Ranking Multidocument Event Descriptions for Building Thematic Timelines

Hieu Kiem Nguyen

16 PUBLICATIONS **108** CITATIONS

SEE PROFILE

Xavier Tannier
Sorbonne Université

133 PUBLICATIONS **1,301** CITATIONS

SEE PROFILE

Ranking Multidocument Event Descriptions for Building Thematic Timelines

Kiem-Hieu Nguyen[†], **Xavier Tannier**[†$], **and Veronique Moriceau**[†$]
[†]LIMSI-CNRS
[$]Univ. Paris-Sud
Orsay, France
{nguyen,xtannier,moriceau}@limsi.fr

Abstract

This paper tackles the problem of timeline generation from traditional news sources. Our system builds thematic timelines for a general-domain topic defined by a user query. The system selects and ranks events relevant to the input query. Each event is represented by a one-sentence description in the output timeline.

We present an inter-cluster ranking algorithm that takes events from multiple clusters as input and that selects the most salient and relevant events. A cluster, in our work, contains all the events happening in a specific date. Our algorithm utilizes the temporal information derived from a large collection of extensively temporal analyzed texts. Such temporal information is combined with textual contents into an event scoring model in order to rank events based on their salience and query-relevance.

1 Introduction

We aim at building thematic timelines from multiple documents relevant to a specific, user-generated query. For instance, for the query "*Libya conflict*", our system will return important events related to the Libya conflict in 2011 involving Kadhafi forces, rebels, NATO intervention, etc. (Figure 1). Such a timeline can then be visualized as a textual, event-based summary, or through any existing graphical timeline visualization tool.

The main contribution of this paper is a two-step inter-cluster ranking algorithm aimed at selecting salient and non-redundant events from temporal clusters, which are sets of sentences describing events related to the query and that occurred at the same day. In the first step, a scoring model is proposed to rank sentences describing events, according to their relevance and salience to the topic. In the second step, the ranked events are iteratively reranked based on their content in order to reduce information redundancy. We finally obtain an extendable, chronological summary of important events concerning the query.

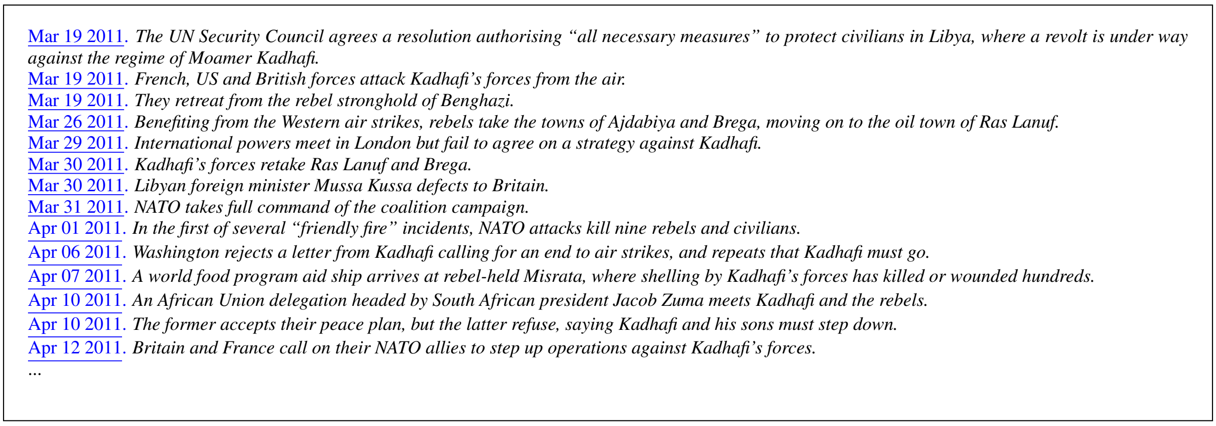
This paper is organized as follows: §2 introduces related work. §3 presents the resources used and gives an overview of the system. The salient date algorithm proposed by Kessler et al. (2012), that we used to build our temporal clusters, is briefly summarized in §4. §5 and §6 describe our ranking approach to event selection and a content-based reranking algorithm, respectively. The evaluations are presented in §7. §8 is dedicated to the conclusion and future work.

