Expanding News Timeline Summarization

Ross Devito, William Hogan, Tianyang Zhang
Jacobs School of Engineering
University of California, San Diego
rdevito@ucsd.edu, whogan@ucsd.edu, tiz010@ucsd.edu

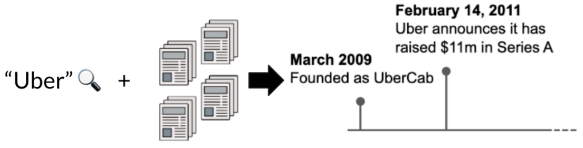
Abstract

News timeline summarization (TLS) is the task of creating a timeline of important events for a topic given its keywords and a set of documents on that topic. The resulting timelines include the most important dates and a short description of the key events that happened on that date, similar to the way a human editor may outline a topic. In the following work, we expand news timeline summarization (TLS) with three main contributions: expanding the available datasets for training and evaluation, introducing more representative evaluation metrics, and improving on existing state-of-the-art date-wise and clustering TLS approaches.

1 Introduction

News timeline summarization is the process of summarizing news topics into a sequence of events with corresponding summaries. It takes an input of a topic query (key phrases, e.g. "Uber"), and a collection of articles that match the topic query, and outputs a list of $\langle date, summary \rangle$ pairs which, combine to form a news timeline summary. The outputs are constrained by length of l dates and m sentences, or $k = m/l$ sentences per date.

Figure 1: Visualized inputs and outputs for the news timeline summarization task. A topic query and a corresponding collection of articles are used to create a timeline of dates and summaries that are constrained by length.



1.1 Why is this problem important?

A single paper or report only reveals a few dimensions of a complex story. The comprehensive depiction of a real world event usually lies behind a combination of multiple reports from different perspectives. With troves of news articles are released everyday from hundreds of news sources around the world, this task can be challenging for humans to summarize a sequence of events within a specific news topic. Either one or more sources can be accidentally overlooked or misinterpreted, or a sequence of sub-events can be obtained in fallacious order. This can lead to a summary that misses some crucial components of a story and eventually result in a misrepresented event. For this problem, we turn to machine learning and, specifically, deep-learning models that generate concise and accurate news timeline summaries. By improving the news timeline summarization, we can depict the story without the potential of losing important information.

1.2 Current shortcomings of news TLS

Current efforts in news timeline summarization struggle in three distinct areas: available data for training, metrics that are truly representative of timeline summary quality, and, generally, in model performance. There are currently three datasets used for news TLS, as described in Section 3. Each are relatively small in size which makes training advanced TLS models difficult. Additionally, the current metrics used to establish the state-of-the-art in news TLS are a combination of ROUGE scores to assess summary quality, and F1 scores to assess accuracy of date selection. However, as we discuss in more detail below, these metrics are flawed and lead to models that produce repetitious summaries. Furthermore, the current state-of-the-art (SOTA) news timeline summaries do not achieve a level of performance to be useful to humans. In our work, we attempt to address each of these shortcomings.

1.3 Our contributions

Our contributions to news TLS are the following:

- We create two new datasets, namely *Entities New* and *Topics*, which help expand the available data to train and evaluate news TLS.
- We introduce new metrics to the task that better indicate news TLS quality.
- We make improvements on two news TLS methods and establish new state-of-the-art performance.

2 Related Work

There are currently three main approaches used in news timeline summarization: one-stage summarization, clustering summarization, and date-wise summarization. One-stage summarization, also known as *direct summarization*, is proposed by Martschat et al.(2018). he approach models the problem similar to a single-document summarization problem. It combines a collection of articles into a single article. A timeline summary is generated directly from the combined articles. The clustering method implemented by Ghalandari et al. (2020) detects events by clustering similar articles together. Once clustered, the method identifies the l most important clusters (events) and then summarizes each cluster separately. The date-wise method implemented by both Tran et al. (2015b) and Ghalandari et al. (2020) first selects l important dates. The list of important dates are used to trim the collection of articles into a subset articles that are either published on or mention l dates. With this subset of articles, it generates a summary for the top n important dates.

