Building Document Graphs for Multiple News Articles Summarization: An Event-Based Approach

Wei Xu[1], Chunfa Yuan[1], Wenjie Li[2], Mingli Wu[2], and Kam-Fai Wong[3]

[1] Department of Computer Science and Technology
Tsinghua University, China
vivian00@mails.tsinghua.edu.cn, cfyuan@mail.tsinghua.edu.cn
[2] Department of Computing,
The Hong Kong Polytechnic University, Hong Kong
{cswjli, csmlwu}@comp.polyu.edu.hk
[3] Department of System Engineering,
The Chinese University of Hong Kong, Hong Kong
kfwong@se.cuhk.edu.hk

Abstract. Since most of news articles report several events and these events are referred in many related documents, we propose an event-based approach to visualize documents as graph on different conceptual granularities. With graph-based ranking algorithm, we illustrate the application of document graph to multi-document summarization. Experiments on DUC data indicate that our approach is competitive with state-of-the-art summarization techniques. This graphical representation which does not require training corpora can be potentially adapted to other languages.

1 Introduction

The main issue of extractive summarization is how to judge the important concept that should be described in the summary. Existing Graph-based ranking algorithms are used to simulating the functioning of human intelligence and are proved to be efficient to identify the salient elements from graph. A graphic representation of documents provides a natural way to model textual units and the relationships that interconnect them on different levels of abstraction. According to the fact that most of news articles report several events and these events are referred in many other documents that are related to the topic, it is better to build event-centric graphs by choosing textual units as event elements (including actions and the entities that participate in the events), events or sentences containing events. In addition, graph solves the problem of reduplicate information by assessing weights of links between nodes.

In this paper, we propose to extract event information and derive intra-event relations between event elements in news articles without deep natural language processing techniques. A weighted document graph is then built to represent the cohesive structure of text, specially emphasizing on events. We evaluate the capability of graph representations on multiple news articles summarization with PageRank [1] ranking algorithms. To focus on the efficiency and potential of event-centric document graphs, we do not consider the other features known to be helpful when creating summaries. We close with the discussion of future work.

Y. Matsumoto et al. (Eds.): ICCPOL 2006, LNAI 4285, pp. 181 – 188, 2006.

2 Related Work

Graph is a relational structure capable of representing the meaning and construction of cohesive text with associative or semantic information, corresponding naturally to human memory. Text visualization has been used to represent the underlying mathematical structure of a text or a group of texts [8]. At the same time, graph-based ranking algorithms has been successfully used in hyperlink analysis [1] and social networks [2], and recently turned into application on natural language processing. These algorithms decide on the importance of a node within a graph through link structure, rather than relying only on local node-specific information.

Extractive summarization emphasizes on how to determine salient pieces from original documents and therefore benefits much from graph-based ranking algorithm. To rank entire sentences for sentence extraction, most of previous works add a node to the graph for each sentence in the text. Different measurements are used to determine how to represent sentence and how to define connections between sentences. The similarity between two sentences according to their term vectors is used to generate links and define link strength in [4]. Similarly, [3] weighed links by the content overlap of two sentences normalized by the length of each sentence. Yoshioka and Haraguchi [6] went one step further taking events into consideration. Two sentences are linked when they share similar events, which are mostly judged by the similarity of words and consistency of date. However, choosing sentences as nodes within graph limits the representation ability of information in documents and the flexibility for further applications. In [5], the importance of the verbs and nouns constructing events is evaluated with PageRank as individual nodes aligned by their dependence relations. Unfortunately, dependency analysis requires syntax processing techniques.

Event-based summarization has been investigated in recent research. As introduced above, [5] and [6] both extracted events information by dependency structure of sentences and then formed a graph for summarization. In contrast, Filatova and Hatzivassiloglou suggested extracting atomic events to capture information about name entities and the relationships between these name entities, avoiding deep structure analysis of sentences [7]. They evaluated sentences only by times of appearance of pairs of name entities and atomic event connectors. The proposed approach claimed to out-perform conventional tf*idf approach on summarization and demonstrated that defining events based on named entities is feasible. However, their event definition is too strict to capture adequate information from texts.