2 Related Work

Our work is closely related to event detection and tracking (EDT) and multidocument summarization (MDS). This section introduces some important work in these fields.

2.1 Event Detection and Tracking

EDT on news streams has been intensively studied. Early work concentrates on detecting events from article texts using vector-based techniques (Allan et al., 1998; Petrović et al., 2010) or graphical models



Mar 19 2011. *The UN Security Council agrees a resolution authorising "all necessary measures" to protect civilians in Libya, where a revolt is under way against the regime of Moamer Kadhafi.*
Mar 19 2011. *French, US and British forces attack Kadhafi's forces from the air.*
Mar 19 2011. *They retreat from the rebel stronghold of Benghazi.*
Mar 26 2011. *Benefiting from the Western air strikes, rebels take the towns of Ajdabiya and Brega, moving on to the oil town of Ras Lanuf.*
Mar 29 2011. *International powers meet in London but fail to agree on a strategy against Kadhafi.*
Mar 30 2011. *Kadhafi's forces retake Ras Lanuf and Brega.*
Mar 30 2011. *Libyan foreign minister Mussa Kussa defects to Britain.*
Mar 31 2011. *NATO takes full command of the coalition campaign.*
Apr 01 2011. *In the first of several "friendly fire" incidents, NATO attacks kill nine rebels and civilians.*
Apr 06 2011. *Washington rejects a letter from Kadhafi calling for an end to air strikes, and repeats that Kadhafi must go.*
Apr 07 2011. *A world food program aid ship arrives at rebel-held Misrata, where shelling by Kadhafi's forces has killed or wounded hundreds.*
Apr 10 2011. *An African Union delegation headed by South African president Jacob Zuma meets Kadhafi and the rebels.*
Apr 10 2011. *The former accepts their peace plan, but the latter refuse, saying Kadhafi and his sons must step down.*
Apr 12 2011. *Britain and France call on their NATO allies to step up operations against Kadhafi's forces.*
...

Figure 1: A chronology about "*Libya conflict*" written by journalists.

(Sayyadi et al., 2009). These papers do not consider time, which is an essential dimension of event timelines.

Attempts to use temporal information for EDT are significant in the literature. To name but a few, Alonso et al. (2009) apply time-based clustering on search results. Yan et al. (2011) use document timestamps to calculate temporal proximity for timeline generation from web documents. Similarly, Zhao et al. (2007) use text similarity and time intensity for event clustering on social streams. Kessler et al. (2012) exploit temporal analysis to detect salient dates of an event from raw text. Following this direction, Battistelli et al. (2013) apply sequential pattern mining to select a one-sentence description for each salient date of an event.

2.2 Multidocument Summarization

Sentence extraction is essential in extractive text summarization. In the unsupervised approach, sentences are scored using term weight and term proximity induced from a document collection (Goldstein et al., 2000). In the supervised approach, training data generated from reference summaries are used to learn classification or ranking models. New sentences are selected based on their confidence value on learned models (Wan et al., 2007). As information comes from documents on the same topic, it should be noticed that it is also important to reduce redundancy in MDS (Carbonell and Goldstein, 1998).

Filippova (2010) builds a co-occurrence word graph from a collection of related sentences and generates a generic summary from the graph based on shortest path finding. Her algorithm is a hybrid method between extractive and abstractive approaches to MDS.

3 Resources and System Overview

3.1 Corpus and Chronologies

For this work, we use a corpus of newswire texts provided by the AFP French news agency. The English AFP corpus is composed of 1.3 million texts that span the 2004-2011 period (511 documents/day in average and 426 millions words). Each document is an XML file containing title, document creation time (DCT), set of keywords, and textual content split into paragraphs.

AFP "chronologies" (textual event timelines) are a specific type of articles written by AFP journalists in order to contextualize current events. These chronologies may concern any topic discussed in the media, and consist in a list of dates (typically between 10 and 20) associated with a text describing the related event(s). Figure 1 shows an example of such a chronology. Note that several important events can occur at the same date.

