Timeline Summarization from Relevant Headlines

Giang Tran, Mohammad Alrifai, and Eelco Herder

L3S Research Center & Leibniz University Hannover
Appelstr. 9, 30167 Hannover, Germany
{gtran,alrifai,herder}@L3S.DE

Abstract. Timeline summaries are an effective way for helping newspaper readers to keep track of long-lasting news stories, such as the Egypt revolution. A good timeline summary provides a concise description of only the main events, while maintaining good understandability. As manual construction of timelines is very time-consuming, there is a need for automatic approaches. However, automatic selection of relevant events is challenging due to the large amount of news articles published every day. Furthermore, current state-of-the-art systems produce summaries that are suboptimal in terms of relevance and understandability. We present a new approach that exploits the headlines of online news articles instead of the articles' full text. The quantitative and qualitative results from our user studies confirm that our method outperforms state-of-the-art system in these aspects.

1 Introduction

More than two years after the Egyptian revolution of 2011, political conflicts in Egypt were back again in the breaking news headlines in 2013. While trying to relate current events to past events, newspaper readers may ask themselves several questions, such as: *How and Why did the Egyptian revolution start back in 2011? What happened in Egypt since then? Why are there many new protests again in Egypt?* A compact summary that represents the development of the story over time, highlighting its most important events - possibly with links to sources for further details - would be very beneficial for fulfilling readers' information needs.

Timeline summarization (*TS* for short) has become a widely adopted, natural way to present long news stories in a compact manner. News agencies often manually construct and maintain timelines for major events, but constructing such visual summaries often requires a considerable amount of human effort and does not scale well. Existing approaches for TS aim to tackle one of two problems: (i) select a subset of important dates as the major points of the timeline (e.g, [12], [4]) and/or (ii) generate a good daily summary for each of these dates (e.g, [6], [27], [4]). In this study, we set our focus on the second problem.

Previous work on the generation of daily summaries usually focuses on the extraction of relevant sentences from article text. The main drawback of such approaches is that it does not guarantee good *understandability* as well as high *relevance* for the daily summary. Low relevance is often caused by the nature of textual data - it is hard to select the right sentence from a large number of sentences; low understandability is

(A.1) It will end as soon as the people vote on a constitution, **he** told state television...

(A.2) ...**President Mohamed Mursi** hopes will help to end a crisis..."

(B.1) On **Wednesday** , two protesters were killed Aden , a southern port city.....

(B.2) On **Thursday** , dozens of people were reportedly injured in clashes.....

(C.1) Anti-government protesters in Yemen have resumed demonstrations to try to force Ali Abdullah Saleh , the president , to quit , ...

(C.2) The students , some of whom were also armed with batons , responded .

Table 1. Examples of summaries with low understandability

often caused by inconsistencies and lack of continuity between the selected sentences. The following examples in Table 1 present 3 summaries generated by a state-of-the-art system, ETS [27], showing a few understandability problems: "he" in sentence (A.1) is ambiguous and can be misunderstood as "Mohamed Mursi" in (A.2) (*daily summary A)*, time inconsistency between sentences (B.1). and (B.2.), which should not be used in the same daily summary *(daily summary B)* and content incoherence between (C.1) and (C.2) of *daily summary C*.

In addition to this, finding a good order for selected sentences to make a coherent summary is on itself already a difficult task in the NLP community (for example, see [3], [5]). This makes it even more challenging to generate a summary with good understandability by ordering selected sentences.

Headlines of online news articles have shown to be a reliable source for adequately providing a high-level overview of the news events[2]. Headlines are comprehensible to the reader without requiring too much reading time [20],[7]. The information provided in headlines is usually self-contained, timely and complete, and therefore suitable for creating coherent daily summaries. For this reason, we consider headlines as good candidates for TS generation.

