Joint Graphical Models for Date Selection in Timeline Summarization

Giang Tran
L3S Research Center
Leibniz-University Hannover
gtran@l3s.de

Eelco Herder
L3S Research Center
Leibniz-University Hannover
herder@l3s.de

Katja Markert
L3S Research Center
Leibniz-University Hannover
and School of Computing
University of Leeds
markert@l3s.de

Abstract

Automatic timeline summarization (TLS) generates precise, dated overviews over (often prolonged) events, such as wars or economic crises. One subtask of TLS selects the most important dates for an event within a certain time frame. Date selection has up to now been handled via supervised machine learning approaches that estimate the importance of each date separately, using features such as the frequency of date mentions in news corpora. This approach neglects interactions between different dates that occur due to connections between subevents. We therefore suggest a joint graphical model for date selection. Even unsupervised versions of this model perform as well as supervised state-of-the-art approaches. With parameter tuning on training data, it outperforms prior supervised models by a considerable margin.

1 Introduction

Major events (such as the Egypt revolution starting in 2011) often last over a long period of time and have impact for a considerable time afterwards. In order to find out what happened when during such an event, time-related queries to search engines are often insufficient as traditional IR does not handle time-related queries well (Foley and Allan, 2015). To provide readers with comprehensive overviews of long events, many news outlets employ *timeline summaries*: a timeline summary is a list of selected dates with a few sentences describing the most important events on each date. An example can be seen in Table 1. Timelines allow the reader to gain a quick overview over a complex event and to answer questions such as: *How and when did the event start? What were the main consequences of the initial events? What happened to the main protagonists in the event?* In addition, timelines are frequent means in education (such as history teaching) so that their generation is relevant for education as well as journalism.

(a1) 2011-01-25 Egyptians hold nationwide demonstrations against the authoritarian rule of Hosni Mubarak, who has led the country for nearly three decades.
(a2) 2011-01-26 A large security force moves into Cairo's Tahrir Square
(a3) 2011-01-28 Protesters burn down the ruling party's headquarters, and the military is deployed.
(a4) 2011-02-11 Mubarak steps down and turns power over to the military.
(a5) 2011-03-19 In the first post Mubarak vote, Egyptians cast ballots on constitutional amendments ..., including scheduling the first parliamentary and presidential elections
(a8) 2012-04-20 The presidential campaign officially begins.
(a10) 2012-06-24 Election officials declare Morsi the winner
(a26) 2013-07-03 Egypt's military chief says Morsi has been replaced by Adly Mansour, the chief justice of constitutional court.

Table 1: A timeline about the Egypt revolution published by the Associated Press (AP). We leave out intermediate dates due to space constraints. The whole timeline includes 30 dates between 2011-01-25 and 2013-07-07.

Though convenient for the reader, the manual creation of a timeline can take a long time even for experts. For example, the creator of the startup Timeline says that it initially took a multi-person team a full work day to create a single timeline.[1] Therefore, automatic timeline summarization (TLS) has emerged as an NLP task in the past few years (Tran et al., 2013a; Kessler et al., 2012; Nguyen et al., 2014; Yan et al., 2011b; Yan et al., 2011a; Wang et al., 2012; Tran et al., 2013b; Tran et al., 2015). TLS has been divided into two subtasks: (i) ranking the dates between beginning

[1] http://www.niemanlab.org/2015/02/timeline-is-providing-historical-context-to-the-news-but-is-there-a-business-model-to-support-it/.

*Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics
and the 7th International Joint Conference on Natural Language Processing*, pages 1598–1607,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

and end of the timeline in order of importance, to achieve date selection and (ii) generating a good daily summary for each of the selected dates. In this paper, we tackle the first task. Date selection is challenging, as normally only a small set of the available dates is chosen for inclusion in the timeline (see Table 1). Date selection may be partially subjective: different journalists might include different dates.[2]

Existing approaches to date selection (Kessler et al., 2012; Tran et al., 2013a) use supervised machine learning, where each date receives a score for ranking the dates. Features used (such as frequency of date mention) are extracted from a corpus of event-related newspaper articles. Though the features are well-explored, the models score each date independently of other dates.

In contrast, we argue that interaction between dates should be taken into account. Timeline summaries tend to include "substories" in which the majority of selected dates are part of a chain of events that share major actors or demonstrate cause-effect. Table 1 shows at least two such chains: the (a1-4-5) chain of protests leading to Mubarak's resignation and the necessity of new elections, as well as the similar (a8-10-26) chain on Mursi. These chains can also be observed in the corresponding news articles. For example, some background articles on Mubarak's step-down will likely explain the reasons behind it. However, extracting such causal information can be difficult, as demonstrated by the still low results for discourse relation extraction (Lin et al., 2014; Braud and Denis, 2014). Instead, we use *date reference graphs*, which model which date refers to which other date. In our example, articles published on Mubarak's resignation date might refer to the date when the protest started. Although weaker than direct causal links, these links are easy to extract and we will show that they are very useful. In addition, references from important dates (such as Mubarak's resignation date) should be weighted higher than other references. This is akin to IR models such as PageRank, which weigh links from popular pages higher than links from less popular pages.

