

Towards Event Timeline Generation from Vietnamese News

Van-Chung Vu¹, Thi-Thanh Ha^{1,2}, and Kiem-Hieu Nguyen¹

¹ School of Information and Communication Technology,
Hanoi University of Science and Technology,

² Thai Nguyen University of Information and Communication Technology
vanchung1995@gmail.com, htthanh@ictu.edu.vn, hieunk@soict.hust.edu.vn

Abstract. Event timeline generation is an active research area in many languages. In this paper, we describe our attempts towards event timeline generation from Vietnamese newswire documents based on the work in [1]. Our main contributions are: *i)* To our knowledge, we are the first to tackle the problem for Vietnamese language. *ii)* Experiments were conducted on a large-scale corpus from 145 popular online news agents in the period from 2007 to 2017 which provides extensive redundant information on various themes. *iii)* We manually built a dataset of 17 timelines for evaluation³. Experimental results show that the proposed method works reasonably well for Vietnamese; and using n-grams on unsegmented texts achieves comparable performance to using word segmentation.

1 Introduction

Text summarization is an active research area in NLP [2]. Event timeline generation could be considered as a branch of text summarization which focuses on summarizing thematic events [1]. It is closely related to query-focused multidocument summarization [3] and topic detection and tracking [4]. The differences are two-fold: Firstly, it works on event-centric documents; Secondly, it considers not only textual information but also temporal information when summarizing. It eventually distinguishes from timeline generation for an individual entity [5].

By *thematic event*, we are referring to a collection of events relevant to a common theme. For example, the thematic event “Vụ con ruồi trong chai Number 1 (The fly in Number 1)”⁴ started with an event “A customer found a fly inside an unopened Number 1 bottle” and continued with another event such as “He then contacted Tan Hiep Phat company to blackmail 1 billion VND”, etc. Our task is to generate a timeline of the most salient events relevant to a thematic

³ The dataset could be downloaded at <http://is.hust.edu.vn/~hieunk/resources/vntimeline17.zip>

⁴ For convenience of readers who are not familiar with Vietnamese, we are going to use the English translation instead of the original texts in Vietnamese in the rest of the paper.

event as in Figure 1. Following previous works on event timeline generation, we treat an event as a pair of timestamp and an event description. For instance, (2014/12/03, “A customer found a fly inside an unopened Number 1 bottle”) is an event. We are not going to use event details such as triggers and participants as in Message Understanding Conference and Automatic Content Extraction (and its subsequent Knowledge Base Population) [6,7,8].

When reporting on a thematic event, such as a politic scandal or a disaster, different journalists tend to agree on salient events and story developments. Such agreement is consequently reflected on redundancy between news sources. We aim at leveraging this redundancy to detect salient events.

The task has been studied in other languages like English and French [9]. However, to our knowledge, there has been no study in Vietnamese so far although there is a large and rapidly increasing volume of event-related newswire contents in Vietnamese.

In this paper, we present our approach towards generating event timelines on Vietnamese newswire documents. Our contributions are three-fold: Firstly, to our knowledge, we are the first to tackle this problem in Vietnamese. Secondly, as redundancy from various sources is crucial in order to judge importance of information related to an event, we have gathered a large-scale corpus of Vietnamese news from 145 popular online news agencies in Vietnam in the period from 2007 to 2017. As illustrated by experimental results, such corpus is adequate for generating timelines of various events. Last but not least, we have manually created a dataset of 17 timelines for evaluation. The dataset consists of various events happening in Vietnam as well as world-wide. Experimental results on the dataset show that the models based on n-grams are on par with the word-based model.

The rest of the paper is organized as follows: Section 2 briefly introduces related work; Section 3 describes our proposed method; Section 4 demonstrates evaluation results; The paper is concluded in Section 5.

2 Related Work

Most works on event timeline generation require redundancy from data. [10] proposed a framework for extensive temporal analysis and used redundant data and machine learning to detect salient dates of thematic events. Following this direction, [1] presented a Maximal Marginal Relevance-like reranking algorithm based on both temporal and thematic clustering [11]. In another approach, [12] proposed a joint graphical model for the problem. The problem has been also tackled in other languages such as French [9]. In an attempt to improve evaluation methodology, [13] proposed a metric based on so-called deep semantic units.

In this work, we aim at applying the approach as described in [1] to a new language, i.e. Vietnamese in our case. Therefore, we follow the essential parts of the method including document acquisition, temporal analysis, event ranking and selection.