There are some technical challenges that make using news headlines for TS far from being straightforward. First, one needs to distinguish informing news updates from other non-informing news headlines, which includes background information, reviews and opinions[1]. In this work, we focus on informing news headlines, which tell *what* happens in the story instead of opinions or background. Second, one needs to identify duplicates among the headlines, to minimize redundancy in the produced summary. Because headlines are often short and do not follow syntactic structures, duplicate detection among headlines is a challenging task. Third, one needs to make a selection of the most relevant headlines for making daily summaries that are as informative as possible. To our knowledge, there are no previous studies on generating TS from headlines.

The contribution of this paper is a novel approach for the generation of timeline summaries of news stories, based on the headlines of news articles. We present a *headline selection algorithm* based on a random walk model (Section 3). Further, we show the results of *quantitative and qualitative evaluations* of the proposed methods in comparison with the-state-of-the-art methods (Section 4)

[1] see Freund et al. (2011) [10] for news genre taxonomy

2 Related Work

There is a plethora of research on the generation of timeline summaries. Typical studies in this domain include Swan and Allan [24], Allan et al. [1], Chieu at al.[6], Yan et al. [27], Tran et al. [4]. These studies share the same approach of extracting the most relevant and descriptive sentences from the full article texts. Experimental evaluations of these approaches have shown that the n-gram overlaps (*typically using ROUGE scores*) between the generated summaries and some manually created summaries for the same time period (or dates) is significant. Nonetheless, to the best of our knowledge, none of these approaches has been evaluated using qualitative analysis.

Our assumption is that the full-text extraction approach that is adopted in the aforementioned research works does not guarantee the (subjective) quality or readability of the produced summaries, as this cannot be measured using the ROUGE score. We use a different approach, directly based on the news article headlines. Our qualitative user evaluation shows that users tend to rate the summaries produced by existing solutions with lower quality scores.

Timeline summarization is a special case of *multi-document summarization* (MDS for short), which organizes events by date. Basically, TS can be generated by MDS systems by applying summarization techniques on news articles for every individual date to create a corresponding daily summary. However, because MDS techniques do not make use of the inter-date connections between news articles, they tend to be less robust than state-of-the-art methods specifically designed for TS generation (e.g., as discussed in[27]). Beside the difference in the approach (using headlines instead of the full text), our framework differs from MDS in that it takes the relations among events across dates into account. As there is already a rich body of research on multi-document summarization (for example,[22],[21], [16], [8], [17]), in this study we also investigate how good they are in producing daily summaries using only headlines, in the same setting as our approaches.

3 Problem Statement and Selection Model

The focus of this study is on generating timeline summaries that represent *what* happened in a news story. More formally, we focus on the following problem:

Problem 1 (Selection of Headlines for TS.). Let H_d be the set of headlines from published news articles of a dated, select c most relevant headlines to make daily summary of that date.

In this section, we discuss aspects of headlines that are relevant for the creation of TS: the headline's Informing value, its Spread and Influence. After that, we develop a random walk model based on personalized Pagerank on the top of these aspects. In summary, the model estimates duplicates among headlines (the *Spread*) and creates a graph in which the nodes represent the headlines and the edges are weighted by the probability that two corresponding headlines are duplicated. The model biases the random walker to prefer headlines with high *Influence* scores. Finally, we conduct a greedy algorithm based on submodularity to select a set of relevant headlines using the *Informing* aspect and the backward probabilities (i.e, rank).

3.1 Aspects of Relevant Headlines

In this section, we describe three important aspects that characterize relevant headlines: their Informing value and their Spread and Influence.

Informing. We consider a headline as an Informing news headline when it informs about a news event[2] An Informing headline typically delivers self-contained information to the readers, as it explicitly describes an event that has occurred. By contrast, non-informing news headlines often provide author opinions or reviews on the event. Although opinions or reviews are helpful in highlighting different aspects of the events, especially when they come from influential columnists, they are typically provide opinionated, subjective views of the authors and hence introduce some bias to the TS. We leave opinion-based TS for another study.