Our work differs from these previous studies in two key respects. First, we propose a novel approach to extract semi-structured events with shallow natural language processing. Second, we build event-centric document graphs to make conceptual information visible and rank textual units for summarization on different granularities.

3 Event-Based Document Graph

3.1 Extraction of Event

Events described in texts link major elements of events (people, companies, locations, times etc.) through actions. In this paper, we use the definition of event proposed in

[8]. Events are anchored on major elements representing as named entities and high frequently occurring nouns, kind of named entities that can not be marked by general named entity taggers. A verb or an action noun is deemed as an event term only when it appears at least once between two named entities. Event terms roughly relate to the actions of events. Thus, we extract events based on named entities and co-occurrence of event elements without syntactic analysis.

Events are extracted from documents by using following steps:

1. Mark texts with named entities and POS tags.
2. Add a frequent noun into the set of named entities (NE) when its appearance times are above a certain threshold.
3. Detect pairs of named entities in every sentence and extract verbs and action nouns as event terms (ET), ignoring stopwords.
4. Scan documents again to extract events as event terms with adjacent named entities. These events take the form as triple $(et_x \mid ne_i, ne_j)$, if the event terms between a pair of named entities; or as couple $(et_y \mid ne_k)$, if the event terms is neighboring with only one named entity in a sentence.

Original:
The <Organization>Justice Department</Organization> and the 20 states <VB>suing</VB> <Organization>Microsoft</Organization> believe that the tape will <VB>strengthen</VB> their <HN>case</HN> because it shows <Person>Gates</Person> saying he was not <VB>involved</VB> in plans to take what the <HN>government</HN> alleges were illegal steps to <VB>stifle</VB> <AN>competition</AN> in the Internet <HN>software</HN> <HN>market</HN>.

Events:
1. {sue | Justice Department, Microsoft}
2. {strengthen | Microsoft, case}
3. {involve | Gates, government}
4.5. {stifle, compete | government, software}

Fig. 1. Example of Event Extraction from a sentence

This approach complements the advantages of statistical techniques and captures semantic information as well. Figure 1 shows an original sentence of news article and five extracted events. The event "sue" represents the structure of Subject-Verb-Object (SVO), whereas the other four events only carry partial relationship of SVO, and "software" is not as proper as "the Internet software market". However, graph-based ranking algorithm calculates the weights of nodes and roughly gets rid of unimportant event elements and extra elements added by mistake.

3.2 Building Document Graph

To form the document graph, we take these events by choosing event elements (event terms and named entities) as nodes. The edges between event elements are established by co-occurrence in a same event. A piece of a graph built by our system for cluster d30026 (DUC 2004) is shown in Figure 2.

The document graph is weighted but undirected. Different from previous work on intra-event relevance [7] [9], the relationship between event elements is measured not only by counting how many times they co-occur in events, but also by taking linguistic structure of sentence into consideration. We observe in real texts that two named entities can be far apart in a long sentence and more than one event terms emerge between them (e.g. "stifle" and "compete" event in Figure 1; event terms in joined rectangles in Figure 2). These adjacent event terms which are associated with same pair of named entities are mostly because of complicate sentence structure, such as subordinate clause. The strength of link between action and named entity within an event is indicated as $L_{event}(et_x, ne_i) = L_{event}(ne_i, et_x) = 1/n$, when n is the number of adjacent event terms between the same named entity (pair). The weight of connection within graph is calculated as $\mathrm{R}(et_x, ne_i) = \mathrm{R}(ne_i, et_x) = \sum L_{event}(ne_i, et_x)$ .Figure 3 enlarges a part of document graph in Figure 2 to show the weight of each edge.