The main contributions of this work are: (i) we leverage interaction between dates via date reference graphs as a basis for date selection in TLS (ii) we provide a novel random walk model on this graph that incorporates both topical importance of referring sentences as well as frequency and temporal distance of references. We propose both unsupervised as well as supervised versions of this model.

We show that the proposed date selection approach outperforms previous approaches with evaluations on four real-life, long-term news events. We also discuss variations in timeline construction over different events, as well as by different journalists.

2 Related Work

Timeline summarization is a special case of *multi-document summarization* (MDS). As TLS organizes events by date, timelines can be generated by MDS systems (such as (Radev et al., 2004b; Radev et al., 2004a; McKeown et al., 2003; Erkan and Radev, 2004; Metzler and Kanungo, 2008; Hong and Nenkova, 2014) by applying their summarization techniques on news articles for every individual date to create corresponding daily summaries. However, manually written timelines normally only include a small number of dates; in addition, the temporal component imposes constraints on sentence selection for timeline summarization, such as the preference for little overlap between sentences selected for different dates (Yan et al., 2011b).

Many studies specific to timeline summarization, such as (Swan and Allan, 2000; Allan et al., 2001; Chieu and Lee, 2004; Yan et al., 2011b; Tran et al., 2015), focus on the extraction of salient sentences or headlines for generating the textual content of timelines. They assume either that the dates are given in advance or they use simple measures such as *burstiness* (Chieu and Lee, 2004; Yan et al., 2011b) for date selection, where burstiness relies on the number of date mentions.

Prior approaches dedicated specifically to date selection are Tran et al. (2013a) and Kessler et al. (2012).[3] They use supervised machine learning methods that score dates *independently of each other*. Features are extracted from a corpus of event-related newspaper articles, including frequency-based features (such as how often the date is referred to in the corpus), temporal distance features (such as how long into the future a date

[2] Note that the date selection task uses dates as proxies for important events on that date.

[3] Kessler et al. (2012) is also used in Nguyen et al. (2014)'s system.

keeps being referred to) and topical features (such as whether the date mention is associated with the most significant keywords of the event). We, however, score dates *jointly*, making use of interactions between dates in a graphical model. This improves substantially over prior approaches. We also propose unsupervised variations that perform competitively to prior supervised models.

3 Problem Definition and Approach

Similar to Kessler et al. (2012) and Tran et al. (2013a), we use the day as the timeline time unit (so, for example, we exclude hourly timelines).

3.1 Problem Definition

Given a main event and a time window $[t_1, t_2]$ within the event duration, our task is to select the top k dates $(d_1, d_2, ..., d_k) \in [t_1, t_2]$, when the most important (sub)events occurred. Therefore, timelines of variable length can be constructed. Like (Kessler et al., 2012; Tran et al., 2013a), we also assume that we have a corpus $\mathcal{C}$, consisting of news articles about the main event. This corpus gives evidence about the dates in $[t_1, t_2]$.

3.2 Proposed Approach

We build a *date reference graph*, which is a *fully directed graph* $\mathbf{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of dates mentioned in any text in corpus $\mathcal{C}$, including publication dates. The edges $\mathcal{E} = \{e(d_i, d_j)\}$ indicate that at least one text published on d_i refers to the date d_j.

We represent each such link as a multi-value tuple $e(d_i, d_j) = (M_{ij}, freq(d_i, d_j), I_{temporal}(d_i, d_j), I_{topical}(d_i, d_j))$ to integrate different measures of date importance. The first value, $M_{ij} = \frac{1}{N}$ expresses the prior stochastic transitional probability between 2 dates where $N = |\mathcal{V}|$. The others express the strength of the connection between d_i and d_j modelled by the following aspects: frequency ($freq$), temporal influence ($I_{temporal}$) and topical influence ($I_{topical}$). We also suggest different combinations of these parameters.

Then we introduce a random walk model that uses these perspectives to rank the collection of dates.

Frequency of References. When a date d_j is referred to from either a past or future news article (published on d_i), it is likely involved in the events that are reported in that article. An example published on Mubarak's resignation date and referring back to the protest start can be seen below:

(1) On January 25, an uprising of Egyptians erupted calling for Mubaraks resignation as president. Protests continued to grow . . . (CBS Detroit, 2011-02-11)

We hypothesize that the more frequent such references are, the stronger this involvement is. Hence, we compute $freq(d_i, d_j)$ as the number of references to d_j from news articles published on d_i. While prior work (Kessler et al., 2012) uses aggregate frequency of references to d_j over the whole corpus as a feature, they do not handle the interaction between dates and can therefore not score dates jointly.

Topical Influence. In Example 1 above, the reference sentence mentions only major actors in the Egypt crisis (Mubarak, Egyptians) as well as only major subevents (uprising, protests). This makes for a link between 2011-02-11 (publication date) and 2011-01-25 (referred date) that is relevant to the main event and emphasises the importance of the referred date. In contrast, Example 2 also talks about less salient entities in context of the Egypt revolution and makes for a less topical link between 2011-02-02 (publication date) and 2011-01-25 (referred date).

(2) Mr Ghonim is Google's head of marketing for Middle East and North Africa and was in Egypt when the protests started on Jan 25 (DailyMail, 2011-02-02).