Each of these approaches are outlined and evaluated head to head by Ghalandari et al. (2020). Ghalandari et al. establish their date-wise method as the state-of-the-art (SOTA) in news TLS. Prior to this paper, the SOTA was a one-stage summarization approach (Martschat and Markert, 2018). In addition to scalability and runtime benefits, the date-wise method was shown to outperform the direct summarization method in all existing metrics and across all test sets.

In this work, we focus our efforts on improving both clustering and date-wise methods. We discuss our proposed improvements in Section 4.

3 Datasets

Previous to our dataset expansion efforts, the following three datasets were the only datasets available for news TLS:

"Entites" Dataset: The "Entities" dataset is a collection of 45k news articles and 47 timeline summaries which correspond to 47 unique news topics (Gholipour Ghalandari and Ifrim, 2020).

"T17" Dataset: The "T17" dataset is a collection of 4.5k news articles, 19 timeline summaries, and 9 unique news topics (t17, 2013).

"Crisis" Dataset: The "Crisis" dataset is a collection of 9.2k news articles, 22 timeline summaries, and 4 unique news topics (Tran et al., 2015a).

Table 2 helps to illustrate the considerable variance across all three datasets. The *Entities* dataset has the most topics (47) and timelines (47). It also has the longest average of timeline length. *Crisis* has the most articles per topic (2310), and *T17* has the most sentences per summary (2.9).

4 Methods

4.1 Dataset Expansion

There are three existing datasets that can be used to evaluate our model. Trans et al. provided 17 Timelines (T17) (t17, 2013) and the Crisis (Tran et al., 2015a) datasets. Ghalandari et al. introduced a better dataset with more topics and longer time duration for each topics (Gholipour Ghalandari and Ifrim, 2020). However, there are two problems remaining on the dataset: 1. the amount of total topics is still relatively small for both training and evaluation 2. The articles inside the datasets is not closely related to the ground truth timeline, which resulted in a lot of noise. Therefore, we extended the dataset with two different approachs and ends up in two new collected datasets: *Topics* and *Entities-new*.

4.1.1 Old Methodology

Firstly we obtained a completely new dataset based on the methodology provided in the previous work (Gholipour Ghalandari and Ifrim, 2020).

Ground-Truth Timelines: The ground-truth timeline for the new dataset is also obtained from CNN Fast Facts. But, unlike the *Entities* mainly focus on the category of people, our new dataset *Topics* coveres a broad range of type of topics, from history event to a geographic region, and each of them have considerable length of time duration.

Queries: Unlike the People category, there are no surnames exists in our new topics to use as the query keyphrases Q. Therefore, we use the full

	Topics	TLs	Avg Articles	Avg # Sents/Summary	Avg TL-Dates	Avg Time Span
Crisis	4	22	**2310**	1.3	29	343 days
T17	9	19	508	**2.9**	**36**	212 days
Entities	**47**	**47**	959	1.2	23	**12 years**

Table 1: Statistics on the three news TLS datasets. *Entities* has the most timelines and topics, and the longest average timeline length. *Crisis* has the most articles per topic, and *T17* has the shortest average timeline length.

name of the event or location instead. It is usually the default title of the timeline.

Input Articles: TheGuardian API is used to obtain news articles related to the timeline. TheGuardian is continuously populating its news collection and it currently provides access to news articles published from 1998. We firstly use the query to perform a search on the API, then search for articles that contain exactly match of query keywords in the body text.

Adjustments and Filtering: Just like the methodology in the previous work, we modify and drop some of the Timelines to make sure they are usable for TLS:

- Timeline entries with no specific year, month, and day are removed.
- Entries that outside the date coverage of articles input is also removed.
- If there are no articles published within 2 days before/after the timeline date, then it is also removed.

We also eliminate the topics that does not fulfill requirements below:

- Each topic must have a timeline with at least 5 entries.
- At least a half of the timeline entries, there should be a textual reference in the article input set.
- Each topic should have at least 100 and at most 3000 articles.

Based on this old methodology, we obtained 35 new topic with 35 new ground-truth timelines, and 39708 new articles. There is an average of 1135 articles per topic.