We calculate the Informing aspect by using a machine learning classification approach. For the sake of simplicity, we follow Yu and Hatzivassiloglou [28] as it performed well on our testing set. Let F(h) denote the probability of a headline *h* being an informing news headline. When a headline *h* is classified as *positive*, we assign F(h) = 1, otherwise F(h) = 0. For training purposes, we use 20K headlines as positive examples that are randomly extracted from news articles using APIs of the *WikiTimes*[3] system [25]. Those news articles are references to actual events in the Wikipedia Current Events portal [4]. In contrast, negative examples are 20K headlines of articles from the New York Time corpus that are annotated as opinion, reviews or other non-informing categories until 2007. By using these two sets of headlines for training the SVM model, instead of sentences from the full text of news articles, the machine learning model is fitted well with our headline input. Our experimental results show that the model reaches 76% accuracy by cross-validation. Due to space limitations, we do not go further into details.

Influence. An event is likely to be relevant for timeline summaries when it is influential in what will happen in the future. For example, *Mubarak resigns* will lead to a *new election event*, then lead to the *presidency of Mohamed Mursi*, and so on. We observed that influential events are those that are most often mentioned in news articles that are published in the future.

We compute Influence as follows. Let I(h) quantify the influence of headline *h*. We analyze temporal information in the content of the respective news article to heuristically locate references to this particular headline in news articles that are published after that. Let $\mathcal{E}_{V \to u}$ be the cluster of all sentences that are not published in u but refer to date u. Using the Heideltime toolkit [23] for temporal tagging, given a headline h of date u, we define its influence on future events by the similarity of its word distribution, $\theta(h)$

[2] We only focus on actual news stories, not on other articles such as Photo essays, Infographics or Weather reports.

[3] `http://wikitimes.l3s.de`

[4] `http://en.wikipedia.org/wiki/Portal:Current_events`

to the word distribution of the cluster $\theta(\mathcal{E}_{V\rightarrow u})$. The computation is done as follows:

$$I(h)_u = \sum_{w \in h} p(w|\theta(h)) * p(w|\theta(\mathcal{E}_{V\rightarrow u})) \tag{1}$$

where $p(w|\theta)$ is probability of word w in θ.

Spread. Cluster hypothesis suggests that headlines that are similar to one another confirm the relevance of each other [26], as they are virtually members of the same clusters. We observed that a relevant event is typically spread among various headlines, as it is very often reported by different news agencies. The following example shows how the event "Mubarak resigns" is reported in different headlines:

- **Huffington Post:** *Mubarak Steps Down Tahrir Square , Egypt Erupts In Cheers.*
- **The Guardian:** *Hosni Mubarak resigns and Egypt celebrates a new dawn.*
- **CNN:** *Egypt's Mubarak resigns after 30-year rule.*
- **NBC:***'Egypt is free,' crowds cheer after Mubarak quits.*

We quantify the *Spread* of a headline by measuring p_{ij} as *the probability that two headlines* h_i *and* h_j *are duplicated* (i.e., they report about the same event). Intuitively, more duplications and higher confidence (by mean of probability) indicate higher *Spread* value. Obviously, *Spread* is transitive: h_i and h_k may be duplicated if they both are duplicated with h_j. Due to this transitivity, the *Spread* value of a headline can be propagated through its duplicated headlines. Using this graph, a random walk model on the duplication graph of headlines is able to estimate the *Spread* value. We will present an algorithm for that estimation in a combined model with other aspects in Section 3.2.

Estimation of Duplication Probability . Now we describe how we computed the duplication probability p_{ij} using a Logistic Regression model. It is worth mentioning that even though this task is similar to sentence paraphrase detection, headlines are of shorter length and sometimes do not follow grammatical rules (but are fancy and catchy). In addition, here we only care about the core message reported in the headline, while in sentence paraphrase detection, the meaning of the entire sentence is taken into account. That makes available labeled corpora for sentence paraphrase detection not a good fit for our learning strategy. Therefore, we constructed our own training data by leveraging the wisdom of the Wikipedia crowd. Due to space limitations, we will only summarize the steps we followed: (1) extract positive examples: pairs of headlines from any pair of news articles on *an event* in Wikipedia's current events portal[5] ; (2) extract negative examples: pairs of *cross-event* headlines (i.e, each headline is from an event). In the end, we obtained a dataset of 16K with a ratio between positive and negative examples of about 50/50. Our intuition is that headlines of news articles that are references of an event are likely to be duplicated.