3.2 System Overview

Figure 2 shows the general architecture of the system. When the user submits a query, sentences are retrieved by the Lucene search engine and are clustered by the dates appearing in those sentences (step ①

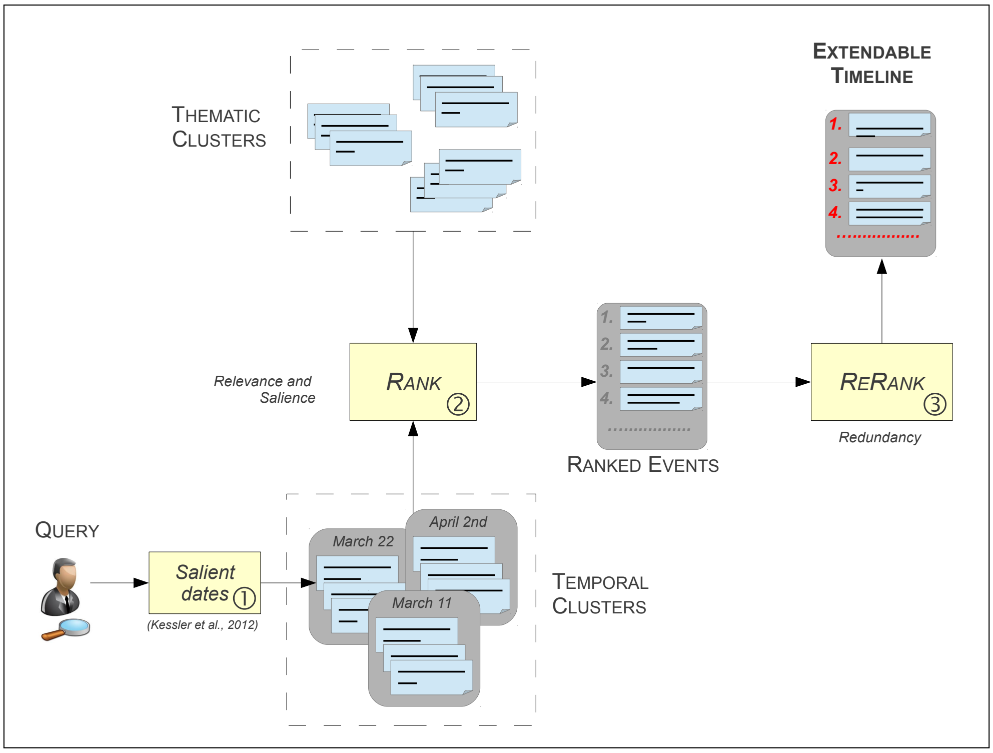


Figure 2: System overview.

in the Figure (Kessler et al., 2012)).

Then, all sentences are ranked by the relevance and salience of described events. This is done by modeling event relevance and salience as a scoring function (step ②). *Thematic* clusters are created by applying clustering on the set of events on the same date. Finally, sentences are reranked by an iterative algorithm aiming at reducing redundancy from the initial list (step ③) to achieve an extendable timeline.

4 Temporal Clusters

As stated in the introduction, our main contribution in this paper is to rank and select salient and non-redundant sentences from clusters, in order to build query-based timelines. We rely on the algorithm proposed by Kessler et al. (2012) for building temporal clusters. This section is a quick overview of their approach.

4.1 Preprocessing

A temporal analysis is performed on all documents from the AFP corpus (see §3.1) with the Heideltime (Strötgen and Gertz, 2013) parser. The main purpose is to collect as much temporal information as possible. Absolute dates and DCT-relative dates are extracted and normalized (full dates represented in a common format). DCT-relative dates are those which are relative to the date on which the document is published, such as "*Yesterday*" (day before DCT), "*next Friday*" (first Friday following the DCT) or "*on Friday*" (can be first Friday preceding or following the DCT, depending on the tense of the verb that governs the temporal expression).

In a corpus containing 426 millions words, 845,000 absolute dates and 4.6 millions relative dates were detected and normalized.

4.2 Temporal Cluster Building

At query time, temporal clusters (or "salient date sets") are then built with the help of a search engine (Lucene in that case[1]). Articles are indexed by Lucene at sentence-level (a document = a sentence).

[1] http://lucene.apache.org

...launched the first air strike on	March 19	*, has deployed around 20 Rafale and Mirage...*
...last week, said they were arrested on	Saturday	*along with Getty photographer Joe Raedle...*
...since coalition air strikes began	Saturday	*, a figure that could not be confirmed...*
...United State to launch air raids on	March 19	*, are in a hurry to get out of whether NATO...*

Figure 3: All the temporal expressions in the following sentences were normalized at date level as 'Mar 19 2011'.

