Structured Summarization for News Events

Giang Binh Tran
supervised by Prof. Wolfgang Nejdl
and Dr. Mohammad Alrifai
L3S Research Center & University of Hannover
Appelstrasse 9a, 30167 Hannover, Germany
gtran@l3s.de

ABSTRACT

Helping users to understand the news is an acute problem nowadays as the users are struggling to keep up with tremendous amount of information published every day in the Internet. In this research, we focus on modelling the content of news events by their semantic relations with other events, and generating structured summarization.

Categories and Subject Descriptors

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

General Terms

Data Mining, Algorithms, Summarization, Cognition

Keywords

News Events, Relation Extraction, Structured Summarization, Temporal, Causal, Spatial, Hierarchical

1. INTRODUCTION

"What is the Arab Spring? Where and when it start occurring? Why did Mubarak resign? How is the Egypt revolution related to the Arab Spring?" These are examples of typical questions that are likely to be asked by people when seeking for information about Arab Spring or similar news topics on the Web. Guiding the users to the answers will help them gain better understanding of the topics. Unfortunately, the immense number of articles talking about these topics bring the users difficulty in grasping all necessary information, and consequently, they often miss the big picture about the events they are interested in. While traditional search engines can effectively deliver a load of news articles related to an event, they fail in removing duplicates, fusing news content and providing the users important aspects of the event. Hence, it calls for engaging tremendous flood of information of the news events and structure them from a high view of semantics.

Why summarization? Summarization is one of the most effective ways to mingle information to avoid information overload. For example, (multi-)document summarization has a strong impact on many applications like headline or snippet generation. It is also valuable for many real life tasks, such as preserving information for future generation.

Why structured summarization? Structured summarization provides better information for the users than the traditional one. For instance, due to the chronological characteristic of online news, temporal structured (timeline) summarization has become a natural way to present a story. A timeline summarization highlights the most important events during the development of the story over time, such as, during the active period of the Arab Spring revolution, the timeline can capture events like "Egypt president Hosni Mubarak resigned after several protests on 11 Feb 2012.", "On 25 Feb, Libyan opposition claimed to control the power over Muammar Gaddafi.", etc. However, there has not been many studies on structured summarization except timeline summarization, which leverages temporal aspect to advantage traditional summarization.

In this research proposal, we aim at structured summarization for news topics in order to help the users answer *"What, When, Why, Where, How"* questions. Extending from traditional summarization, we consider various relational types such as causal, temporal, spatial and topic hierarchical relations since they are intuitively close to typical *"Wh-questions"* that people likely to ask when they are looking for information about the news topics. Moreover, these types of relation among events are shown to be key factors contributing to the content comprehension [20, 13, 3], which is an increasing demand on newswire systems. We intend to blend them together in a structured summarization framework. Our proposed methodology includes the techniques from content analysis and cognitive science to learn the event structure in a close way to the human cognition. Our ultimate goal is to support users to have an excellent navigation to travel through news articles and understand the whole picture of an event they care about.

Our expected contributions are:

- Improve state of the art in event relation extraction across documents.
- Novel algorithms for news summarization from different semantic aspects, such as temporal, causal, spatial and topic hierarchical relation between events.

2. RELATED WORK

This section reviews the related work to our research. There are a substantial number of studies close to what we are proposing to do such as: event relation extraction, hierarchical classification, text summarization.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

Event relation extraction In literature, event relation extraction task has received considerable attention from research community and many of them follow from discourse theory [14]. To name some,[4, 10, 7, 11] on causal relation, [17] on spatial relation, [23, 24, 5] on temporal relation.

Unlike major spatial, temporal and causal studies in linguistics , our research focuses on high conceptual events (i.e, event is something happens, e.g., Arab Spring, Iphone 5 release) instead of lower lexicon-based events (i.e, linguistically when the verb appears in the sentences or phrases). Since their focus was on inter-sentence and inter-document event relations, we plan to extract the relations at the multi-document (or intra-document) level.

Document hierarchical classification News document hierarchical classification has been extensively studied for long time. The common approaches classify texts into a hierarchical labelled category predefined by human or inferred from some knowledge ontology (for example, see [12]). Typically, machine learning algorithms such as SVM, NaiveBayes are widely used. This research is a hint for us to extract the hierarchical relation between news events. We are not planning to improve the state of the arts in this problem but employ them for our extraction task.

Learning topic relation between events The task of learning topical relation between events is basically related to topic detection and tracking (TDT) or event evolution (for TDT example, see [26, 15]). However, TDT draws a bead on the natural change of an event, thus, it differs to what we propose in that our task is to determine whether two events hold a same content topic and indicates some content transition between them. This problem is related to topical shift catching which is useful in information ordering task [3].

