0223	0.0038	0.0319	0.0048
Late Skip							
EmbeddingRank	**0.2298**	0.2839	0.0397	0.0236	0.0041	0.0321	0.0046
Centroid-opt	**0.2298**	0.2842	0.0396	0.0234	0.0041	0.0319	0.0046
Submodular	**0.2298**	0.2842	0.0398	0.0237	0.0041	0.0322	0.0046

TABLE VII
ECL- AND FCL-ROUGE RESULTS. *Early* METHOD. TEXTRANK-BM25 SUMMARIZER.

Target language	Date F1		Concat R2		Agreement R2		Alignment R2	
	ECL	FCL	ECL	FCL	ECL	FCL	ECL	FCL
English	0.0	0.2298	0.0	0.0337	0.0	0.0049	0.0	0.0054
Spanish	0.0526	0.3440	0.0604	0.1008	0.0148	0.0226	0.0181	0.0240
French	0.2007	0.3869	0.1082	0.1088	0.0112	0.0187	0.0158	0.0219
Italian	0.1802	0.3648	0.1369	0.1070	0.0072	0.0133	0.0092	0.0147

Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: ACL, 2015, pp. 1598–1607.

[16] H. L. Chieu and Y. K. Lee, "Query based event extraction along a timeline," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '04. New York, NY, USA: ACM, 2004, p. 425–432.

[17] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang, "Evolutionary timeline summarization: A balanced optimization framework via iterative substitution," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11. ACM, 2011, p. 745–754.

[18] G. Binh Tran, M. Alrifai, and D. Quoc Nguyen, "Predicting relevant news events for timeline summaries," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 91–92.

[19] D. Gholipour Ghalandari, "Revisiting the centroid-based method: A strong baseline for multi-document summarization," in *Proceedings of the Workshop on New Frontiers in Summarization*. Copenhagen, Denmark: ACL, 2017, pp. 85–90.

[20] J. Strötgen and M. Gertz, "Multilingual and cross-domain temporal tagging," *Language Resources and Evaluation*, vol. 47, no. 2, pp. 269–298, 2013.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[22] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[23] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, 2011, pp. 510–520.

[24] H. Zheng and M. Lapata, "Sentence centrality revisited for unsupervised summarization," in *Proceedings of the 57th Annual Meeting of the ACL*. Florence, Italy: ACL, 2019, pp. 6236–6247.

[25] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2020, pp. 4512–4525.

[26] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the ACL*, vol. 8, pp. 726–742, 2020.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. ACL, 2004, pp. 74–81.

[28] S. Martschat and K. Markert, "Improving ROUGE for timeline summarization," in *Proceedings of the 15th Conference of the European Chapter of the ACL: Volume 2*. ACL, 2017, pp. 285–290.

[29] G. Tran, M. Alrifai, and E. Herder, "Timeline summarization from relevant headlines," in *European Conference on Information Retrieval*. Springer, 2015, pp. 245–256.

[30] K. Nguyen and H. Daumé III, "Global Voices: Crossing borders in automatic news summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. ACL, 2019, pp. 90–97.

[31] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.

[32] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford *et al.*, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[33] A. Tixier, P. Meladianos, and M. Vazirgiannis, "Combining graph degeneracy and submodularity for unsupervised extractive summarization," in *Proceedings of the workshop on new frontiers in summarization*, 2017, pp. 48–58.

[34] L. Cagliero, P. Garza, and E. Baralis, "Elsa: A multilingual document summarization algorithm based on frequent itemsets and latent semantic analysis," *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 2, pp. 1–33, 2019.

[35] E. W. Noreen, *Computer-intensive methods for testing hypotheses*. Wiley New York, 1989.