For training the Logistic Regression model, we use state-of-the-art semantic similarity measures that are popular in paraphrase detection: corpus-based similarity, as proposed by Mihalcea et al. [18] and Malik et al. [15], and Wordnet-based paraphrase similarity [9]. In addition, we extracted prior co-occurrence probabilities of any verb

[5] to save the engineering cost, we use *WikiTimes* data: `http://wikitimes.l3s.de`

pair in the whole WikiTimes dataset as a signal for two corresponding headline pairs being duplicated. A verb pair is counted as one co-occurrence if both verbs appear in two headlines of the same event. That model results in 77% accuracy with 10% improvement gained by additionally using prior co-occurrence probability feature.

3.2 Headline Selection Model

Overview. Our target is to select headlines that maximize all three aspects Influence, Spread and Informing value. Among available propagation algorithms, personalized PageRank [11] on a graph of headlines appears to be suitable for this task, as it both takes the link graph structure (Spread aspect) into account and considers the personalized probability (Influence aspect) while performing random walks. Then, by using PageRank score as the probability of being relevant for TS, we formulate headline selection as an optimization problem that can be solved by submodular-based techniques, which we describe in the remainder of this section.

Formation of Headlines Graph From the set of headlines $H = \{h_1, h_2, ..., h_n\}$ of a day, we create an undirected event-based similarity graph G = (E, V), in which each node of V is a headline in H and each edge between 2 nodes (i, j) is weighted by the duplication probability $p_{ij} \in [0, 1]$.

Influence-based Random Walk In order to integrate the multiple aspects of the headline, we use a random walk model that follows the personalized PageRank method for ranking headlines. Headline relevance (R) is estimated by its probability of being visited by the random walker in the model, which is iteratively computed using the equation 2.

$$R(j) = d \sum_i \frac{p_{ij}}{\sum_k p_{ik}} * R(i) + (1 - d) * \frac{I(h_j)}{\max_{h \in H} I(h)} \tag{2}$$

where the damping factor d = 0.85 and the transitional probability is normalized from the duplication probability to satisfy the Markov property. We guide the random walker to headlines that have high influence scores *I(h).*

Algorithm 1: Algorithm for selection of relevant headlines

```
$S \leftarrow \emptyset$
$Q \leftarrow H$
while $Q \neq \emptyset$ and $|S| < c$ do
$h_i \leftarrow \arg\max_{h \in Q} \mathrm{R}(S \bigcup h) - \mathrm{R}(S)$
$p(h_i, S) \leftarrow \max_{h_j \in S} p_{ij}$
subject to: $p(h_i, S) < \theta \wedge F(h_i) = 1$ (#no duplication and be informing news)
$S \leftarrow S \bigcup h_i$
$Q \leftarrow Q \setminus h_i$
end while
```

Submodular method for selecting events. Based on the R scores of all headlines, we greedily select the top headlines as long as they do not violate any constraint: no pair of selected headlines is duplicated and selected headlines must be informing. Formally, we have to solve the following optimization problem:

$$\begin{aligned} \underset{S \subseteq H_d}{\text{maximize}} \quad & \mathrm{R}(S) \\ \text{subject to} \quad & \mathrm{R}(S) = \sum_{h_i \in S} \mathrm{R}(i) \\ & |S| = c \\ & F(h) = 1 \; \forall h \in S \\ & p_{ij} < \theta \; (i,j) \in \Omega_S \end{aligned}$$

Given our constrains: (a) no duplicated pairs in selected subset $S \subset H$, (b) budget size(S) = **c** as the number of headlines for each day summary, and (c) all selected headlines should be Informing news headlines. Our objective function is monotone and submodular [13], and therefore we may use the greedy Algorithm 1 to solve it with accuracy guarantee $1 - \frac{1}{e}$, where θ is the threshold for identifying whether one pair is duplicated, determined by the trained Logistic Regression model for duplication probability estimation.