4.1.2 New Methodology

To increase the coherence of the articles and the ground-truth timelines, we invented some new rules in addition to the old methodology.

Ground-Truth Timelines: We use the ground-truth Timelines in the existing *Entities* dataset to obtain a comparability between old and new methods.

Queries: Unlike the previous work, we use human labor to manually label keyword and phrases for each timeline entries, and eventually result in a more accurate query phrase. We split the giant single query from the old methodology to multiple sub-queries, one for each timeline entry. The date duration for each query was adjusted to 14 days before/after the timeline date.

Input Articles: Still using TheGuardian API for obtaining articles. Each timeline entry query is used to search for articles, and all articles that does not have a exact match of query keywords are eliminated. The first 100 articles are obtained for each entry.

Adjustments and Filtering: Some of the timeline entries are dropped or modified to fit the new methodology.

- Timeline entries with empty query feedback from TheGuardian API is removed.
- Timeline entries with date prior than 1998 is also removed.

After collecting all the articles, topics that does not match the following criteria is removed.

- Each topic must have 5 or more timeline entries that have a non-empty article set.
- Each topic should have less than 3000 and more than 100 articles.

By using the new methodology on the old *Entities* timeline ground-truth, we obtained a new set of articles for each of the 46 topics. 23334 articles are collected for these topics and there are an average of 507 articles per topic.

	Topics	TLs	Avg Articles	Avg Sentences	Avg TL-Dates	Avg Time Span
Entities-new	**46**	**46**	507	20787	23	12 years
Topics	35	35	**1135**	**62425**	**24**	**14 years**

Table 2: Statistics on the two news TLS datasets.

4.2 New Evaluation Metrics

4.2.1 Current Evaluation Metrics

Currently, the standard for evaluating news TLS performance is to use F1-score to evaluate date selection and event aligned ROUGE-1 and ROUGE-2 F1-scores for text quality (Martschat and Markert, 2018; Gholipour Ghalandari and Ifrim, 2020).

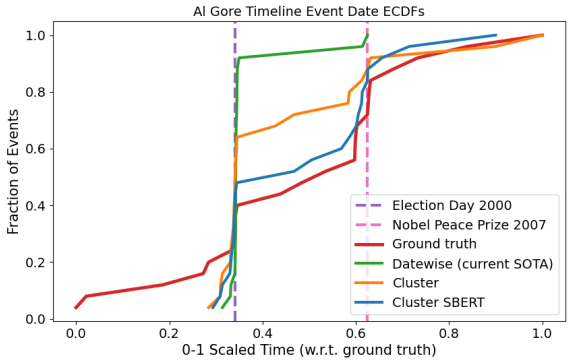


Figure 2: Empirical cumulative distribution functions for event dates in ground truth and predicted timelines on Al Gore.

The date F1-score only considers exact date matches to be true positives. This all or nothing reward, combined with the general difficulty of guessing correct dates, leads to models which boost their number of correct dates by predicting a burst of many events around the day of a ground true event. This burst behavior is shown in Figure 2 by the current SOTA date-wise model. While the ground truth timeline for Al Gore dedicates less than 20% of its events to the time immediately around the 2000 election, the date-wise model dedicates nearly 90% of its events to the same period. The maximum number of events per predicted timeline is limited to the number of events on the ground truth timeline, so over predicting events in this way means that there will be important time periods and ground true events left off the timeline. Despite this undesirable behavior, the date-wise model achieves a date F1 of 0.20 for this example compared to 0.12 and 0.08 for the cluster and cluster SBERT models respectively.

The aligned ROUGE scores align the text summaries from the predicted timeline with the ground truth in a many-to-one manner by both date and semantic similarity in terms of ROUGE-1 as outlined in (Martschat and Markert, 2017). ROUGE-1 and ROUGE-2 F1-score are then used to assess the similarity of these aligned pairs of text to produce final scores. While this does not require exact matching of dates, it suffers from the general weaknesses of ROUGE metrics. ROUGE only evaluates the co-occurrence of n-grams between the two texts. It was not designed to evaluate if texts with different surface forms convey the same semantic meaning (Zhao et al., 2019). This is crucial, as the primary objective of a predicted timeline is to provide the same semantic information as the ground truth. Furthermore, ROUGE has been shown to be a poor evaluator of text quality, with a generally low correlation with human evaluations of summary quality (Liu and Liu, 2008; Reiter, 2018).

