Predicting Relevant News Events for Timeline Summaries

Giang Binh Tran, Mohammad Alrifai
L3S Research Center, Germany
{gtran, alrifai}@L3S.de

Dat Quoc Nguyen
UET, Vietnam National University, Hanoi
datnq@vnu.edu.vn

ABSTRACT

This paper presents a framework for automatically constructing timeline summaries from collections of web news articles. We also evaluate our solution against manually created timelines and in comparison with related work.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; H.3.1 [**Information Storage and Retrieval**]: Content Analysis

General Terms

Algorithms, Design

Keywords

News Events, Timeline, Summarization, Supervised

1. INTRODUCTION

Timeline summaries (TS) have become a common way for providing a simple and effective solution to explore and navigate through temporally related events.

The challenges in automatic generation of TS lie in extracting important points of the story, both in temporal and content dimensions. For instance, a good TS of the Arab Spring revolutions should capture events like "Egypt president Hosni Mubarak resigned on 11 February 2011.", "Muammar Gaddafi was killed on 20 October 2011.", etc.

While there has been a rich body of research in automatic summarization, little work recently tackles the problem of timeine summarization [7, 2, 1]. In this paper, we propose a supervised approach for automatically generating TS of web news articles.

Our work employs human created TS from news agencies, in order to better meet perception of a good TS. Furthermore, our work differs from previous studies in providing a unified framework to summarize the information both in time and content dimensions.

2. TIMELINE SUMMARY GENERATION

Typically, a TS consists of a chronologically ordered list of day summaries. Given a collection A_q of news articles related to a specific news topic q, the goal is to produce TS with a maximum number of n day summaries, where each day summary has a maximum number of m sentences. The value of n and m is defined by the desired compression ratio. We use machine learning approach to predict the importance score of each date expression and each sentence extracted from A_q. Next, we return the top m sentences for each of the top n dates.

2.1 Dataset and training setting

We collected several TS published by popular news agencies such as CNN, BBC, NBCnews, etc. about famous topics such as "BP Oil Spill", "Influenza H1N1" and "'Arab Spring"'. We only consider timelines where the timestamps are explicit dates (including day, month and year) such as *07 July 2011*. Finally, we obtained 17 different TS from 9 different topics.

For each TS, we used Google to retrieve news articles from the news agency that published the timeline (i.e. BBC news articles for BBC-published timeline) using topics as queries and time filter option and retained top 400 returned articles that are published during the TS timespan. At the end, we obtained 4650 news articles in total after duplication removal. Next, we used some cleaning tools and additional hand-crafted rules to extract the main content of the news articles.

We use *leave one out* approach based on topics to split our corpus into training and testing sets. At each round, we train a LinearRegression model with TS of 8 topics and test on TS of 1 topic left. In this experiment, we used the LinearRegression implementation provided by Weka toolkit [3].

2.2 Feature Selection for Dates

The follwoing features are extracted for each date d:
articles published on d.
articles published after d and have reference to d.
articles published before d and have reference to d.
sentences published on d and refer to d.
sentences published after d and refer to d.
sentences published before d and refer to d.

The intuition here is that the relevance of a date d is determined by the number of refernces to d.

During the training phase, we assign each date d a target score of 1 if d is in the manually created timeline and 0 otherwise.

2.3 Feature Selection for Sentences

Table 1 shows the summary list of 22 features we extract

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

for each sentence $s \in S_{d_i}$. We group them into 4 different categories as following:

Table 1: List of features and their category

Feature	Category
Length	surface
stop/non-stop words	surface
#pronoun	surface
position	surface
#causal signals	coherence
#temporal signals/expression	coherence
#logical signals	coherence
sum/avg TFIDF/pos * TFIDF	topic
cross entropy (sentence v.s day news)	topic
TF top 10,30,50,100	topic
sum/top/avg logodd	topic
popularity	time-related
hasTempExp	time-related

In manually created timelines sentences are often paraphrased and differ the original sentence from news article. We compute the target score of each sentence s in the articles published on date d by measuring its semantic similarity to the manually created summary DS_d of that date.

$$target(s) = \max_{k=1}^{|DS_d|} sim(s, s_k) \quad (1)$$

where $sim(s, s_k)$ measures the content similarity between two sentences s and s_k. We adopt sentence similarity measurement method proposed by [5].

$$sim(s, s_k) = \frac{\sum_t w_s(t).w_{s_k}(t)}{\sqrt{w_s^2(t)w_{s_k}^2(t)}} \quad (2)$$

where t is a term occurring in both s and s_k; $w_s(t)$ is sentence-based TF*IDF weight of term t.

3. EVALUATION

We evaluate our solution by measuring the similarity of the generated TS with the manually created ones and in comparison with other methods.

3.1 Date selection

We use *Mean Average Precision* (MAP) metric for comparison. Let the set of relevant dates for TS be $d_1, ..., d_n$, and R_k be the ranked list of dates from d_q to d_k

$$MAP = \frac{1}{n} \sum_{k=1}^{n} Precision(R_k) \quad (3)$$

We compare our solution with a baseline system that ranks the dates by number of news articles published on that date. The average MAP result of 17 timelines is shown in the table 2, indicating that our method outperforms the baseline.

Table 2: Comparison of date selection methods

Method	MAP
$\#Docs_{published}$	0.47
LinearRegression	**0.54**

3.2 Sentence selection

We use the ROUGE scores [4], including ROUGE-N-1 (R1), ROUGE-N-2 (R2) and ROUGE-S* (S*), as the metric for evaluation (using ROUGE toolkit (1.5.5) [4]). Let's take the $R1$ score of the timeline is computed as follows:

$$R1 = \frac{2 * R1 - P * R1 - R}{R1 - P + R1 - R} \quad (4)$$

where, $R1 - P$ and $R1 - R$ are the average R1-Precision and R1-Recall scores over all day summaries. Then, R2 and S* are computed in a similar manner.

We evaluated our system against the work of Chieu et al. [2], MEAD [6], and ETS [7].

The results are reported in Table 3.

Table 3: Comparison of summary generation methods

Method	R1	R2	S*
Random	0.128	0.021	0.026
Chieu	0.202	0.037	0.041
MEAD	*0.208*	*0.049*	0.039
ETS	0.207	0.047	*0.042*
LinearRegression	**0.218**	**0.050**	**0.046**

4. CONCLUSION

In this work we present a supervised machine learning method for automatic summarization of news event timelines. Our solution exploits timelines created manually by professional journalists to train a Linear Regression model for selecting relevant time points and content to be included in the timeline summary.

5. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of SIGIR'01*, pages 10–18, 2001.

[2] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of SIGIR'04*, pages 425–432, 2004.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[4] 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.

[5] R. Nelken and S. M. Shieber. Towards robust context-sensitive sentence alignment for monolingual corpora. In *In 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[6] D. R. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. ÃĞelebi, S. Dimitrov, E. DrÃąbek, 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.

[7] 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.