We quantify the topical influence between dates as follows: Let $\mathcal{S}_{i\to j} = \{s_{ij}\}$ be the set of sentences that are published in d_i and refer to d_j. We are interested in how relevant this connection is to the overall news event, looking at the content in $\mathcal{S}_{i\to j}$. To do so, we represent the overall content of the news collection by a set of keywords $Q = \{q_1, q_2, ..., q_n\}$, which are computed via TextRank (Mihalcea and Tarau, 2004).[4] We compute a relevance score for each sentence s_{ij} in $\mathcal{S}_{i\to j}$ by the famous Okapi BM25 function (Robertson et al., 1994), which ranks a sentence more topical if it contains more as well as more of the most salient collection keywords Q.[5] We compute topical influence ($I_{topical}$) as either the maximum value or the sum value of the relevance scores of all s_{ij}.

$$I_{\max topical}(d_i, d_j) = \max_{s_{ij} \in \mathcal{S}_{i\to j}} BM25(s_{ij}, Q) \quad (3)$$

$$I_{freq*topical}(d_i, d_j) = \sum_{s_{ij} \in \mathcal{S}_{i\to j}} BM25(s_{ij}, Q) \quad (4)$$

[4] We set n=20 in practice.

[5] We use the standard BM25 parameter settings $k_1 = 1.2$ and $b = 0.75$

Intuitively, $I_{freq*topical}(d_i, d_j)$ is proportional to the size of $\mathcal{S}_{i\rightarrow j}$ as well as to the relevance scores of its sentences whereas $I_{\max topical}(d_i, d_j)$ does not consider reference frequency at all.

When d_j is not mentioned by any articles published on d_i, the value of the topical influence is equal to zero.

Temporal Influence. The longer ago an event happened the more likely it is to have been forgotten. Only very important events are referred to over long time frames. We therefore hypothesise that a date d_j is more influential (for another date d_i) if d_i mentions d_j and the temporal distance between the two dates is high. Overall, d_j gathers importance with several long-term references. Ex. 5 showcases an example:

(5) Military generals took over power from Mubarak when he stepped down on *February 11 last year.* (Daily Mail, 2012-01-25).

We define the temporal influence of an existing edge $I_{temporal}(d_i, d_j)$ as either the absolute value of temporal distance between the two dates or by the product of the temporal distance with the number of references $freq(d_i, d_j)$. In the second computation, the temporal influence between two dates increases when d_i references d_j more than once.

$$I_{|temporal|}(d_i, d_j) = \Delta t = |d_i - d_j| \quad (6)$$

$$I_{freq*temporal}(d_i, d_j) = freq(d_i, d_j) \cdot |d_i - d_j| \quad (7)$$

When d_j is not mentioned by any articles published on d_i, the temporal influence is set as zero.

Random Walk Model for Date Ranking. A random walk on a given graph is a Markov process, where each node represents a state and a walk transiting from one state to another state is based on a transition probability matrix. One well-known random walk algorithm is PageRank (Page et al., 1999), which models web surfer behavior to determine the importance of web pages with the following formula:

$$x_t(j) = \alpha \sum_{i \in L_j^-} M_{ij} x_{t-1}(i) + (1-\alpha) v_j, \quad (8)$$

where M_{ij} is the stochastic transition probability from page p_i to p_j, $x_t(j)$ is the importance score of page p_j at step t, α is a damping factor that controls how often the walker jumps to an arbitrary node, v_j is the initial probabilistic importance score (generally set to $1/N$, where N is the number of nodes in the graph), and L_i^- is the set of incoming links of page p_i. When t is iterated enough, the importance score vector reaches a stationary distribution that can be used for ranking pages.

The traditional PageRank process in Eq. 8 captures only the observed linking characteristics of nodes but ignores other sources of information which can be indicators for their importance.

We extend the model by introducing an influence-based random walk model (**IRW**) that allows the random walker to take into account multiple sources of information and perform voting more effectively. The random walk process we propose can be defined by the following formula:

$$x_t(j) = \alpha \sum_{i \in L_j^-} \mathcal{I}(i,j) \cdot M_{ij} \cdot x_{t-1}(i) + (1-\alpha) v_j \quad (9)$$

where $\mathcal{I}(i,j)$ is the normalized influence factor that indicates how influential the edge $d_i \rightarrow d_j$ is in the global context of the event. The normalization is done by scaling the range of value from [0, 1]. M is the stochastic transitional matrix. In our case, $\mathcal{I}(i,j)$ can be just the value of $freq(d_i, d_j)$, $I_{topical}(d_i, d_j)$, $I_{temporal}(d_i, d_j)$ alone or a linear combination of them. Note that, $(\mathcal{I} \cdot M)$ in most case is not stochastic and *must not* be transformed into a stochastic transitional matrix, as the transformation will collapse the global context of $\mathcal{I}$. IRW is different to PageRank on weighted graph, weighted or personalized PageRank and their variations e.g, (Xing and Ghorbani, 2004; Haveliwala, 2002), among others. In particular, weighted PageRank integrates influence scores into the stochastic transitional matrix. Thus, the random walker contributes the voting impact of a node X to its neighbor with an influence score normalized by the sum of scores on all outgoing connections. That process leverages how good this connection is in the sub-graph (G*) which consists of X and its outgoing neighbors. In contrast, our proposed model uses the non-normalized value of the influence score to leverage how good this connection is on the entire graph instead of G*. To give an example, if date X1 mentions only X2 with a raw temporal distance score of 20 and X3 mentions only X4 with a score of 100, then in weighted Page Rank both would be normalized to a weight one, losing the information that X4 is mentioned after a much longer time period than X2. The process for combination in our model is defined as the following:

$$
\begin{aligned}
x_t(j) = \alpha\omega & \sum_{i \in L_j^-} W_1(i,j) \cdot M_{ij} \cdot x_{t-1}(i) \\
& + \alpha(1-\omega) \sum_{i \in L_j^-} W_2(i,j) \cdot M_{ij} \cdot x_{t-1}(i) \\
& + (1-\alpha) v_j
\end{aligned}
\tag{10}
$$

where $W_1(i,j) = \frac{I_{topical}(d_i,d_j)}{\max_{uv} I_{topical}(d_u,d_v)}$ is the normalized value for topical influence, and $W_2(i,j) = \frac{I_{temporal}(d_i,d_j)}{\max_{uv} I_{temporal}(d_u,d_v)}$ is the normalized value for temporal influence.[6]

Here, the hyper-parameter $0 \leq \omega \leq 1$ controls the proportion of the topical influence from d_i to d_j. When $\omega = 0$, no topical influence is taken into account. No temporal influence is considered when $\omega = 1$. Intuitively, at every step, the random walker can follow the outgoing nodes and either carry topical influence (the first part) or temporal influence (the second part) to contribute to the rank of the outgoing nodes. Otherwise, it teleports to an arbitrary node with probability $(1-\alpha)$.

Convergence Property. Starting from Eq. 10, we now show that the IRW model converges to a stationary distribution.

Let Λ and Λ' be the matrix with elements $W_1(i,j)$ and $W_2(ij)$ respectively, with any edge (d_i, d_j), $\mathbf{I}$ be the $n \times n$ identity matrix, and $\mathbf{v}$ be the transpose of $1 \times n$ uniform stochastic vector. $\mathbf{M}$ denotes the transitional matrix for G.

Proposition 1. $(\mathbf{I} - \alpha(w\mathbf{M}^T\Lambda + (1-\omega)\mathbf{M}^T\Lambda'))$ *is invertible for all* $\mathbf{M}, \Lambda, \Lambda', \alpha, \omega$.

Proof. Let $\mathbf{P} = w\mathbf{M}^T\Lambda + (1-\omega)\mathbf{M}^T\Lambda'$, we need to prove that $\mathbf{I} - \alpha\mathbf{P}$ is invertible. Equivalently, we prove its transpose $\mathbf{I} - \alpha\mathbf{P}^T$ is invertible, which can be proved by showing that $(\mathbf{I} - \alpha\mathbf{P}^T)\mathbf{y} = 0$ only has the trivial solution $\mathbf{y} = 0$.

$$
\begin{aligned}
(\mathbf{I} - \alpha\mathbf{P}^T)\mathbf{y} &= 0 \\
\mathbf{y} &= \alpha\mathbf{P}^T\mathbf{y} \\
y_i &= \alpha \sum_j P_{ji} y_j \\
&= \alpha \sum_j ((\omega W_1(i,j) + (1-\omega) W_2(i,j)) M_{ij} y_j).
\end{aligned}
\tag{11}
$$

Let $u = \arg\max_j y_j$. When $i == u$, Eq. 11 infers,

$$
\begin{aligned}
y_u &\leq \alpha \sum_j ((\omega W_1(u,j) + (1-\omega) W_2(u,j)) M_{uj} y_u). \\
y_u &\leq \alpha y_u \sum_j ((\omega W_1(u,j) + (1-\omega) W_2(u,j)) M_{uj} \\
y_u(1 - \alpha\mathcal{F}_u) &\leq 0.
\end{aligned}
\tag{12}
$$

where $\mathcal{F}_u = \sum_j ((\omega W_1(u,j) + (1-\omega) W_2(u,j)) M_{uj}$. Clearly, $\mathcal{F}_u \leq 1$ because $W_1(u,j) \leq 1$ and $W_2(u,j) \leq 1$ and $\sum_j M_{uj}$ = 1. Since $\alpha < 1$ and $\mathcal{F}_u \leq 1$, $(1 - \alpha\mathcal{F}_u) > 0$. Therefore $y_u \leq 0$. Similarly, let $v = \arg\min_j y_j$ we have that $y_v \geq 0$. As $y_v \leq y_u$, this implies $y_u = y_v = 0$ to satisfy all inequalities. Consequently, $y_i = 0$ for all i, or $\mathbf{y} = 0$. Thus, $\mathbf{I} - \alpha\mathbf{P}^T$ invertible. Equivalently,$(\mathbf{I} - \alpha(\omega\mathbf{M}^T\Lambda + (1-\omega)\mathbf{M}^T\Lambda'))$ is invertible. □

Proposition 2. *The iteration in Eq. 9 converges to* $(1-\alpha)(\mathbf{I} - \alpha(\omega\mathbf{M}^T\Lambda + (1-\omega)\mathbf{M}^T\Lambda'))^{-1}\mathbf{v}$.