Vụ con ruồi trong chai nước Number 1

- 2014/12/03: Ông Võ Văn Minh (35 tuổi, Tiền Giang) lấy chai number one bán cho khách hàng và phát hiện con ruồi trong chai chưa khai nắp.
- 2014/12/05: Ông Minh gọi điện cho công ty TNHH thương mại dịch vụ để nghị công ty đưa 1 tỉ đồng nếu không đưa sẽ kiện ra hội đồng bảo vệ người tiêu dùng.
- 2014/12/06: Nhân viên Tân Hiệp Phát đến lần 1 không đưa tiền mà chỉ tặng sản phẩm và xin lại chai nước.
- 2014/12/16: Nhân viên Tân Hiệp Phát đến lần 2, lập biên bản ghi nhận. Ông Minh đề nghị đưa 500 triệu
- 2015/01/20: Nhân viên Tân Hiệp Phát đến lần 3. Ông Minh đề nghị 500 triệu
- 2015/01/27: Nhân viên Tân Hiệp Phát giao 500 triệu và ông Minh đưa chai nước, nhận tiền và bị công an bắt. Ông Minh bị khởi tố cưỡng đoạt tài sản.
- 2015/02/13: Thanh tra sở y tế tỉnh Bình Dương thanh tra đây truyền của Tân Hiệp Phát.
- 2015/02/14: Kết luận Thanh tra không vi phạm.
- 2015/05/27: Công an gia hạn tạm giam ông Minh, Kết quả giám định nắp chai không còn nguyên.
- 2015/09/08: Công an kết luận ông Minh phạm tội cưỡng đoạt tài sản và đề nghị truy tố.
- 2015/10/09: Ra cáo trạng và chuyển hồ sơ sang tòa.
- 2015/12/17: Xét xử sơ thẩm, Minh bị lĩnh 7 năm tù
- 2015/12/18: Tòa án ND tỉnh Tiền Giang tuyên án 7 năm tù với Nguyễn Văn Minh
- 2015/12/19: Phó Tổng giám đốc xin lỗi người tiêu dùng
- 2016/01/21: Người dân phát hiện ruồi trong chai Number 1 và được đưa ra hội bảo vệ người tiêu dùng.
- 2016/12/23: Tổng giám đốc xin lỗi người tiêu dùng

Fig. 1. A timeline of “The fly in Number 1” written by journalists.

Applying to a resource-scarce language like Vietnamese is not trivial. The main issues are:

1. Acquiring a large, redundant news corpus over a long period.
2. Processing temporal analysis.
3. Dealing with unsegmented texts, i.e. texts in which word delimiters are ambiguous. This is a specific problems in some Asian languages like Chinese, Japanese, and Vietnamese.
4. Creating reference timelines for evaluation.

We will discuss in more details these issues and our proposed solutions in subsequent sections.

3 Our Event Timeline Generation Method

In the first phase, from a large, redundant news corpus, temporal analysis is processed on the whole corpus. That is to say, whenever there is a temporal expression, it will be detected and will be normalized. Sentences containing at least one temporal expression are then collected. We use Lucene⁵ search engine to index events as pairs of date and description sentences. Moreover, sentences containing the same normalized time will be gathered into a cluster. We use Lucene ranking algorithm to rank and select salient events for an input query and to generate the final timeline. In this paper, we follow previous works to use date as temporal interval. Each date cluster hence is consists of all sentences

⁵ <https://lucene.apache.org>

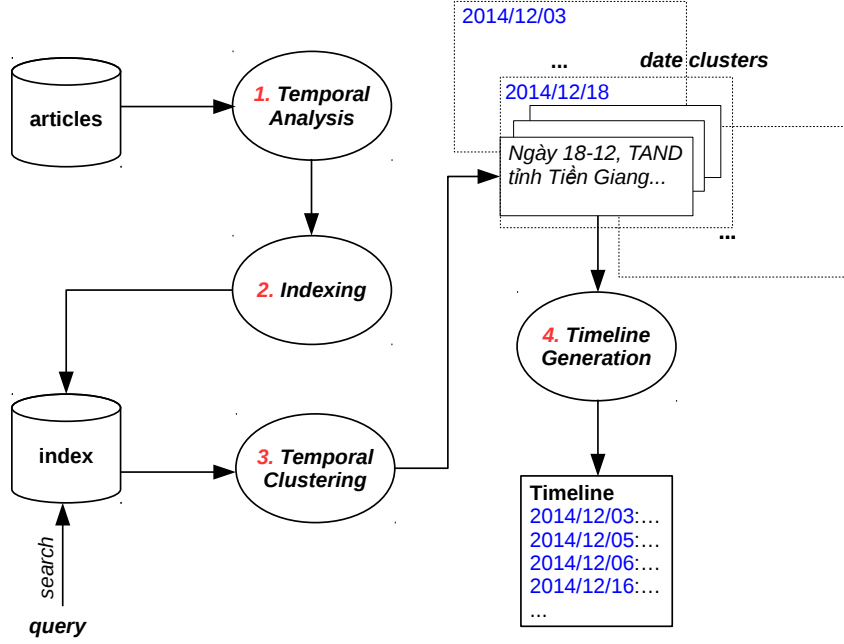


Fig. 2. Our event timeline generation method

belong to an individual date. Our method is demonstrated in Figure 2. Readers could refer to [1] for more details on the original method.