4.2.2 Date Distribution Metrics

To better evaluate date selection, and by extension event detection performance, we propose focusing on the difference between the distributions of dates for the ground truth and predicted timelines. To accomplish this, we primarily use the Kolmogorov–Smirnov statistic, which quantifies a distance between the two distributions equivalent to the supremum distance between the two timelines' date ECDFs. We can also use the KS statistic in conjunction with the number of dates that form the distributions to perform a two-sample Kolmogorov–Smirnov test. When this test results in a p-value less than our significance level of 0.05, we reject the null hypothesis that the dates for both timelines were drawn from the same distribution. The KS test results are reported as the fraction of predicted timelines where the predicted date distribution was statistically significantly different than that of the ground truth in this way.

We also use the earth mover's distance (the Wasserstein metric) as a measure of distance between the two distributions. This distance measure is more sensitive to variations in start and end time between predicted and ground truth, which can be

beneficial as the KS statistic is least sensitive towards the extremes. When calculating the earth mover's distance (EMD) all dates are scaled such that the start date of the ground truth timeline is 0 and the last date is 1. Even with this scaling, there can be cases where poorly predicted timelines in a dataset result in EMDs an order of magnitude larger than typical, which throws off the use of this distance score in the aggregate. This can be seen in Table 8. As a result, we recommend the Kolmogorov–Smirnov statistic and test results as the primary date selection quality metrics.

4.2.3 Text Quality Metrics

When evaluating the quality of timeline text, the primary objective is that the predicted text conveys the same semantic meaning about the same underlying events as the text of ground truth timeline, regardless of surface form. To that end, we choose to use MoverScore (Zhao et al., 2019), a generalization of Word Mover's Distance (Kusner et al., 2015), as our text evaluation metric. To determine the similarity of two texts, MoverScore first soft aligns similar words between the pair of texts by their contextual BERT embeddings. Based on these embeddings and alignments, MoverScore finds the minimum effort required to transform the texts into each other by treating it as a constrained optimization problem. The intuition behind using a distance metric in this way, is that is will be better at indicating how much a generated text deviates from the ground truth, as opposed to just capturing their shared content.

We use MoverScore to evaluate both the global and aligned quality of timeline texts. To compute the global similarity, reported as MoverScore, we join the text from each timeline in order and then computer the MoverScore between the two chronological summaries of the topic.

To measure event aligned text quality, we first compute the MoverScore between all pairs of ground truth and predicted event texts. We then greedily align events in the two timelines by their semantic similarity, as determined by the MoverScore, in a one-to-one manner. The average and median MoverScores of the aligned pairs are then used to compute aligned scores (reported as A-MoverScore Avg/Med). In the formulation of the news TLS task we are using, a model may not predict more timelines events than are on the ground truth timeline. If a model predicts less than the ground truth number of events (e.g. due to not being able to detect enough events), unaligned ground truth events contribute scores of 0 towards the average and median.

4.3 Improving TLS

4.3.1 Date-wise improvements

The date-wise news TLS method has two distinct sub-tasks: date selection followed by summarization of those dates. We investigate whether it is possible to improve on the SOTA date-wise method by improving the date-selection sub-task. To do this, we constructed 5 variations of date-predictive models: a logistic regression model, a fully-connected network (FCN), a deep fully-connected network (Deep FCN), a wide fully-connected network (Wide FCN), and a convolutional neural network (CNN). The models dimensions are as follows:

- FCN: 2×100 linear layers
- Deep FCN: 8×100 linear layers
- Wide FCN: $2\times4{,}00$ linear layers
- CNN: 2×1d convolutions, 1×1d pooling, $2\times$dense layers

We compare each model to the current date-wise SOTA baseline which leverages a simple linear regression model for date prediction. We train our date-predictive models on each dataset, resulting in 3 trained models per architecture. This differs from the current SOTA date-wise method which trains a linear regression model for each topic within a dataset, resulting in a total of 60 trained models. Our approach is designed to make more data available to train our more advanced date-predictive models. However, one downside to our approach, is that it may not perform well if there is a large variance between topics within a dataset. For all of our neural net architectures, we use Xavier weight initialization, ReLU activations, 0.5 dropout, batch normalization, L2 normalization on the inputs, and a sigmoid output. All models are trained on random train/val/test (0.8/0.1/0.1) splits.