Proof. We can re-write Eq. 9 in matrix form:

$$
\begin{aligned}
\mathbf{x}_t &= \alpha\mathbf{P}\mathbf{x}_{t-1} + (1-\alpha)\mathbf{v} \\
&= (\alpha\mathbf{P})^t\mathbf{x}_0 + (1-\alpha)(\sum_{i=1}^{t} (\alpha\mathbf{P})^{i-1})\mathbf{v}
\end{aligned}
\tag{13}
$$

We will show that $\lim_{t\to\infty} \mathbf{x}_t = (1-\alpha)(I - \alpha\mathbf{P})^{-1}\mathbf{v}$.

$$
\begin{aligned}
\sum_i (\alpha P)_{ij}^t &= \sum_i \sum_k (\alpha P)_{ik} (\alpha P)_{kj}^{t-1} \\
&= \sum_k (\alpha P)_{kj}^{t-1} \sum_i (\alpha P)_{ik} \\
&= \sum_k (\alpha P)_{kj}^{t-1} \alpha(\mathcal{F}_k) \\
&\leq \sum_k (\alpha P)_{kj}^{t-1} \alpha \\
&\leq (\alpha)^t
\end{aligned}
\tag{14}
$$

Here, $\mathcal{F}_k = \sum_i ((\omega W_1(k,i) + (1-\omega) W_2(k,i)) M_{ki} \leq 1$ (proof similarly to Proposition 1

Because $\alpha < 1$, this column sum converges to zero when $t \to \infty$. We then derive $\lim_{t\to\infty} (\alpha\mathbf{P})^t\mathbf{x}_0 = 0$. When $t \to \infty$, given Proposition 1 and Neumann series, Eq. 13 becomes:

$$
\mathbf{x}_t = (\alpha\mathbf{P})^t\mathbf{x}_0 + (1-\alpha)(I - \alpha\mathbf{P})^{-1}\mathbf{v}
$$

hence, $\lim_{t\to\infty} \mathbf{x}_t = (1-\alpha)(I - \alpha\mathbf{P})^{-1}\mathbf{v}$. Convergence proved. □

4 Experiments

4.1 Ground Truth and Data Preprocessing

Kessler et al. (2012) use 91 timelines from AFP as ground truth along with the AFP news corpus for feature extraction. However, their dataset is not publically available. In addition, although they consider a wide spread of events, each event is only represented by a single timeline from a single source, making that method somewhat vulnerable to journalism bias (as discussed by themselves in their paper). The data collected by us previously (Tran et al., 2013a) is publically available at `http://l3s.de/~gtran/timeline/` and has since been extended by us (Tran et al., 2015). Similar to Kessler et al. (2012), it contains ground truth timelines as well as a corpus of news articles covering each event. The dataset is suitable for our purpose because of the following reasons: (1) it is a heterogeneous dataset which contains news articles and expert timeline summaries from different news agencies. Thus, it is more likely to avoid the issue of bias. Also, each event is represented

[6] In the case of linear combinations we incorporate frequency into topical or temporal influence as described above.

by more than one timeline; (2) it covers long-term stories that have been happening since 2011, making the date selection problem non trivial for any system.

Timelines. The groundtruth contains 21 timelines for 4 main events (Egypt Revolution, Libya War, Syria War, Yemen Crisis), created by professional journalists. Table 2 shows statistics about the timelines. Only a small number of all possible dates in a time range is included in at least one timeline (for example, only 122 dates among a possible 918 dates for the Egypt Revolution).

News Corpus. The news articles have been collected from 24 well-known news outlets by querying Google with the event name together with the outlets' sitename and time range specification. The crawl time range starts from the first of the month of the earliest event in any timeline (for example, 2011-01-01 for the Egypt revolution) and ends at crawl date. The top-ranked 300 news articles from each news site were collected, if still available. The article creation date is parsed from the answers returned by Google. The corpus contains 15,534 news articles. Its statistics are summarised in Table 3. The overlap between timeline date ranges and news corpus date ranges is only partial: on the one hand, the corpora have many articles published after the timelines end; on the other hand, sometimes the corpus has no articles published near the beginning of the timeline (Syria War). The distribution of document frequency leans towards the end date of the news collection. The reason could be that most search engines rank recent documents higher than those published longer ago.

Story	Time Range	#News
Egypt	2011/01/11 - 2013/11/10	3869
Libya	2011/02/16 - 2013/07/18	3994
Syria	2011/11/17 - 2013/07/26	4071
Yemen	2011/01/15 - 2013/07/25	3600

Table 3: Overview of the news corpus

Preprocessing. Accurate date extraction including both implicit (like *last Friday*) and explicit (like *11 Feb*) temporal expressions is vital to our approach as well as for competitor systems. We use the Heideltime state-of-the art toolkit (Strötgen and Gertz, 2010) for this task.

4.2 Experimental settings

As can be seen from Table 2, different timelines for the same event can contain varying dates, due to different ranges timelines might cover but also due to selection preferences by individual writers. Therefore, we consider the union of all timelines for an event. The set of input dates for ranking are all dates from the start t_1 and end t_2 of the union of timelines.[7] We call that input time range $\mathcal{TR}_e$, depending on main event e.