4 Experiments

In this section, we evaluate the proposed framework by measuring the *relevance* and *understandability* of the TS output and comparing it to that of state-of-the-art systems. Our evaluation methodology is based on *human evaluation* instead of automatic n-gram based overlap metrics like Rouge scores [14], especially because headlines exhibit different characteristics than article text sentences, and n-gram based measures hardly capture paraphrases in event reporting, for instance, *'Egypt is free,' crowds cheer after Mubarak quits.* v.s *Hosni Mubarak resigns and Egypt celebrates a new dawn.*

The relevance score measures how well the selected headlines perform in reporting and summarizing important events of the news story, while the understandability score measures the readability and comprehensibility of the summary that is constructed from the selected headlines. In other words, we consider one summary better than another if it covers more relevant events and/or if users understand its description of the events better.

4.1 Dataset and Experimental Setting

We constructed a dataset that consists of news articles, which serve as input, and expert timeline summaries, which serve as ground-truth summaries (the ideal output). The articles focus on long-span stories on the *armed conflicts* Egypt Revolution, Syria War, Yemen Crisis and Libya War [6].

News articles We collected news articles by simulating users searching for articles relevant for the timelines of the aforementioned news stories - for this purpose, we used Google and targeted the same 24 news agencies that were used for creating the timelines used as the ground truth. We constructed several queries, such as “Egypt (revolution OR crisis OR uprising OR civil war)”, as queries with the time filter option [Jan/2011 - July/2013] and the “site” specification. For each query, we took the top-300 answers. Using this method, we obtained 15.534 news articles, of which the distribution is summarized in the #News column of Table 2.

[6] Available at `http://www.l3s.de/~gtran/timeline/`

Expert Timeline Summaries Arguably, timeline summaries that have been published by well-known news agencies are the most trustful base for ground-truth timeline summaries, as they have been manually created by professional journalists. We manually collected 25 timeline summaries from 24 populair news agencies, including the BBC, CNN and Reuters. These ground-truth timeline summaries are offered to the participants of our study as a baseline for deciding whether the automatically selected headlines are relevant or not. Table 2 gives an overview of these timelines.

Story	#TL	#Timepoint	#GT-Date	TL-Range	#a.sent	#News
Egypt Revolution	4	112	18	01/2011-07/2013	2	3869
Libya War	7	118	51	02/2011-11/2011	2	3994
Syria War	5	106	15	03/2011-08/2012	2	4071
Yemen Crisis	5	81	22	01/2011-02/2012	2	3600

Number of timelines (#TL), total number #Timepoints of all timelines, number of groundtruth dates(#GT-Date), the time ranges and rounded average sentences per date of each timelines (#a.sent.), number of news articles (#news)

Table 2. Overview of expert timeline summaries

4.2 Systems for comparison

We compare our approach with systems for TS generation as well as for traditional MDS. In addition, we consider two other baselines, SumSim and Longest. To make the generated summaries comparable with expert summaries in term of length, we use the same setting $c = 2$ for all systems in our evaluation, which means that each system will generate daily summaries of 2-sentence length.

Timeline Summarization. We choose two state-of-the-art methods for TS generation that focus on daily summaries: ETS and Chieu et al. Both systems have originally been designed to work with article texts. However, in addition to that, we developed one version of Chieu et al. for headlines only, named SumSim. Due to the design of the algorithm and the spare word distribution, it is not easy to adapt ETS to work with just headlines. We leave it for future investigation.