Given the query, a number of sentences are retrieved by search engine. Dates are extracted from these sentences. These dates are then ranked by their "salience" in the set of documents. The idea behind the notion of salient date is that if a date is important in a sub-corpus (Lucene output), then we can say that important events occurred at this date, and then that these events must appear in a timeline.

In practice, salience is mostly defined by the number of occurrences of the date in the documents from the search engine, as well as some other features that are used to feed a machine learning classifier.

The output of this salient date algorithm is then a ranked list of dates, where each date comes together with a set of sentences that contain this date and that are relevant to the query. We call *temporal clusters* these sets of sentences linked to a specific date (see Figures 2 and 3).

5 Event Ranking

Our ranking mechanism relies on the mutual relation between relevance and salience. It aims at ranking events based on these two factors. The problem of information redundancy will be addressed by a reranking step in §6. Our principal motivation is that an event has more chance to be selected into a timeline if it is both relevant to the topic and important, or in other words, salient w.r.t other related events. The concepts of relevance and salience are realized in our ranking function by considering term proximity and date frequency, respectively.

Previous works in event detection normally formalize events as individual terms or syntactic patterns, which facilitates the use of text content. Instead, as our method utilizes both time and text content, we come to a formalization of an event as a pair of its mentioned date and its one-sentence description.

Given an input query, the aim of ranking is to select the most relevant and salient events. The relevance of an event is calculated by vector-based query similarity, and augmented by the average relevance of its containing thematic cluster. Salience is contributed by date frequency and averaged term weight. As a result, the overall score of an event e given a query q is the multiplication of the following four factors:

$$\begin{aligned} score(e|q) =& rel_e(e|q) \\ & * rel_{cl}(cl|q) \\ & * salience_e(e|d,q) \\ & * salience_d(d|q), \end{aligned} \tag{1}$$

where:

- $rel_e(e|q)$ is the relevance of e to q (see §5.1).
- $rel_{cl}(cl|q)$ is the relevance of a thematic cluster cl to q, which is the averaged relevance of its members (see §5.2).
- $salience_e(e|d,q)$ is the salience of e w.r.t the date d that the event happens. It is calculated as the average salience of the terms in its one-sentence description. Term salience, in turn, is calculated based on term frequency in the date cluster (see §5.3).
- $salience_d(d|q)$ is the salience of d w.r.t to q. Date salience is the averaged salience of all the events in that date (see §5.4).

5.1 Event Relevance: $rel_e(e|q)$

The motivation behind considering relevance is that if an event is relevant to the query then it is an important event. We use the conventional TFIDF vector space model with bag-of-word assumption to represent document and query vectors. For relevance, the similarity between document and query vectors is the built-in Lucene score formula[2],

$$rel_e(e|q) = cosine(\vec{e}, \vec{q}) * norm_L(\vec{e}, \vec{q}). \tag{2}$$

5.2 Thematic Cluster Relevance: $rel_{cl}(cl|q)$

Date salience does not always correctly reflect the importance of event. For instance, the date of Haiti earthquake considers the earthquake itself as the main event. However, related events such as the sorrow expression of UN Secretary General also happen immediately after the earthquake but still in the same date. Such satellite events will have the same date salience as the central event. In another case, a date when there is no central event but there are many "consequent" events will also have a high salience value. E.g., on the day after the earthquake, international aids are planned; number of victims is estimated; aftermath events are invoked, etc.

Those examples show that the "one event per date" assumption is weak in reality. To overcome this weakness, we apply an hierarchical clustering technique, in which two clusters are merged if their normalized Manhattan distance is lower than a threshold θ, to generate thematic sub-clusters inside a date cluster [3]. In in-house experiments, we observed that different values of θ did not significantly vary performance. We hence selected $\theta = 0.5$ for our system.The score of each thematic cluster is then calculated as averaged document relevance,

$$rel_{cl}(cl|q) = \frac{\sum_e rel_e(e|q)}{|cl|}. \tag{3}$$

5.3 Event Salience: $salience_e(e|d, q)$

An important event tends to contain salient terms. Those terms, in turn, tend to occur frequently on a date. We hence come to measure term salience as its frequency of occurrence on the date $f(t|d, q)$, and event salience as the averaged salience of its terms. For term normalization, stopwords are removed and tokens are normalized by the Porter stemming algorithm (Porter, 1997).