Text summarization has been extensively investigated since the last decade (e.g., [8, 21]). There are various of methods to sort out this problem. However, we plan to provide more structured representation which can capture important semantic aspects of the article collection.

Structured summarization There is a large body of research in generating structured representation for news event as timeline summarization. In most cases, the timeline generation system aims at extracting *important* events or sentences and put them on a chronological time span (e.g., see [6, 1, 22, 25]) There are several approaches for important sentences extraction. For example, in sentence level, Chieu et al [6] used "interest" and "burstiness" scores, which intuitively indicate the popularity of the reported event in the sentence and the time point; Allan et al [1] relied on "Usefulness", "Novelty" of the sentences, which show the how a sentence is related to the news topic and how it differs to prior sentences. This kind of summarization system can capture the important sentences from news articles and their temporal aspect but works only for simple stories which are linear in nature as claimed in [18]. In a nutshell, it lacks many other aspects like cause-effect, spatial relation, intention and thus reduces the semantic relations between events. In our research, we don't ignore the temporal aspect because it is a necessary for users but we aim at provide new algorithms to improve state of the art and to ease the process of mingling with other aspects.

3. PROPOSED RESEARCH

This section describes our proposed research. We will introduce our research questions and then discuss an approach to manage them.

3.1 Research Questions

Although event analysis has been long time studied and the concept of event is intuitively clear but there has not been an formal definition of event, up to our knowledge. In our research, we adopt the common used definition of event provided by Allan et al [2] below:

Definition 1: Event is something that happens at a particular time and place.

Starting from our motivation in section 1, our main research question clearly addresses on: how to improve the user's experience while navigating news articles related to an event using event content modelling based on relation analysis? Our research will focus on some following aspects:

(1) Extract event relation across documents: The relation between 2 events may range from *sub-parent* to *temporal*, *cause-effect*, *spatial* and *topic hierarchical* relations. Although in literature there exists some studies that extract (typically one of) these relations, the task of event relation extraction across documents, such as temporal or causal relation, is still open and challenging. Additionally, there are not many published studies (if any) working on determining how each type of event relationship affect each others. Answers for these two problems can benefit not only our research topic but also related research areas.

(2) Generate structured summarization for users: There are many questions regarding the way we generate the structured summarization of an event to the users: How to incorporate all types of relations above into one framework? How to meaningfully present the events to the users? Literature has shown both traditional summarization or the common list-like structure (i.e, timelines) have some limitations. It becomes challenging when we have different relations in hand that requires some blends. Hence, we seek for better solutions so that users can navigate easily the set of news articles and can quickly grasp the important information of the articles from different aspects.

(3) Leverage wisdom of the crowd: To remind, we attempt to build a representation which is close to human's cognition because the closer to human's cognition it is, the better users comprehend its content. While *wisdom of the crowd* is a very much valuable resource, how can we employ that knowledge to improve our system? How to borrow knowledge from social computing resource like Wikipedia (where people often create some sorts of representation for main events, e.g, summarization or timelines of Arab Spring, The Pirate Bay trial)? Could we build a system that can exploit human inference skills to improve our relation extraction and summarization task?

3.2 Research Approach

Referring our research questions in the Section 3.1, we propose a solution to tackle our research problem as the following discussion.

3.2.1 Extract Event Relation across Documents

Our current plan is to extract relationships between events namely causal, temporal, spatial, topic hierarchical relation. To do so, first we follow previous work on event detection and tracking to annotate all possible events in the document collection. We extract the characteristics of an event such as protagonists, predicate, time, location with some semantic role labeling and temporal tagger tools. In this way, we can have characteristics of the event based on its lexicon appearance. The more difficult issue is extracting these properties of the higher conceptual events (for example, "Iphone 5 release") because most of such existing systems fail at this case. We suggest using semantic analysis to find the similar lexicon-based events for the higher conceptual ones (e.g, Iphone 5 release v.s. Apple will introduce the new phone this September) to deal with this case. We believe it is a challenging task but possible to tackle.

For each event relationship, we plan to create heuristic rules to capture explicit relations and then use machine learning or probabilistic methods to infer implicit relations. We consider the effects of each relation to each other to foster our model.