ETS is by far one of the best unsupervised TS systems in the news domain. It takes advantage of the similarity between the word distributions in a sentence and the word distribution in an entire corpus as well as within the neighboring dates. We implemented the ETS algorithm described by the authors in [27].

Chieu et al.[6] utilize the popularity of a sentence on date t_i as the sum of TF-IDF similarity scores with other sentences that are published in around t_i +- k days to estimate how important this sentence is. We select k=10, following the author's setting.

Traditional Document Summarization. Since ETS and Chieu et al. extract sentences from the full text of news articles for timeline summaries, as shown in the experiments of Yan et al. [27], we also would like to see how good (multi-)document summarization would work on the headlines dataset. We consider the following state-of-the-art methods: Centroid [22], LexRank [8], TextRank [19] [7]

[7] for Centroid, we used the MEAD toolkit, for LexRank and TextRank, we used the sumy toolkit https://github.com/miso-belica/sumy.

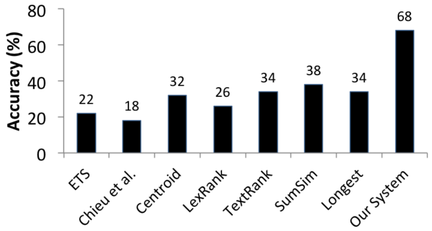
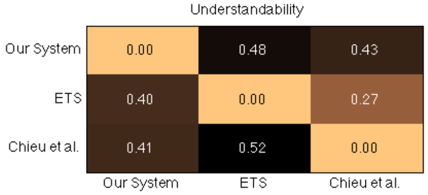


Fig. 1. Relevance evaluation of the produced summaries by the different systems in comparison with expert manual summaries


Understandability

	Our System	ETS	Chieu et al.
Our System	0.00	0.48	0.43
ETS	0.40	0.00	0.27
Chieu et al.	0.41	0.52	0.00

Fig. 2. Pairwise comparison of the *understandability* of summaries produced by the different systems

SumSim selects top news reports and non-duplicated headlines that maximize the sum of TF-IDF similarity with other headlines that are published in the previous and next 10 days. Conceptually, it works similarly to Chieu et al., but on the headline level.

Longest selects top news reports and non-duplicated headlines ordered by their length. Conceptually, it assumes that the longest headlines are the most important ones.

4.3 Relevance Evaluation

We examine the performance of our approach for producing day summaries by comparing it with the aforementioned baselines. As discussed, we rely on human evaluation. We recruited 3 annotators, who confirmed to be familiar with the news stories used in our study, to annotate the relevance of the collected headlines in our dataset. We extracted all headlines of the news articles from 106 dates that appear in the ground truth TS (the expert timeline summaries of news agencies) and ignored dates on which fewer than 10 news articles were found. We asked the 3 annotators to label each headline as relevant ('1') or not ('0'), based on whether the headline reports events mentioned in the expert summary of that date. Among the annotators, one is a co-author of this study and the other two are graduate students. In total, 1319 headlines were annotated with an average agreement of $\kappa = 0.89$ between any two annotators. We kept only the dates that contain at least one relevant headline and kept the major judged answers among annotators. At the end, 1123 headlines were annotated for 47 dates.

Judging relevance for short summaries produced by the baseline systems can be a little more difficult than that of headlines. Therefore, to collect more judgments, we used CrowdFlower[8] for recruiting users to judge the relevance of the daily summaries produced by *ETS* and *Chieu et al.*. The users were requested to read the ground-truth summaries of a given date and to specify the relevance of sentences from the summaries from ETS or Chieu et al. Before working, each user was trained with at least 12 questions that we used as gold questions. During the job, they were secretly requested to answer gold questions. In total we collected 5104 judgments. Only answers from users who passed gold questions with a high agreement (≥ 0.85) were taken into account. We gathered between 5 and 10 trustful answers from separate trustful users for each question.

Results*:* Figure 1 shows the *Accuracy@2* of selected headlines (our system and MDS) and sentences (by Chieu et al. and ETS systems).