$$salience_e(e|d, q) = \frac{\sum_{t \in e} f(t|d, q)}{|e| \sum_{t' \in d} f(t'|d, q)} \tag{4}$$

5.4 Date Salience: $salience_d(d|q)$

The use of temporal clusters, i.e. date clusters, is motivated by the observation that an important event happens on a salient date. Date salience is the total relevance of all events happening on that date (the numerator):

$$salience_d(d|q) = \frac{\sum_e rel_e(e|q)}{\sum_d \sum_e rel_e(e|q)}. \tag{5}$$

The denominator is used to normalize date salience so that it is comparable to other factors in (1).

6 Event Reranking

The score described in previous section leads to a ranked list of salient and relevant events. However, it does not consider the fact that some information can be redundant between events. The reranking algorithm presented in this section strives to reduce such redundancy. In principal, information redundancy is

[2] https://lucene.apache.org/core/3_6_2/api/core/org/apache/lucene/search/Similarity.html

[3] In our implementation, for each one-sentence document, we used the whole texts of its containing article to create its document vector. Manhattan distance is the sum of the absolute difference of term weight between two clusters

Rank	Date	Event Description
		NO RERANK
1	Mar 31 2011	*The North Atlantic Treaty Organisation takes over formal command of the military operation.*
2	Mar 31 2011	*NATO took command of operations over Libya on March 31.*
3	Mar 31 2011	*NATO takes command of the coalition campaign.*
4	Mar 19 2011	*[...] French, US and British forces launch UN-mandated air strikes and push them back.*
5	Mar 30 2011	*Libyan foreign minister Mussa Kussa defects.*
...	...	...
		RERANK
1	**Mar 31 2011**	*The North Atlantic Treaty Organisation takes over formal command of the military operation.*
2	Mar 19 2011	*[...] French, US and British forces launch UN-mandated air strikes and push them back.*
3	Mar 30 2011	*Libyan foreign minister Mussa Kussa defects.*
4	Mar 23 2011	*US Defence Secretary Robert Gates on Wednesday held talks in Cairo on the conflict in Libya [...]*
5	Apr 04 2011	*[...] photographer Manu Brabo disappeared on April 4 while covering the conflict in Libya.*
...	...	...
{2}	Mar 31 2011	*NATO took command of operations over Libya on March 31.*
...	...	...
{3}	Mar 31 2011	*NATO takes command of the coalition campaign.*
...	...	...

Figure 4: The effect of reranking on the order of events (by score).

Algorithm 1 Reranking algorithm

```
1: $out \leftarrow \phi$
2: while $(!terminate)$ do
3:     for $e \in S(q) \setminus out$ do
4:         $score(e|q)$
5:     end for
6:     $e^* = \operatorname{argmax}_{e \in S(q) \setminus out} score(e|q)$
7:     $out \leftarrow out \cup e^*$
8:     $d^* = date(e^*)$
9:     for $t \in e^*$ do
10:        $used(d^*) \leftarrow used(d^*) \cup t$
11:    end for
12: end while
```

estimated by the distinction between *used* and *unused* terms. The algorithm iteratively recomputes event salience (hence the overall event score) based on used/unused terms as follows:

$$salience^*_e(e|d,q) = \frac{\sum_{t^* \in e} f(t^*|d,q)}{|e| \sum_{t' \in d} f(t'|d,q)}, \tag{6}$$

where t^* is an unused term on the date d. A used term is the one that already occurred in better-ranked sentences. This formula is different from (4) in the distinction between used and unused terms. Each time a new event is selected, its appropriate list of used terms is updated with the terms in the one-sentence description of the selected event. Each date has its own list of used/unused terms.

The algorithm for reranking is provided in Algorithm 1. At first, the score of all sentences related to the query $S(q)$ is calculated using the formula (1) with event salience defined in (6) (lines 3-5). Then, the highest scored sentence is selected into the output (lines 6-7) and is removed from the pool. In line 8, d^* is the date when the event e happens: $d^* = date(e^*)$. The list of used terms on its date is updated with the terms from that selected sentence (lines 9-11). A new iteration restarts by recalculating score of unselected sentences according to new lists of used terms. The algorithm terminates after K iterations, i.e. when K events have been selected into timeline.