For complete hyperparameters and implementation details, please reference the link to the project's repository in Section 5.3.

4.4 Clustering based improvements

The clustering approach for timeline generation is based in the intuition that an important real world

event is likely to result in a cluster of semantically and temporally similar news articles. Once clusters of articles are found, we can use the information each cluster contains to rank them by importance, determine their associated dates, and then generate summaries for the top l most important underlying events. Here, we focused on improving the clustering method and retained the date assignment, ranking, and summarization methods from (Gholipour Ghalandari and Ifrim, 2020) for ease of comparison. The centroid-based extractive summarization method used, Centroid-Opt (Gholipour Ghalandari, 2017), is also used by the date-wise models.

For article clustering, we adopt the temporally constrained Markov Clustering (TCMC) framework from (Gholipour Ghalandari and Ifrim, 2020). To perform TCMC, a graph is created in which nodes are news articles. Edges are added between all articles within a time window of one another, the weights of which represent the similarity of the articles. When this graph is complete, the random walk based Markov Cluster Algorithm (MCL) (Dongen, 2000) is used to detect clusters. We focused on the effects of varying the time constrain window and using a different similarity measure.

In (Gholipour Ghalandari and Ifrim, 2020), a temporal constraint of one day is always used. When generating timelines for topics that span multiple years, as is the case with their entities dataset and our topics dataset, this seemed overly restrictive, leading us to explore using larger time windows.

In the original TCMC scheme, the similarity weight assigned to edges is the cosine similarity of the TF-IDF vectors for the pair of news articles. We also evaluated using the cosine similarity of sentence-BERT (SBERT) embeddings of news articles as the similarity edge weight (Reimers and Gurevych, 2019). We used a DistilRoBERTa model pretrained on millions of paraphrase examples by Ubiquitous Knowledge Processing (UKP) Lab. Up to the first 512 word pieces tokens, or about 300-400 words, per article were used to compute the embedding.

5 Results

In the following section, we provide the results from our attempts to improve news TLS performance. All references to "old metrics" are referring to the metrics all previous news TLS works use to establish SOTA performance, namely Aligned-ROUGE-1 and Aligned-ROUGE-2 for summary quality, and Date-F1 scores for date selection. All references to "new metrics" are referring to the metrics we are proposing as better evaluation metrics for the news TLS task, namely Moverscore for measuring summary quality, and the Kolmogorov–Smirnov statistic and Earth Mover's Distance for evaluating date selection.

5.1 Date-wise improvements

Tables 3, 4, and 5 contain the results comparing the current SOTA date-wise performance (labeled "base" in each table) to the performance of date-wise methods using our new date-predictive models.

Crisis Results: From the experiments with the *Crisis* dataset (Table 3), we observe that the logistic regression model outperforms all other models in the old evaluation metrics. However, using the new metrics, the neural networks offer superior performance in both date selection and summarization. The Wide FCN achieves the highest Moverscore of 0.1702.

T17 Results: From the experiments with the *T17* dataset (Table 4), we observe both the Wide FCN and the logistic regression models perform best with the old metrics. The CNN and Deep FCN both perform well in predicting dates and in summarization.

Entities Results: From the experiments with the *Entities* dataset (Table 5), we observe that no new models were able to out-perform the SOTA baseline in the old metrics. However, with the new metrics, the Wide FCN predicts more representative dates and the logistic regression model produces better summaries.