We consider two evaluation settings:

relaxed setting: A date from $\mathcal{TR}_e$ selected by an algorithm is counted as correct if it is included in the union of timelines, therefore in at least one individual timeline.

strict setting: A date from $\mathcal{TR}_e$ selected by an algorithm is counted as correct if it is included in at least two individual timelines.

The first setting is the one used in previous work such as (Kessler et al., 2012; Tran et al., 2013a). It is also the only one that can be used if only one timeline per event is considered as in Kessler et al. (2012). We therefore include it for comparison purposes. However, we think it is better to consider several timelines as it allows us to consider agreement between timeline writers. If more than one writer agrees on a date being important we have more evidence that a system should find that date. Finding dates that only a single writer includes is less important and could even be due to bias or system overfitting. Therefore, our second setting is preferable as it emphasizes highly important dates selected by multiple journalists.

Each system selects the top k dates during the input time range. We evaluate the systems by Mean Average Precision at k (MAP@k) for k = 5, 10, 15, 20 over all four events.

4.3 Systems.

Baseline. We use three unsupervised baselines. The baseline *Document Frequency* ranks dates according to the number of news articles published on that date. Our assumption is that on a date where one or more important events happened, there would be a spread of information over different news agencies in the world. Therefore, this date has more news articles published. This baseline is related to the burstiness date selection used by Yan et al. (2011b).

The baseline *MaxLength* ranks dates by the maximum article length of all articles published on that date. Our hypothesis is that important events

[7] Prior work also uses start and end date of timelines for delimiting input (Kessler et al., 2012; Tran et al., 2013a).

Story	#TL	#atLeastOnce	#atLeastTwice	avgL	maxL	minL	Time Range	#dates
Egypt	4	122	18	36	57	24	2011/01/01 - 2013/07/07	918
Libya	7	118	56	34	62	22	2011/02/14 - 2011/11/22	281
Syria	5	106	17	60	26	13	2011/03/15 - 2013/07/06	844
Yemen	5	81	26	24	42	10	2011/01/22 - 2012/02/27	401

Number of timelines (#TL), number of dates occurring in at least one timeline (#atLeastOnce), number of dates that appear in at least 2 timelines, average (avgL), max (maxL) and min (minL) length of timelines; the Time Range of the union of timelines and all potential dates (#dates) within the time range.

Table 2: Overview of groundtruth timelines

often receive more attention from writers, leading to longer articles.

Date Frequency ranks a date d by the total number of sentences referring to d that are not published on d. This is a simple measure of d's influence without joint scoring of dates or integration of temporal distance or topic.

Competitors. We reimplement *Kessler et al. (2012)*'s model. It first detects all sentences with date references and filters out certain types of sentences according to linguistic features (such as presence of modality as this can put the factuality of the event into question). Then, the importance score of a date is determined by the product of the Lucene score of referring sentences and an ML-predicted score that takes into account date reference frequencies, temporal distance of date references and topical importance of referring sentences. To use the same setting as for our systems, we use the list of keywords extracted by TextRank (Mihalcea and Tarau, 2004) to formulate a topic query for the Lucene index.

We reimplement *Tran et al. (2013a)* who use a supervised ML approach based on a more detailed consideration of date reference frequencies.

Both *Kessler et al. (2012)* and *Tran et al. (2013a)* are retrained and tested via 4-fold cross-validation on events. In addition, we noted that the two supervised systems could profit from the fact that for certain dates in $\mathcal{TR}_e$ no published news articles exist in the news collection and that they are therefore a priori unlikely to be relevant. We therefore also run those systems with a stricter input time range, which intersects $\mathcal{TR}_e$ with the dates that are the publication date of at least one article in the news collection. We indicate these systems as *Kessler et al. (2012) (Pub)* and *Tran et al. (2013a) (Pub)*.

Our Approach. Our system builds graphs with all dates referenced in the news corpus for an event as nodes. We select the top k highest ranked nodes that also fall within $\mathcal{TR}_e$. We measure the performance with different strategies for the *Influence* factor $\mathcal{I}$. We use the following five unsupervised strategies, where we just set the damping factor α to 0.85 as suggested by Page et al. (1999).[8]

IRW_{freq} only uses the frequency aspect. This corresponds to a joint modelling version of the *Date Frequency* baseline.

$IRW_{\max topical}$ uses topical influence, disregarding frequency aspect in its computation.

$IRW_{freq*topical}$ uses topical influence, incorporating the frequency aspect in its computation.

$IRW_{|temporal|}$ uses temporal influence, disregarding the frequency aspect.

$IRW_{freq*temporal}$ uses temporal influence incorporating the frequency aspect.

Furthermore, we are interested in combining topical and temporal influence (with or without frequency aspects). Here, our model is parameterized by ω which controls the impact of *topical influence* vs. *temporal influence*. This parameter is tuned on the training set via 4-fold cross-validation and, therefore, the next two models have a small element of supervision.

$IRW_{\max topical+freq*temporal}$ combines topical and temporal influence, integrating the frequency aspect into temporal influence.

$IRW_{freq*topical+|temporal|}$ combines topical and temporal influence, integrating the frequency aspect into topical influence.