Figure 4 illustrates the effect of reranking on the order of events in a timeline. The upper shows the top events ranked by score without reranking. The date 'Mar 31 2011' appears three times in 1st, 2nd, and 3rd events. The lower shows the ranking of events after the highest scored event has been selected. As an effect of reranking, the two events previously ranked 2nd and 3rd now fall down the list.

Mar 19 2011. **(2)** *With the forces of Libyan leader Moamer Kadhafi threatening rebel-held Benghazi, French, US and British forces launch UN-mandated air attacks and push them back.*
Mar 19 2011. **(9)** *Norwegian Prime Minister Jens Stoltenberg said Saturday Norway would contribute six F-16 warplanes to the international military operation –led by the United States, France and Britain– to enforce a no-fly zone over Libya.*
Mar 19 2011. **(11)** *Residents of another western town, Yafran, say nine people died there in an offensive that began on Monday.*
Mar 21 2011. **(4)** *Kadhafi's forces retreat from the rebel stronghold of Benghazi.*
Mar 22 2011. **(1)** *In Western Libya fighting intensifies in Misrata, which has been in the hands of rebels for a month.*
Mar 24 2011. **(10)** *When I ask: What is the next stage? Do you have a road map? I see they do not, he said Thursday.*
Mar 25 2011. **(12)** *Ping returned early Friday from Europe after meeting with French Foreign Minister Alain Juppe and an envoy sent by the European Union's Chief Diplomat Catherine Ashton.*
Mar 28 2011. **(6)** *Qatar follows France in recognising the rebel shadow government.*
Mar 29 2011. **(3)** *Kadhafi loyalists push the rebels back.*
Mar 30 2011. **(8)** *Kadhafi's forces push back.*
Mar 31 2011. **(5)** *NATO takes command of the coalition campaign.*
Apr 04 2011. **(13)** *Italy joins France and Qatar in recognising the rebel Transitional National Council.*
Apr 13 2011. **(7)** *A Libya contact group of 20 countries and organisations, including the rebels, meets in Qatar.*
Apr 23 2011. **(14)** *The United States carried out its first predator drone strike in Libya on Saturday, the Pentagon said, declining to give details on the targets or location.*
...

Figure 5: Timeline for the query "*Libya conflict*" created by the Rank-Rerank method. Events are shown in chronological order, each accompanied with its rank starting from 1, displayed as a number between ().

7 Evaluations

Our system for building timelines is named as RaRE, as short for "*Rank and RErank*". We use a set of 91 chronologies manually written by expert journalists from the AFP news agency (Figure 1) as golden reference summaries for evaluation. As our generated timelines are extendable, we need to define its length for evaluation. Considering the characteristics of reference summaries, we decide that if a reference summary of a timeline contains k events, we appropriately use only the k highest ranked events in the timeline for evaluation (Figure 5). The evaluations of the date selection and summary generation are presented in §7.1 and §7.2, respectively.

7.1 Evaluate Date Selection

We evaluate the dates selected by timelines returned by our system. The purposes of this evaluation are two-fold: *i)* Since time (as date in our case) is an essential dimension of chronological timeline, it is necessary to evaluate the time selected by timelines; *ii)* The novelty of this work w.r.t Kessler et al. (2012) is the mixture of content and temporal information. We need to show empirical evidences that at least, this mixture does not break the performance of date selection.

The dates occurring in a timeline are compared with the dates occurring in its reference timeline using Mean Average Precision (MAP) metric. It should be noted that by using MAP@k as evaluation metric, a date with higher rank has more impact than another date with lower rank. We use two systems presented in Kessler et al. (2012), named as DFIDF and ML in Table 1, for comparison as follows:

- DFIDF is an unsupervised system solely relying on date frequency with a tfidf-like scoring function. This method uses the AFP corpus, the same as the one used in our work. As the AFP corpus is temporally analyzed, the method indexes all the occurrences of dates in the corpus. Dates are then scored and ranked with so-called DFIDF, a tfidf-like scoring mechanism.