In Section 3.1, we describe our news corpus. Sections 3.2 and 3.3 present temporal analysis and indexing, respectively. Section 3.4 is dedicated to temporal clustering. Timeline generation is discussed in Section 3.5.

3.1 The News Corpus

A large, redundant corpus is crucial in this work. To obtain such a corpus in Vietnamese, we first surveyed popular online news agencies in Vietnamese. From that, we selected 145 most popular sites. Articles from these sites between the year of 2007 and 2017 were gathered, resulting in totally 3.8M articles. HTML documents were parsed and all contents and meta-data such as title, document creation time (DCT), tags, URL were stored in XML format. Note that DCT is required for temporal analysis. Most of the articles comes from dominant agencies in Vietnam such as VnExpress, Tuoitre, Dantri, Vietnamnet. The content varies from social-economics, politics to hi-tech, entertainments, world wide events.

3.2 Phase 1: Temporal analysis

Temporal analysis consists of detecting and normalizing temporal expressions in texts. It is shown in [10] that only 7% of temporal expressions in texts are absolute dates (i.e. with date, month, and year), the rest DCT-related dates require normalization. For example, in the sentence “Ngày 18-12, TAND tỉnh Tiền Giang tuyên Minh 7 năm tù vì tội Cường đoạt tài sản”, we need to detect “Ngày 18-12” and normalize it as 2004/12/18. Other expressions such as “ngày hôm qua” (yesterday) or “Thứ Sáu tuần trước” (last Friday) are even more challenging.

In our experiments, we used HeidelTime, a multilingual temporal analyzer which supports Vietnamese [14]. To our knowledge, it is the only temporal analyzer to date for Vietnamese. It uses JvTextPro⁶ for word segmentation and part-of-speech tagging as prerequisite for temporal analysis.

After temporal analysis, we remove all sentences without temporal expressions as well as sentences with normalized temporal expressions that are incomplete (e.g. 2015/10/xx where date is missing). These sentences are indexed using Lucene search engine, each one as a *document*, resulting in an index of totally 11.6M documents. Statistics of the corpus is shown in Table 1.

Table 1. Corpus statistics.

Item	Number
Sites	145
Articles	3,801,523
Sentences	54,932,191
Sentences with temporal expressions	11,596,796

3.3 Phase 2: Indexing

Vietnamese texts don’t have explicit word boundaries. Spaces, which are natural word boundaries in languages such as English, only serve as boundary between syllables in Vietnamese. Word segmentation is required for detecting word boundaries, i.e. deciding whether or not a space is a word boundary or is inside a multi-syllable word. Word segmentation is useful for downstream problems like syntactic parsing and semantic parsing. It has been shown in [15] that using n-grams is comparable to using words for information retrieval of Chinese texts.

To investigate the impact of word segmentation, in our experiments, we built three indices using unigram, bigram, and word. For example, “truy hồi (retrieval) thông tin (information)” is indexed as {‘truy’, ‘hồi’, ‘thông’, ‘tin’}, {‘truy hồi’, ‘hồi thông’, ‘thông tin’}, and {‘truy hồi’, ‘thông tin’} using unigram, bigram, and word, respectively. VnTokenizer [16] was used for word segmentation.

⁶ <http://jvntextpro.sourceforge.net/>

3.4 Phase 3: Temporal clustering

When a query, such as “The fly in Number 1”, is fed into the Lucene index, it will return relevant documents to the query. In our experiments, we used the built-in tf-idf scoring function of Lucene, and limited to 10K documents for each query. All events having the same date value are gathered into a *date cluster*. Date cluster is a central notion in our method. If a thematic event lasts over a long period, the dates on which many events happen tend to be more salient than the others. Moreover, within a date cluster, not all the events are equivalently important. Salient events take an essential part in the whole story. Marginal events, for instance, could be some kind of reactions, or could describe minor details. The most important events are duplicated in several descriptions because they are reported by many news sources.