The results overall show that better date selection methods does indeed lead to improved date-wise summarization. Considering the old metrics, our new models marginally outperform SOTA date-wise in most cases. However, considering our new metrics, the new models outperform SOTA date-wise in all cases. We believe the datasets are still not large enough for deep-learning models to significantly outperform simpler methods such as linear and logistic regression.

5.2 Clustering improvements

Clustering based model performance is shown is Tables 6, 7, and 8 alongside the current SOTA base

Date-wise experiments: Crisis

	Base	LogR	FCN	Deep FCN	Wide FCN	CNN
A-R1: F1	0.0891	**0.0897**	0.0279	0.0589	0.0848	0.0853
A-R2: F1	0.0261	**0.0263**	0.0072	0.0137	0.0248	0.0252
DateF1	0.2945	**0.3038**	0.0733	0.1923	0.2656	0.2694
KS Stat	0.4173	0.4028	0.4098	**0.3259**	0.4017	0.4041
KS Signif	0.5455	0.5455	0.5455	**0.3182**	0.4545	0.5000
Earth Movers Dist	0.2381	0.2316	0.1763	**0.1561**	0.2337	0.2295
Moverscore	0.1642	0.1681	0.1434	0.1575	**0.1702**	0.1678
A-Moverscore Avg	0.1299	0.1319	0.1225	0.1249	**0.1376**	0.1291
A-Moverscore Med	0.1316	0.1352	0.1225	0.1346	**0.1375**	0.1345

Table 3: Results comparing various date-selection models using the date-wise method on the *Crisis* dataset.

Date-wise experiments: T17

	Base	LogR	FCN	Deep FCN	Wide FCN	CNN
A-R1: F1	0.1201	**0.1210**	0.1151	0.1018	0.1207	0.1187
A-R2: F1	0.0351	0.0348	0.0317	0.0272	**0.0352**	0.0345
DateF1	0.5436	0.5530	0.5213	0.4418	**0.5449**	0.5195
KS Stat	0.2445	0.2551	0.2110	0.2299	0.2389	**0.1978**
KS Signif	0.3158	0.3684	0.1579	0.1579	0.2632	**0.1053**
Earth Movers Dist	0.1211	0.1276	0.1010	0.1138	0.1187	**0.0937**
Moverscore	0.1417	**0.1476**	0.1429	0.1413	0.1445	0.1440
A-Moverscore Avg	0.1084	0.1086	0.1090	0.1113	0.1105	**0.1116**
A-Moverscore Med	0.1117	0.1159	0.1135	**0.1178**	**0.1178**	0.1163

Table 4: Results comparing various date-selection models using the date-wise method on the *T17* dataset.

date-wise model and another date-wise model that performs well on each task.

Crisis and T17 Results: These datasets share similar characteristics and had similar results with clustering models. When compared to the *Entities* and *Topics* dataset timelines, the *Crisis* and *T17* timelines are on average an order of magnitude smaller while containing more events. In this setting, increasing the time window beyond one day was not shown to increase performance. Using SBERT embedding similarity to weigh edges did lead to the best global MoverScore of the unsupervised models, but at the expense of performance in our other metrics. This imbalance in performance across the three areas our metrics evaluate (overall similarity, aligned similarity, and date distribution) is undesirable, and thus we do not consider this an improvement.

Entities Results: Here, clustering based models beat the current SOTA date-wise model by all metrics except the old aligned ROUGE and date-F1 scores. These models were able to outperform all of our date-wise models in all new metrics, except for the wide FCN date-wise model, which had the best overall aligned MoverScores.

Increasing the time window in which articles are connected by a similarity edge up from one day led to better performance, however, using too large of a window could lead to a loss of granularity in event detection and in turn worse timeline performance. The optimal window length is likely a function of factors including the length and density of the timeline being predicted and the distribution of articles in the associated corpus. Future works could look to address this by estimating the optimal window length on a per case basis or only assigning edges for articles above a similarity threshold.