First, it can be seen that the results of the TS baselines ETS and Chieu et al. are not as good as those produced by our system and other headline-based baselines. The main

[8] http://www.crowdflower.com/

reason for this is that ETS and Chieu et al. select sentences from the full text of news articles and do not exploit the fact that the headlines themselves quite often serve as high-quality expert-created summaries of these articles. Their approach benefits from the rich distribution of the words, but - as a consequence - the huge list of sentences makes the task to create high-quality summaries more complicated.

Second, our system outperforms the MDS systems in selecting good headlines that reflect important events. This result implies that applying state-of-the-art MDS techniques does not ensure highly relevant events in the TS. This observation confirms the need for further investigation on TS generation using just headlines.

Third, *SumSim* perform slightly better than MDS. That is mostly because SumSim uses the information from neighbor articles (from the previous and next 10 dates) while MDS (and also Longest) do not. That is not a surprise, but confirms that the temporal aspect is crucial for TS generation, even for headline-based approaches. SumSim also outperforms its brother Chieu et al., and it shows the benefits of using headlines instead of article full-text here.

Fourth, our system outperforms all others with much higher scores for the generated timelines. The better performance can be explained by the following facts: (1) headlines are written by experts and mostly report the most important event; using the headline is therefore a better solution than selecting sentences from the full text. (2) different from the SumSim and TS baselines, our method leverages temporal information by using the influence aspect of headlines, which focuses on selective sentences with visible temporal tagging instead of all sentences; we observed that sentences with visible temporal tagging often highlight important information. (3) the combination of influence and the network structure (headline graph) produce better estimations of the importance, horizontally (Spread) and vertically (Influence). Last but not least, it is worth mentioning that the improvement is statistically significant.

4.4 Understandability Evaluation

With this experiment, we aim to evaluate the readability and understandability of the summaries from a user perspective. We compare our summaries one by one with the summaries produced by ETS and Chieu et al. More specifically, we investigate whether the selection of headlines produces summaries that are more coherent and comprehensible than extracted summaries that are composed from selected sentences from the article full-text.

Task setting: We provide CrowdFlower users with our collected ground-truth daily summaries from professional journalists, followed by 2 daily summaries, say A and B, which are alternately produced by either our system or ETS or Chieu et al. Users can answer “1” if A is more understandable than B, “-1” if A is less understandable than B, or “0” otherwise. The quality of answers is checked by the agreement with that of a small set of gold questions, secretly delivered to the users during their working sessions. In total, 141 summary pairs are presented to CrowdFlower users.

Result: We collected 2244 judgments from 122 users, of which 1552 judgments are from trusted users, who earned at least 0.85% correct on our gold questions and 0.85 *trust* gained from their work on CrowdFlower. Those 1552 judgments are used for our evaluation. The results are shown in Figure 2, where the value *m[Y][X]* in each matrix is the percentage of users who judged system *X* better than system *Y*. The higher the number, the darker its color. The rest *(1 - m[X][Y] - m[Y][X])*, which is not presented in the figure, is the percentage of users who considered X and Y to be equal.

Analysis: Generally, our headline-based approach results in better understandability than the other systems. We noticed that the confidence, the highest value among *m[Y][X], m[X][Y], and 1- (m[X][Y] + m [Y][X]),* is not very high, which indicates that the comparison of text quality is a hard task. User feedback confirmed that the task was clear (rated 4/5), but that they found it difficult to select the answer (rated 3/5).

While the relevance results showed that ETS is slightly better than the summaries provided by Chieu et al., users tend to rate the Chieu et al. summaries better in term of understandability than ETS. The reason could be that the ETS algorithm provides daily summaries that are related to summaries of the neighbor dates. Therefore, missing a piece of information from the connection between summaries between the neighbor dates can make ETS's day summaries less understandable than Chieu et al., which simply focuses on the daily events.