3.2.2 Generate Structured Summarization

This task will cope with many outstanding problems: first, we need to determine which information is important and should be included in the summary. For example, when taking temporal aspect into account, it requires us to know which date is important. It is challenging in the sense that for hot event such as the Arab Spring, it is likely that very frequently there are some news agencies publishing a articles about it, therefore, there will be a lot of dates appearing on the news article collection. However, we only able to put some of them in the summary. Similarly, determining the important event of a date is also not an easy task.

Second, we need to ensure the text quality of the summaries since gathering different events/information from with different aspects may reduce the coherence of our summarization, hence, it won't help users better understand the content we want to reveal. It is a grueling task in text generation and requires a deep analysis in (psycho) cognitive science or theoretical computational linguistics.

3.2.3 Investigate the Use of Social Computing

We investigate the use of knowledge from social computing to enhance our task of event relation extraction and summarization. First, we aim at leveraging some knowledge resources like Wikipedia or Freebase to improve our information extraction and summarization algorithms. Later, we plan to borrow knowledge from the crowd. Recent trends are delivering tasks to workers on some crowdsourcing platforms such as MTurks or Crowdflowers for later uses. However, in this direction, it is likely a "task-in task-out" (or "oursourcing") approach, e.g, if one person wants to tag an objects in an image from a huge collection of images, he distributes the task to many workers and gather them together. We aim at exploring the use of social computing in another way around, that is, our system should be able to learn some "added value" in the sense that analysing and merging the crowd's input together for discovering something new rather than "task-in task-out". Our idea is to build a Web service to help the Web users complete their work, but their interaction with the Web service can be used for improving our algorithms.

4. METHODOLOGY

Our methodology of this PhD thesis proposal includes content analysis and algorithm design. The working flow is that: model the content of news event by its relations; selecting the important events by different aspects; mingling event to form a summarization with high text quality.

As described in section 1, this research intends to cope with many challenging problems. First, implicit relations across documents such as causal relation normally require inference skills of human, but the problem of automatic modelling the inference skills of human is not mature yet. For that reason, there is a need of efficient algorithms to mimic the way that the human infer things. In a nutshell, the inference chain likely to be affected by the prior background knowledge on events and entities. That suggests us to employ knowledge bases or human knowledge from crowd sources as the background knowledge to support the implicit relation extraction between events. *We rely on an assumption that the knowledge bases can represent the prior knowledge of the human.* We start by studying the literature in human cognition research and plan to embed the information from these studies into our learning models. Nonetheless, there is a side problem related to the performance cost, which is caused by a very large number of relevant articles available on the Web. That requires us a careful algorithmic and storage design. In addition, we apply the approach that we build Web services for helping the Web users and from their interaction we analyze and obtain some "added value" as the "wisdom of the crowd" supporting our system.

Regarding the data for learning and evaluation, we plan to use both existing news corpus, such as NewYork Time corpus, and news collected from the Internet about the "hot" events. We would like exploit social services like Twitter, Wikipedia and/or Lastfm to get the hot events since they are among of the best systems in this category.

To evaluate the quality of our framework, we break our research problem into smaller pieces: event detection, event relation extraction, summarization and then evaluate each of these tasks against related systems in the literature. There has been many related works already done on these subproblems, for example, *"Detection, Representation, and Exploitation of Events in the Semantic Web"* Workshop papers on ISWC 2011 [1] on event detection and event relation extraction. For the summarization task, which is the target of our research, we plan to use the ROUGE evaluation metrics to measure how the generated content correlates with summaries generated by human. Examples include that when we take the temporal aspects into account, we may compare our generated summarization to existing human-generated timelines on the Wikipedia or from some news agencies. In addition, we plan to have human-rated evaluation when there is no ground truth available.

5. PRELIMINARY RESULT

This PhD proposal is still in early stage, hence, in this section, we present some preliminary results on structured summarization where we employ temporal features for our generation model. We refer to this task as timeline summarization generation. Different to existed work on this topic, we used a machine learning approach for learning a criteria and then apply dynamic programming to optimize summarization generation.

[1] http://iswc2011.semanticweb.org/workshops/detection-representation-and-exploitation-of-events/

5.1 Problem Statement

Input: A collection A_C of relevant news articles for a news topic C, and 2 parameters given by users: n, which is maximum number of dates will be included in the timeline, and m, the maximum number of sentences that report main events happening on a day.

Output: A timeline TS_C, which is consist of n day summaries DS_{d_i}, each has date d_i and the day summary S'_{d_i} contains maximum n sentences reporting events on date d_i.

5.2 Proposed Model

5.2.1 *Event*

Following previous studies [6, 25], we model an event as a sentence that reports about it.