When compared to TF-IDF vector similarity, using SBERT embedding similarity to weigh edges improved overall MoverScore without sacrificing performance in our other new metrics. The choice of TF-IDF or SBERT similarity for edge weighting also impacted the effect of different window sizes. Clustering using SBERT leads to a loss of granularity with smaller window sizes than with TF-IDF. We found optimal performance using SBERT with

Date-wise experiments: Entities

	Base	LogR	FCN	Deep FCN	Wide FCN	CNN
A-R1: F1	**0.0566**	0.0523	0.0109	0.0422	0.0112	0.0297
A-R2: F1	**0.0171**	0.0158	0.0026	0.0124	0.0027	0.0073
DateF1	**0.2049**	0.1787	0.0230	0.1264	0.0275	0.0980
KS Stat	0.4161	0.3992	0.4569	0.3894	**0.3733**	0.3954
KS Signif	0.4894	0.4894	0.6383	0.3830	**0.3191**	0.4468
Earth Movers Dist	0.1921	0.1798	0.2000	0.1741	**0.1524**	0.1731
Moverscore	0.1014	**0.1021**	0.0879	0.0972	0.0882	0.0942
A-Moverscore Avg	0.0664	**0.1086**	0.0444	0.0611	0.0477	0.0564
A-Moverscore Med	0.0538	**0.1159**	0.0437	0.0549	0.0487	0.0525

Table 5: Results comparing various date-selection models using the date-wise method on the *Entities* dataset.

Model	**Similarity Type**	**Time Constraint**	**Mover-Score**	**A-MS Avg**	**A-MS Med**	**KS Stat**	**KS Test**	**Date EMD**	**AR1 F1**	**AR2 F1**	**Date F1**
Base	n/a	n/a	0.101	0.066	0.054	0.416	**0.489**	0.192	**0.057**	**0.017**	**0.205**
LogR			**0.102**	**0.109**	**0.116**	**0.399**	**0.489**	**0.180**	0.052	0.016	0.179
TCMC	TF-IDF	1	0.105	0.067	0.064	0.357	0.277	0.164	**0.051**	**0.015**	**0.174**
		7	0.108	0.070	**0.072**	0.346	**0.234**	0.155	0.047	0.013	0.161
		11	0.109	0.070	0.064	0.344	0.255	0.155	0.045	0.012	0.157
		21	0.108	0.070	0.064	0.345	0.298	0.148	0.042	0.012	0.155
		28	0.107	**0.071**	0.067	0.329	**0.234**	**0.141**	0.042	0.010	0.150
TCMC	SBERT	1	0.104	0.070	0.063	0.381	0.489	0.164	0.049	**0.015**	0.165
		6	0.109	0.070	0.064	0.357	0.319	0.155	0.042	0.012	0.157
		7	**0.111**	**0.071**	0.064	0.353	0.319	0.150	0.043	0.012	0.156
		11	**0.111**	0.070	0.063	0.341	0.277	0.147	0.040	0.010	0.148
		14	0.108	0.069	0.060	**0.323**	**0.234**	0.145	0.037	0.009	0.142

Table 6: Results comparing temporally constrained Markov Clustering based models and high performing date-wise models on the *Entities* dataset. Time constraint is in number of days. The best date-wise and clustering scores are in bold.

a window size around a week and a significant drop off in performance over around two weeks. With TF-IDF there was little difference using windows ranging from one to four weeks.

5.3 Code and datasets

Our newly curated datasets and all the code for this project is available on GitHub: https://github.com/RossDeVito/news-tsl-cse291

6 Discussion and Conclusion

News TLS is a challenging problem. Ground-truth important dates and representative summaries are subjective targets that are difficult to quantify. Ground-truth timelines and their associated news corpora may not contain all the same events. Event summaries vary in length from a few words to a paragraph. Furthermore, there's a lot of variance in timeline length (200 days to 12 years). Current timeline summaries are still not accurate enough to be useful. Crucial tasks including cluster date selection and cluster ranking are still done heuristically which is something future work may improve on.