5 Conclusion

We presented a novel framework for automatically constructing a timeline summary for a news story from a collection of news articles. Different from previous work, where the proposed solutions extract sentences from article texts, our framework makes use of headlines. The intuition is that a careful selection of news headlines can result in summaries that are more informative and understandable than summaries that are composed of selected sentences from different parts of the news articles. Indeed, the qualitative user study showed that users prefer the timeline summaries produced by our headline-based approach over the summaries that are produced by other extractive approaches.

Unlike traditional MDS, our approach exploits temporal information to estimate the impact of an event on the future development of a new story. Therefore, it is worth mentioning that our approach best fits scenarios of retro-active summarization. Experimental evaluations have shown that the use of temporal information resulted in summaries of more relevant events than the ones selected by MDS methods.

Acknowledgements The first author thanks Dr. Katja Markert for her valuable comments. The work was partially funded by the European Commission for the FP7 project EUMSSI (611057) and the ERC Advanced Grant ALEXANDRIA (339233).

References

1. J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of SIGIR'01*, pages 10–18, 2001.
2. S. L. Althaus, A. J. Edy, and P. Phalen. Using substitutes for full-text news stories in content analysis: Which text is best? *American Journal of Political Science*, pages 707–723, 2001.
3. R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

4. G. Binh Tran, M. Alrifai, and D. Quoc Nguyen. Predicting relevant news events for timeline summaries. In *Proceedings of WWW companion 2013*.
5. S. R. K. Branavan, N. Kushman, T. Lei, and R. Barzilay. Learning high-level planning from text. In *The 50th Annual Meeting of the ACL*, pages 126–135, 2012.
6. H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of SIGIR'04*, pages 425–432, 2004.
7. D. Dor. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 2003.
8. G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, 2004.
9. S. Fernando and M. Stevenson. A semantic similarity approach to paraphrase detection, 2008.
10. L. Freund, J. Berzowska, J. Lee, K. Read, and H. Schiller. Digging into digg: Genres of online news. In *Proceedings of the iConference*, 2011.
11. T. H. Haveliwala. Topic-sensitive pagerank. In *WWW*, pages 517–526, 2002.
12. R. Kessler, X. Tannier, C. Hagège, V. Moriceau, and A. Bittar. Finding salient dates for building thematic timelines. In *Proceedings of ACL'12*.
13. S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. 1999.
14. C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NAACL'03 - Volume 1*, pages 71–78, 2003.
15. R. Malik, L. V. Subramaniam, and S. Kaushik. Automatically selecting answer templates to respond to customer emails. In *IJCAI 2007*.
16. K. McKeown, R. Barzilay, J. Chen, D. K. Elson, D. K. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman. Columbia's newsblaster: New features and future directions. In *HLT-NAACL*, 2003.
17. D. Metzler and T. Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *Proceedings of the 2008 ACM SIGIR LR4IR Workshop*, 2008.
18. R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, pages 775–780, 2006.
19. R. Mihalcea and P. Tarau. Textrank: Bringing order into text. In *EMNLP*, 2004.
20. C. A. Perfetti, S. Beverly, L. Bell, K. Rodgers, and R. Faux. Comprehending newspaper headlines. *Journal of Memory and Language*, 26(6):692 – 713, 1987.
21. D. R. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. elebi, S. Dimitrov, E. Drbek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. Mead - a platform for multidocument multilingual text summarization. In *Proceedings of LREC'04*, 2004.
22. D. R. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. pages 919–938, 2004.
23. J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the SemEval '10*, pages 321–324, 2010.
24. R. C. Swan and J. Allan. Timemine: visualizing automatically constructed timelines. In *SIGIR*, page 393, 2000.
25. G. B. Tran and M. Alrifai. Indexing and analyzing wikipedia's current events portal, the daily news summaries by the crowd. In *WWW (Companion Volume)*, pages 511–516, 2014.
26. C. J. van Rijsbergen. Information retrieval. 1979.
27. R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of SIGIR'11*, pages 745–754, 2011.
28. H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP '03*, pages 129–136, 2003.