5.2.2 *Temporal Information Extraction*

We aim at extracting the pub_date (the published date of the sentence) and the ref_date (the actual date when the reported event happened) given a sentence. The pub_date is straightforward computed as the published date of the news article containing the sentence. For the ref_date extraction, first, we use Heideltime toolkit [19] to tag possible temporal expression on the news. However, we observed in our corpus that there is more than 70% number of sentences having no date expression.

Let a sentence s is represented as a bag of words $(w_1, w_2, .., w_l)$. $P(d_i|s)$ is the probability of the date d_i given the sentence s.

$$\begin{aligned} ref\_date(s) &= \arg \max_{d_i, i=1..|D_C|} P(d_i|s) \\ &= \arg \max_{d_i, i=1..|D_C|} P(s|d_i) * p(d_i) \\ &\approx \arg \max_{d_i, i=1..|D_C|} (\prod_{j=1}^{l} P(w_j|d_i)) \end{aligned} \tag{1}$$

where $P(s|d_i)$ is the probability of that the sentence s appears on the date d_i in the collection A_C and $P(w_j|d_i)$ is the probability the word w_j appearing on d_i. We estimated its value by uni-gram language model. We then group the list of sentence to their ref_date to form the input collection of sentence S_{d_i} for each date d_i.

5.2.3 *Summarization*

Our day summary generation includes 2 folds: first, we predict the relevance score $(I(s_j))$ of all sentences $s_j \in S_{d_i}$ of the given date d_i by using machine learning (ML) regression approach. To train the ML model, for each sentence s_j, we align to every sentence s_k from the summary created by human in our golden data (S''_{d_i}), to find out how many shared words (including synonyms) they have $(align(s_j, s_k))$. The target score of each sentence is their maximum *align* score.

$$I(s_j) = \max_{k=1}^{|S''_d|} align(s_j, s_k) \tag{2}$$

We extract surface features for each sentence, such as: sentence length, ratio of pronouns, number of stop words, the position in the article, similarity with the first sentence of the article, tf.idf score, etc. These features are feed to the ML algorithm for building the model that can predict the $I(s_j)$score for each sentence s_j later in the testing phrase.

Second, we select a subset of sentence $S'_{d_i} \subseteq S_{d_i}$ such that $|S'_{d_i}| \leq m$ based on their predicted relevant score, the higher the better. To avoid redundancy (since many sentences of S_{d_i} may have the same meaning to each other)), we propose the novelty measure for a day summary S'_{d_i} as the difference between a given sentence and a set of other sentences.

$$N(s_i, S) = \frac{|\{w : w \notin \{s_i \cap S\}\}|}{length(s_i)} \tag{3}$$

5.2.4 *Summary Optimization*

We see the problem of sentence selection to build day summary of a date d_i is an optimization problem. Our proposed algorithms applies dynamic programming for doing optimization with time complexity $O(|S_d|^2)$.

Let $f(i, j, k)$ is the maximum target function of the timeline story while we are building the timeline at the date d_i and if we decide to select the sentence s_{ik} from S_{d_i} for the position $j^{th} \leqq m_i$ in the set S'_{d_i}. Let $histS'_{i,j,k} = \bigcup_{u=1}^{j}(s'_u)$ is the current set of j selected sentences s'_u (u = 1.. j-1) at this step.

$$f(i, j, k) = \max\{f(i, j-1, v) + \delta(s_{ik})\} \tag{4}$$

such that
$s_{ik} \in S_{d_i}$ and $s_{ik} \notin histS'_{i,j,v}$, the selected sentence set of the previous step $j-1$ when we select the sentence $s_{iv} \in S_{d_i}$ for the $(j-1)^{th}$ position (i.e, $s'_{j-1} \leftarrow s_{iv}$)

The notation $\delta(s_{ik})$ is for the added value to the target function f when we choose the sentence s_{ik}.

$$\delta(s_{ik}) = \lambda_1 * I(s_{ik}) + \lambda_2 * N(s_{ik}, histS'_{i,j-1,v}) \tag{5}$$

5.3 Experimental Result

We build a corpus o 31 timeline user-generated summaries from Wikipedia and collected total 8666 related news from the Web. 21 timelines are used for training (with 6456 news) and 10 timelines (with 2210 news) are used for testing. We used ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-SU4 (SU4) as the metric for evaluation. We compare with 2 common state of the art multi-document summarization system, namely LexRank [9], MEAD [16], and another timeline generation system Chieu [6] and the Random (randomly select m sentences for each day summary). The results are promising (see the table 1 for results of 4 timelines as an example).