4.4 Analysis of date reference graphs

Table 4 shows an analysis of the four date reference graphs. In this Table, #sent provides the total number of sentences from all news articles while #hasRef shows the number of sentences that refer to a date (around 15%), suggesting a sustainable part of data can be helpful for the interaction-based approach. The number of nodes shows the unique dates that are involved in a date reference link. The number of edges is equivalent to the number of date reference links (d_i, d_j) that indicate that there exist sentences published on d_i but referring to d_j. *toStrict* and *toRelaxed* is the

[8] We could make these models supervised by tuning the damping factor via cross-validation. However, we found it encouraging that we were able to achieve competitive results without tuning — similar to links between web pages in the traditional PageRank algorithm, links between dates seem to embody strong relations, making the same damping factor suitable.

	#sent	#hasRef(%)	#nodes	#Edges	toStrict	reachStrict	toRelaxed	reachRelaxed
Egypt	143,096	26,428 (18.5)	939	2784	15.55%	100.00%	35.99%	89.34%
Libya	140,753	22,166 (15.7)	971	1797	33.78%	98.21%	56.98%	99.15%
Syria	162,305	26,992 (16.6)	812	1555	7.14%	88.24%	31.00%	73.58%
Yemen	140,156	21,606 (15.4)	1106	1608	18.28%	100.00%	37.00%	100.00%

Table 4: Interaction-based analysis on experimental news collections

proportion of the edges that link to groundtruth dates in the *strict setting* and *relaxed setting*. Those edges cover almost all the groundtruth dates (i.e, *reachStrict* and *reachRelaxed*), i.e almost all groundtruth dates are indeed referenced at least once in our corpus.

4.5 Results

Table 5 shows the average performance of different systems over our four events. Several general observations stand out. First, we notice that the scores wrt. *relaxed setting* of all systems are higher than those wrt. *strict setting*. That is expected, as in *relaxed setting*, a selected date has a higher likelihood to be one of the milestones in the timeline of at least one expert. Second, simple baselines such as *Document Frequency* and *MaxLength* perform reasonably well in the relaxed-setting. That confirms our assumptions that important dates often possess more published news articles and are likely to have at least one article of substantial length. However, these baselines are not enough to distinguish highly important dates (which are selected by more than one journalist) as shown by their performance in the *strict setting* (around 0.3 MAP@k only).

Using *Date Frequency* leads to a substantial performance improvement in the strict setting comapred to the publication-based baselines. Therefore, highly important dates are more likely to be kept mentioning in the future and that supports our research direction to better leverage date interaction for ranking date importance. This is further confirmed by the performance of the IRW_{freq} system which is the joint modelling version of the *DateFrequency* baseline and outperfoms the baseline without inclusion of any further information such as topical salience. It can even compete with prior supervised competitors when their input time range is not modified.

Our supervised competitors (Kessler et al., 2012; Tran et al., 2013a) perform overall well and both profit from modifying their input time range as suggested in the *Pub* versions. However, the unsupervised versions of our system $IRW_{\max topical}$ and $IRW_{freq*topical}$ perform very comparably to the supervised competitors in the strict and relaxed setting, respectively.

The last two lines of Table 5 show the results of our proposed method when using a linear combination of the different influence factors, and the hyperparameter ω having been tuned on the training set. $IRW_{\max topical+freq*temporal}$ shows the result of our system with $\omega = 0.2$ and $IRW_{freq*topical+|temporal|}$ with $\omega = 0.1$ These systems outperform the state-of-the-art systems clearly in the strict setting and for most measures in the relaxed setting.

Stability. We also investigated the stability of the performance of different systems by looking into their results on each event. Table 6 presents the performance of our best system $IRW_{\max topical+freq*temporal}$ and its best supervised competitors Tran et al. (2013a) (Pub) and Kessler et al. (2012) (Pub). All systems perform worse on the Syria story although our dropoff is less than the one of prior systems.

We speculate that the competitor systems are more sensitive to the amount of available published content on a target date than ours. In particular, Tran et al. (2013a) use the frequency of published dates and sentences as one of their features, and Kessler et al. (2012) rely on the returned results from Lucene index which tends towards substories from the publication periods. Different to others, the time range for the *Syria* news collection does not include the time range for the Syria timelines fully or almost fully (see Tables 2 and 3). We therefore are not as dependent on an exact match between timeline dates and news collection dates and can use news articles from later dates more effectively.

5 Conclusion and Future Work

This paper addresses the problem of date selection for timeline summarization. Our approach leverages the interactions between dates via a joint model based on a date reference graph, improving on individual scoring of dates.

We capture the interactions between dates from the number of cross-references between dates, and