3.5 Phase 4: Timeline generation

Timeline generation consists of selecting the most salient dates and selecting the most salient event in each date. Saliency score of a date d is the accumulation of Lucene scores⁷ of all events e in d regarding a query q :

$$saliency(d) = \sum_{e \in d} score_{Lucene}(e, q) \quad (1)$$

Lucene score of an event reflects the relevance of its description to the query.

For event ranking inside a date d , we simply select the one with the highest Lucene score as the representative event of d :

$$\arg \max_{e \in d} score_{Lucene}(e, q) \quad (2)$$

The resulting timeline has K events and could be ordered by saliency or chronology. The number of events K could be varied to show only salient events or to get more details.

4 Evaluations

In this section, we describe the creation of a dataset of 17 reference timelines in Section 4.1. Our evaluation follows the work in [1]. Temporal contents of the timeline are evaluated by salient date detection (Section 4.2). Textual contents are evaluated by text summarization (Section 4.3).

4.1 Creation of reference timelines

We first investigated timelines written in Vietnamese by journalists. There were not many such timelines at the moment we conducted the experiments. Timelines

⁷ https://lucene.apache.org/core/3_6_0/scoring.html

describing events out of the period 2007 and 2017, which are not available in our corpus, are ignored. Our final dataset contains 17 timelines, for both events happening in Vietnam (e.g. “Airplane crashing in Hoa Lac”) and worldwide (e.g. “Missing plane MH370”) as shown in Table 2.

4.2 Evaluating salient date detection

Table 2. Evaluating salient date detection.

No Query	UNIGRAM	BIGRAM	WORD
1 Nguyễn Thanh Chấn (Nguyen Thanh Chan)	26.4	26.1	24.4
2 Cá chết hàng loạt ở miền Trung (Dead fish spreading in Middle Vietnam)	41.9	41.3	43.5
3 Cà phê “Xin Chào” (“Hello” café)	57.9	60.2	54.9
4 Cao Toàn Mỹ - Trương Hồ Phương Nga (Cao Toan My - Truong Ho Phuong Nga)	13.1	16.7	11.7
5 Cháy karaoke Cầu Giấy (Cau Giay karaoke on fire)	44.7	47.9	43.8
6 Leo thang căng thẳng trên bán đảo Triều Tiên (Escalating tensions in Korea Peninsula)	10.7	14.8	12.4
7 Máy bay mất tích MH370 (Missing plane MH370)	34.4	34.1	38.2
8 Philippin - Trung Quốc (Philippines - China)	60.6	49.1	54.9
9 Máy bay rơi Hòa Lạc (Airplane crashing in Hoa Lac)	3.1	2.0	1.3
10 Thẩm mỹ viện Cát Tường (Cat Tuong beauty saloon)	48.0	48.4	43.7
11 Giàn khoan 981 (HD-981 Oil Rig)	70.6	74.2	73.8
12 Bê bối của Tổng Thống Hàn Quốc (South Korean president scandal)	48.7	49.2	51.1
13 Vụ con ruồi trong chai Number 1 (The fly in Number 1)	55.6	52.3	54.9
14 Đắm phà Sewol (Sewol sinking ferry)	48.6	45.1	47.9
15 Đồng Tâm, Mỹ Đức (Dong Tam, My Duc)	11.0	13.8	7.2
16 Huỳnh Thị Huyền Như (Huynh Thi Huyen Nhu)	30.2	34.2	18.8
17 Trịnh Xuân Thanh (Trinh Xuan Thanh)	41.3	54.5	53.1
MAP	38.1	39.0	37.4

We used Mean Average Precision (MAP) to evaluate salient date detection on three systems, each uses one of the three indices: unigram, bigram, and word. For each query, the ranked list of all the dates returned by a system according to Equation (2) is compared against the dates in the reference timeline. If a date in the reference timeline is not retrieved by the system, its average precision is counted as zero.

$$MAP = \frac{1}{N} \sum_{j=1}^N \left(\frac{1}{Q_j} \sum_{i=1}^{Q_j} P(date_i) \right) \quad (3)$$

Q_j : number of relevant dates for query j

N : number of queries

$P(date_i)$: precision at i th relevant date

As shown in Table 2, UNIGRAM and BIGRAM perform slightly better than WORD. Performance varies across the timelines. The main reasons for poor performance are: An event lasts for too long (“Cao Toan My - Truong Ho Phuong Nga” lasts for three years); The event lasts only in a few days and there are many events of the same type happening in the same time but in other locations (“Airplane crashing in Hoa Lac”); There are not many news about the event, such as the riot in Dong Tam, My Duc; The event lasts for too long but the reference timeline only covers a specific period (The timeline about “Escalating tensions in Korea Peninsula” only covers events in 2017). In fact, one interesting aspect of our method is that when a user want to focus on a particular period, he could limit search range accordingly. For example, one could zoom in for events about Korea Peninsula in 2017. Implementing this feature in Lucene is straightforward.