In this project, we expanded the available data for future TLS works with 35 new of topics with newly crawled timelines, keywords, and articles. We also propose using the MoverScore (for text summaries) and the Kolmogorov-Smirnov statistic (for date selection) as new metrics to evaluate future TLS work. Lastly, we were successful in beating the current state-of-the-art for News Timeline Summarization. Our improved date-wise method out-performed current SOTA in the existing evaluation metrics (ROUGE-1 and ROUGE-2) and our improved cluster method exceeded current SOTA in our new evaluation metrics (MoverScore and the KS statistic). Our work revealed that date-wise approaches will benefit from choosing the date-predictive architecture that performs best per

Model	Similarity Type	Time Constraint	Mover-Score	A-MS Avg	A-MS Med	KS Stat	KS Test	Date EMD	AR1 F1	AR2 F1	Date F1
Base	n/a	n/a	0.164	0.130	0.132	0.417	0.546	0.238	**0.089**	**0.026**	**0.295**
Wide FCN			**0.170**	**0.138**	**0.138**	**0.402**	**0.455**	**0.234**	0.085	0.025	0.266
TCMC	TF-IDF	1	0.150	**0.128**	**0.132**	**0.313**	0.318	0.172	**0.061**	**0.013**	0.226
		2	0.144	0.116	0.128	0.334	0.364	0.190	0.057	0.012	0.237
		3	0.149	0.118	0.125	0.333	0.318	0.191	0.052	0.012	**0.238**
		7	0.135	0.098	0.104	0.345	0.364	0.176	0.040	0.008	0.212
		28	0.122	0.066	0.066	0.332	**0.227**	0.199	0.028	0.005	0.169
TCMC	SBERT	1	**0.161**	0.087	0.076	0.331	0.318	**0.141**	0.052	0.009	0.193
		3	0.118	0.038	0.019	0.438	0.273	0.208	0.038	0.004	0.152

Table 7: Results comparing temporally constrained Markov Clustering based models and high performing date-wise models on the *Crisis* dataset. Time constraint is in number of days. The best date-wise and clustering scores are in bold.

Model	Similarity Type	Time Constraint	Mover-Score	A-MS Avg	A-MS Med	KS Stat	KS Test	Date EMD	AR1 F1	AR2 F1	Date F1
Base	n/a	n/a	0.142	0.108	0.112	0.245	0.316	0.121	0.120	**0.035**	0.544
Wide FCN			**0.145**	**0.111**	**0.118**	**0.239**	**0.263**	**0.119**	**0.121**	**0.035**	**0.545**
TCMC	TF-IDF	1	**0.124**	**0.088**	0.078	**0.232**	**0.158**	8.300	**0.082**	**0.020**	**0.407**
		2	0.111	0.084	**0.080**	0.269	**0.158**	2.005	0.076	0.018	0.387
		3	0.116	0.084	**0.080**	0.288	**0.158**	0.709	0.077	0.016	0.356
		7	0.122	0.076	0.070	0.316	0.316	0.361	0.072	0.014	0.358
		28	0.122	0.066	0.051	0.375	0.316	**0.342**	0.063	0.014	0.307
TCMC	SBERT	1	0.122	0.060	0.042	0.428	0.263	0.368	0.075	0.019	0.229

Table 8: Results comparing temporally constrained Markov Clustering based models and high performing date-wise models on the *T17* dataset. Time constraint is in number of days. The best date-wise and clustering scores are in bold.

dataset (e.g. Wide FCN for *Crisis*, CNN for *T17*, etc.).

References

2013. *WWW '13 Companion: Proceedings of the 22nd International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA.

S.M. van Dongen. 2000. *Graph clustering by flow simulation*. Ph.D. thesis, Utrecht University.

Demian Gholipour Ghalandari. 2017. Revisiting the centroid-based method: A strong baseline for multi-document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 85–90, Copenhagen, Denmark. Association for Computational Linguistics.

Demian Gholipour Ghalandari and Georgiana Ifrim. 2020. Examining the state-of-the-art in news timeline summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1322–1334, Online. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2017. Improving ROUGE for timeline summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 285–290, Valencia, Spain. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. 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), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015a. Timeline summarization from relevant headlines. pages 245–256.

Giang Tran, Eelco Herder, and Katja Markert. 2015b. Joint graphical models for date selection in timeline summarization. 1:1598–1607.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.