- ML is a supervised system that learns a classifier and ranks unseen dates based on classification

System	MAP
DFIDF	71.46
ML	79.18
RaRE	77.83

Table 1: Comparison of salient date detection using MAP.

System	P	R
DFIDF*	27.24	25.50
ML*	29.93	27.54
RaRE-no-rerank	28.82	24.47
RaRE	31.23	26.63

Table 2: Comparison of MDS using ROUGE at 95% confidence interval.

confidence. The method leverages the dates in reference summaries to create training data with salient/non-salient examples. Temporal features such as date frequency, DCT, novelty, etc., are extracted to learn an adaptive boosting classifier.

As shown in Table 1, our method is close to ML. This result is encouraging as ML requires training data; and on the other hand, our system is not designed to directly solve the task of date selection. As expected, our system beats the unsupervised system DFIDF by a large margin. This superiority shows that the mixture of temporal information and content leads to an improvement on date selection over using only the former.

7.2 Evaluate Summary Generation

In order to evaluate timelines as text summaries, we ignore dates and consider all the entries in a timeline as one summary. We use ROUGE metric (Lin, 2004) to evaluate generated timelines against reference summaries.

The following baselines are implemented (Table 2): In DFIDF*, salient dates are taken from the outputs of the DFIDF system described in previous section. Each salient date is equivalent to a cluster containing all the events happening in that date. We then select the event the most relevant to the query, i.e. the event with the highest Lucene score, as representative of that salient date. Note that consequently, DFIDF* makes an assumption, which is not assumed in RaRE, that there is only one event happens in a particular date. The same assumption is presumed in Battistelli et al. (2013). However, because their system is particularly designed for French and is intended to parse small corpora, we could not conduct a direct comparison with their method. ML* is built similarly to DFIDF*, except that salient dates are instead taken from the ML system. The RaRE-no-rerank system is identical to RaRE in the ranking step, but the reranking step is omitted.

Our system is superior to DFIDF* as expected. Moreover, it outperforms ML*, even though ML* performs better on the task of date selection. Among these three systems that combine temporal information and textual contents for summary generation, our system is the most successful. Furthermore, RaRE outperforms RaRE-no-rerank, which shows that reducing redundancy by reranking improves the performance of summary generation.

8 Conclusion and Future Work

We presented a two-step inter-cluster ranking algorithm for event selection. The *rank* step sorts events based on their salience and query relevance. The event scoring function is based on both date frequency induced from temporal analyzed texts and term weighting induced from contents to reflect these two factors. The *rerank* step allows to reduce information redundancy by using inter-sentence dependency between the descriptions of events happening in the same time period (i.e. the same date in this work).

Ranking based on sentences may be sensitive to sparsity. In the future, we will expand local contexts, for instance, to neighboring sentences, to acquire richer textual representation of events. One remaining issue is that reference chronologies, written by the journalists, are very subjective, and that we have only one example of chronology per topic. In the future, we will conduct a manual evaluation in order to complete results from this automatic evaluation. With the help of a validation interface, journalists will be provided ranked list of events w.r.t. their queries. They will then be able to select and edit the events that they wish to validate for their future timelines. Such an interface will both help journalists to produce new timelines, and bring a new evaluation methodology for our system.

Acknowledgements

This work has been partially funded by French National Research Agency (ANR) under project Chronolines (ANR-10-CORD-010). We would like to thank the French News Agency (AFP) for providing us with the corpus. We would like to thank anonymous reviewers for comments and suggestions on the paper.

References

James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 37–45, New York, NY, USA. ACM.

Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2009. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 97–106, New York, NY, USA. ACM.

Delphine Battistelli, Thierry Charnois, Jean-Luc Minel, and Charles Teissedre. 2013. Detecting salient events in large corpora. In *Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing*, CICLing'13, Berlin, Heidelberg. Springer-Verlag.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Katja Filippova. 2010. Multi-sentence compression: finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 322–330, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 40–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rémy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. 2012. Finding salient dates for building thematic timelines. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–739, Jeju Island, Korea, July. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. F. Porter. 1997. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc.

Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. 2009. Event detection and tracking in social streams. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009.*

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 2903–2908, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 433–443, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, AAAI'07, pages 1501–1506. AAAI Press.