System	strict setting				relaxed-setting			
	MAP@5	MAP@10	MAP@15	MAP@20	MAP@5	MAP@10	MAP@15	MAP@20
Document Frequency	0.312	0.303	0.299	0.299	0.509	0.550	0.564	0.560
MaxLength	0.349	0.335	0.311	0.287	0.647	0.594	0.566	0.533
Date Frequency	0.555	0.498	0.457	0.427	0.597	0.626	0.625	0.613
(Kessler et al., 2012)	0.567	0.546	0.519	0.491	0.790	0.740	0.723	0.704
(Kessler et al., 2012) (Pub)	0.701	0.620	0.571	0.524	0.912	0.807	0.759	0.731
(Tran et al., 2013a)	0.668	0.565	0.522	0.488	0.740	0.717	0.700	0.673
(Tran et al., 2013a) (Pub)	0.710	0.601	0.551	0.506	0.792	0.771	0.746	0.716
IRW_{freq}	0.646	0.535	0.471	0.431	0.861	0.770	0.711	0.687
$IRW_{\max topical}$	0.763	0.647	0.564	0.510	0.887	0.794	0.724	0.685
$IRW_{freq*topical}$	0.737	0.576	0.498	0.448	0.945	0.836	0.762	0.709
$IRW_{\|temporal\|}$	0.724	0.587	0.522	0.484	0.699	0.597	0.570	0.564
$IRW_{freq*temporal}$	0.724	0.588	0.527	0.486	0.712	0.622	0.581	0.559
$IRW_{\max topical+freq*temporal}$	0.879	0.760	0.658	0.587	0.897	0.842	0.775	0.730
$IRW_{freq*topical+\|temporal\|}$	0.818	0.677	0.596	0.536	0.928	0.866	0.801	0.745

Table 5: Average MAP@k scores of different systems on 4 news collections

	Egypt	Libya	Syria	Yemen
	$IRW_{\max topical+freq*temporal}$			
MAP@5	0.960	1.000	0.713	0.843
MAP@10	0.738	0.969	0.598	0.735
MAP@15	0.600	0.854	0.503	0.676
MAP@20	0.520	0.776	0.433	0.619
	Kessler et al. (2012) (Pub)			
MAP@5	0.703	0.843	0.257	1.000
MAP@10	0.566	0.759	0.203	0.952
MAP@15	0.507	0.697	0.187	0.894
MAP@20	0.450	0.659	0.171	0.816
	Tran et al. (2013a) (Pub)			
MAP@5	0.960	0.910	0.257	0.713
MAP@10	0.803	0.836	0.224	0.541
MAP@15	0.665	0.799	0.227	0.514
MAP@20	0.569	0.758	0.212	0.484

Table 6: Stability of our systems vs. competitors

their temporal and topical influences. We present a novel random walk model that incorporates these perspectives into connectivity-based computation. Experimental results on four news events that span a long time period show that the proposed models outperform state-of-the art approaches. Even unsupervised versions of the model perform on a par with previous supervised methods. We also draw attention to the necessity to take personal bias into account, which leads to differences between manually created timelines for the same event — we encourage future work to always consider several timelines per event in the way that other NLP work uses several annotators to create ground truth.

In future work, we will consider a wider range of events and event types. This will also lead to considering timelines where the day as unit of granularity might not be appropriate or where the unit of granularity might be varying across the timeline. We will also explore in depth the effect of size and type of news corpus on resulting timelines, research further into the issue of human disagreement in timeline creation and explore human evaluation of timeline summarization.

Acknowledgments

The work was partially funded by the European Commission for the FP7 project EUMSSI (611057) and the ERC Advanced Grant ALEXANDRIA (339233).

References

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Proceedings of SIGIR'01*, pages 10–18.

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *COLING 2014: Technical Papers. Dublin, Ireland*, pages 1694–1705.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of SIGIR'04*, pages 425–432.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

John Foley and James Allan. 2015. Retrieving time from scanned books. In *Advances in Information Retrieval*, pages 221–232. Springer.

Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *WWW*, pages 517–526.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of EACL 2014*, pages 712–721.

Remy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. 2012. Finding salient dates for building thematic timelines. In *Proceedings of ACL*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Kathleen McKeown, Regina Barzilay, John Chen, David K. Elson, David Kirk Evans, Judith Klavans, Ani Nenkova, Barry Schiffman, and Sergey Sigelman. 2003. Columbia's newsblaster: New features and future directions. In *HLT-NAACL*.

Donald Metzler and Tapas Kanungo. 2008. Machine learned sentence selection strategies for query-biased summarization. In *Proceedings of the 2008 ACM SIGIR LR4IR Workshop*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*, pages 404–411.

Kiem-Hieu Nguyen, Xavier Tannier, and Véronique Moriceau. 2014. Ranking multidocument event descriptions for building thematic timelines. In *COLING 2014*, pages 1208–1217.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, November. Previous number = SIDL-WP-1999-0120.

Dragomir R. Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drbek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004a. Mead - a platform for multidocument multilingual text summarization. In *Proceedings of LREC'04*.

Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004b. Centroid-based summarization of multiple documents. pages 919–938.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *TREC*.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the SemEval '10 Workshop*, pages 321–324.

Russell C. Swan and James Allan. 2000. Timemine: visualizing automatically constructed timelines. In *SIGIR*, page 393.

Binh Giang Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013a. Predicting relevant news events for timeline summaries. In *WWW'2013*.

Giang Binh Tran, Tuan A. Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013b. Leveraging learning to rank in an optimization framework for timeline summarization. In *SIGIR 2013 Workshop TAIA*.

Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevent headlines. In *Proceedings of ECIR'2015*.

Dingding Wang, Tao Li, and Mitsunori Ogihara. 2012. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *Proceedings of AAAI'2012*.

Wenpu Xing and Ali A. Ghorbani. 2004. Weighted pagerank algorithm. In *CNSR*, pages 305–314.

Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011a. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of EMNLP'11*.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011b. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of SIGIR'11*, pages 745–754.