6. CONCLUSION

We have discussed the task of information extraction and structured summarization with application to the information overloading on the Web problem. To our knowledge, studying in structured summarization or representation is fairly new and attractive. In this PhD thesis proposal, we propose to tackle it in the event-based direction where we propose to extract topic hierarchical, causal, temporal, spatial relations between events from the news corpora and apply them to generate some sort of structured summarization for the content of an specific event. We also address to it the use of human knowledge and wisdom of the crowd to build our algorithms. We aim at extracting a selected number of events that can express important aspects of the events. Our research is related to information extraction, summarization

Table 1: Sentence selection evaluation results

	Timeline 28			Timeline 30		
	R1	R2	SU4	R1	R2	SU4
Random	0.19	0.06	0.03	0.15	0.05	0.01
LexRank	0.31	0.10	0.08	0.30	0.09	0.08
MEAD	0.30	0.09	0.07	0.28	0.09	0.06
Chieu	0.25	0.11	0.05	0.22	0.09	0.03
Ours	0.34	0.11	0.12	0.33	0.08	0.10
	Timeline 31			Timeline 27		
	R1	R2	SU4	R1	R2	SU4
Random	0.25	0.03	0.08	0.28	0.06	0.10
LexRank	0.32	0.06	0.12	0.34	0.08	0.16
MEAD	0.31	0.05	0.11	0.27	0.06	0.07
Chieu	0.28	0.05	0.09	0.31	0.07	0.12
Ours	0.33	0.06	0.12	0.36	0.09	0.16

and social computing. We believe that the thesis outcome can benefit not only Web users but the research in related areas.

7. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *Proceedings of ACM SIGIR 2011*, pages 10–18, New York, NY, USA, 2001. ACM.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of ACM SIGIR 1998*, pages 37–45.

[3] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL*, pages 113–120, 2004.

[4] E. Blanco, N. Castell, and D. Moldovan. Causal Relation Extraction. In *Proceedings of LREC'08*, pages 310–313, 2008.

[5] B. Boguraev and R. K. Ando. Timeml-compliant text analysis for temporal reasoning. In *Proceedings of IJCAI'05*, pages 997–1003. Morgan Kaufmann Publishers Inc., 2005.

[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] Q. X. Do, Y. S. Chan, and D. Roth. Minimally supervised event causality identification. In *Proceedings of EMNLP'11*, pages 294–303, 2011.

[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] 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.

[10] R. Girju. Automatic detection of causal relations for question answering. In *Proceedings of ACL workshop MultiSumQA'03*, pages 76–83, 2003.

[11] R. Girju and D. I. Moldovan. Text mining for causal relations. In *Proceedings of FLAIRS'02*, pages 360–364, 2002.

[12] V. Ha-Thuc and J.-M. Renders. Large-scale hierarchical text classification without labelled data. In *Proceedings of the fourth ACM WSDM*, pages 685–694, 2011.

[13] J. He, W. Weerkamp, M. Larson, and M. de Rijke. An effective coherence measure to determine topical consistency in user-generated content. *Int. J. Doc. Anal. Recognit.*, 12(3):185–203, 2009.

[14] J. R. Hobbs. On the coherence and structure of discourse. In *Technical Report CLSI-85-7, Center for the Study of Language and Information*, 1985.

[15] J. Makkonen. Investigations on event evolution in tdt. In *In Proceedings of HLT-NAACL 2003 Student Workshop*, pages 43–48, 2003.

[16] 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.

[17] K. Roberts, T. Goodwin, and S. M. Harabagiu. Annotating spatial containment relations between events. In *Proceedings of LREC'12*, 2012.

[18] D. Shahaf, C. Guestrin, and E. Horvitz. Trains of thought: generating information maps. In *WWW*, pages 899–908, 2012.

[19] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of SemEval'10*, pages 321–324, 2010.

[20] Z. R. Sundermeier BA, van den Broek P. Causal coherence and the availability of locations and objects during narrative comprehension.

[21] H. Takamura and M. Okumura. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th EACL*, pages 781–789, 2009.

[22] T. A. Tuan, S. Elbassuoni, N. Preda, and G. Weikum. Cate: context-aware timeline for entity illustration. In *Proceedings of the 20th WWW'11*, pages 269–272. ACM, 2011.

[23] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of SemEval'07*, pages 75–80, 2007.

[24] M. Verhagen and J. Pustejovsky. Temporal processing with the tarsqi toolkit. In *Proceedings of COLING'08*, pages 189–192, 2008.

[25] 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.

[26] Y. Yang, J. G. Carbonell, R. D. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.