4.3 Evaluating text summarization

For each query, top K dates from the system are selected. The most relevant sentence in each date is then extracted. The final timeline contains K sentences. Here, K is the number of dates in the reference timeline. We use ROUGE metric [17] to evaluate generated timelines against reference timelines.

Table 3. Evaluating text summarization.

System	ROUGE-1	ROUGE-2	ROUGE-L
UNIGRAM	39.0	16.4	26.3
BIGRAM	38.3	16.1	25.5
WORD	40.1	15.3	24.9

Table 3 again demonstrates that UNIGRAM and BIGRAM perform equally to WORD, while requiring no word segmentation. On the other hands, the results are fluctuated. This is probably because we only have 17 queries and there is only one timeline per query, that could reflect subjectivity in timeline contents.

5 Conclusions and Future Works

This paper presents a timeline generation system for Vietnamese. The experiments were conducted on a large newswire corpus of articles in various domains. We empirically show that using unigrams and bigrams produces timelines of comparable quality without word segmentation.

There are more rooms for improvement. We could applying event clustering to highlight important events and event reranking for diversity as in [1]. Unlike their work, we don't have keywords in reference timeline. Therefore, putting only event *titles*, sometimes too short or too ambiguous, into Lucene could return in many irrelevant results. Query expansion technique could be a useful and feasible solution. Moreover, we are going to expand the reference dataset that has multiple timelines per query for more robust evaluation. One possibility is using existed dates from timelines written in English about world-wide events for evaluating salient date detection, and further manually translating those timelines into Vietnamese for evaluating text summarization. Another direction for enhancement is improving temporal analysis for Vietnamese. TIME named entities are not only crucial for timeline generation but they are also important in other tasks such as event extraction and knowledge base population.

Acknowledgments

We would like to thank Vccorp for kindly supporting us on conducting the experiments.

References

1. Nguyen, K.H., Tannier, X., Moriceau, V.: Ranking multidocument event descriptions for building thematic timelines. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, Dublin City University and Association for Computational Linguistics (2014) 1208–1217
2. Mani, I.: Advances in Automatic Text Summarization. MIT Press, Cambridge, MA, USA (1999)
3. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4. NAACL-ANLP-AutoSum '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 40–48
4. Allan, J.: Topic detection and tracking. Kluwer Academic Publishers, Norwell, MA, USA (2002) 1–16
5. Li, J., Cardie, C.: Timeline generation: Tracking individuals on twitter. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14, New York, NY, USA, ACM (2014) 643–652
6. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 1. COLING '96, Stroudsburg, PA, USA, Association for Computational Linguistics (1996) 466–471

7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ace) program tasks, data, and evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal, European Language Resources Association (ELRA) (2004) ACL Anthology Identifier: L04-1011.
8. Ji, H., Grishman, R.: Knowledge base population: Successful approaches and challenges. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1148–1158
9. Battistelli, D., Charnois, T., Minel, J.L., Teissèdre, C.: Detecting salient events in large corpora by a combination of NLP and data mining techniques. In: Conference on Intelligent Text Processing and Computational Linguistics. Volume 17., Samos, Greece (2013) 229–237
10. Kessler, R., Tannier, X., Hagège, C., Moriceau, V., Bittar, A.: Finding salient dates for building thematic timelines. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jeju Island, Korea, Association for Computational Linguistics (2012) 730–739
11. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for re-ordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98, New York, NY, USA, ACM (1998) 335–336
12. Tran, G., Herder, E., Markert, K.: Joint graphical models for date selection in timeline summarization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics (2015) 1598–1607
13. Bauer, S., Teufel, S.: A methodology for evaluating timeline generation algorithms based on deep semantic units. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, Association for Computational Linguistics (2015) 834–839
14. Strötgen, J., Gertz, M.: Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 321–324
15. Nie, J.Y., Gao, J., Zhang, J., Zhou, M.: On the use of words and n-grams for chinese information retrieval. In: Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages. IRAL '00, New York, NY, USA, ACM (2000) 141–148
16. Phuon, L.H., Thi Minh Huyền, N., Roussanaly, A., Vinh, H.T.: Language and automata theory and applications. Springer-Verlag, Berlin, Heidelberg (2008) 240–249
17. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S.S., ed.: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, Association for Computational Linguistics (2004) 74–81