S1 : The Justice Department and the 20 states suing Microsoft believe that the tape will strengthen their case because it shows Gates saying he was not involved in plans to take what the government alleges were illegal steps to stifle competition in the Internet software market.
S2: It showed a few brief clips of a point in the deposition when Gates was asked about a meeting on June 21, 1995, at which, the government alleges, Microsoft offered to divide the browser market with Netscape and to make an investment in the company, which is its chief rival in that market.
S3: In the taped deposition, Gates says he recalled being asked by one of his subordinates whether he thought it made sense to invest in Netscape.
S4: But in an e-mail on May 31, 1995, Gates urged an alliance with Netscape.
S5: The contradiction between Gates' deposition and his e-mail, though, does not of itself speak to the issue of whether Microsoft made an illegal offer to Netscape.

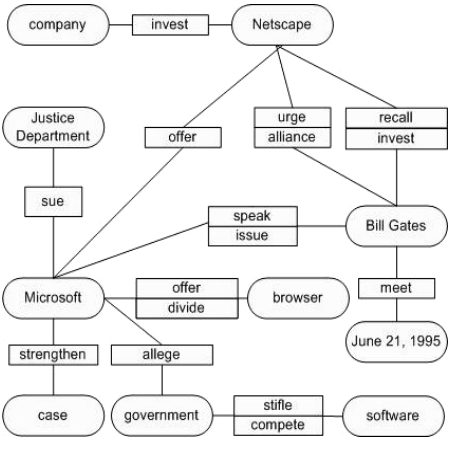

Fig. 2. Document Graph Fragment, on event element level

Since these events are commonly related with one another semantically, temporally, spatially, causally or conditionally, especially when the documents are under the same or related topic, we can derive intra-event relevance between two event terms or two named entities from document graph.

$$\mathrm{R}(et_x, et_y) = [\sum_{ne_i \in \mathrm{NE}(et_x) \cap \mathrm{NE}(et_y)} \mathrm{R}(et_x, ne_i) \bullet \sum_{ne_i} \mathrm{R}(ne_i, et_y)]^{1/2} \quad \text{(E1)}$$

$$\mathrm{R}(ne_i, ne_j) = [\sum_{et_x \in \mathrm{ET}(ne_i) \cap \mathrm{ET}(ne_j)} \mathrm{R}(ne_i, et_x) \bullet \sum_{et_x} \mathrm{R}(et_x, ne_j)]^{1/2} \quad \text{(E2)}$$

Where $\mathrm{NE}(et_x)$ is the set of named entities et_x associates; $\mathrm{ET}(ne_i)$ is the set of event terms ne_i associate.

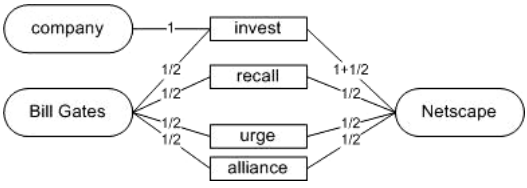

Fig. 3. Weight of link between event terms and named entities

For the convenience to observe organization of document and to investigate certain event or specific sentence with associated contextual information in the future, we design to form document graph on event and sentence level. To determine the strength of events, we have two choices. One is to use a simple cosine similarity based on a measure of event elements overlap and the other is to use the cross strength of relation between event elements. In this paper, we consider only events and neglect other words, thus the second approach is better to make use of event relevancy. As shown in Figure 4 and Figure 5, relations of events are measured by sum all the weights of connections between event elements and similarly, relations of sentence by weights of connections between events.

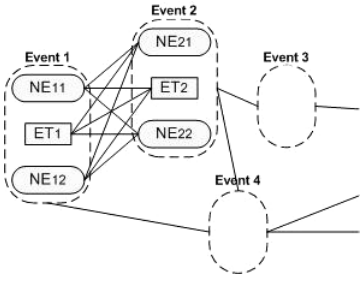

Fig. 4. Sketch Map of Document Graph, on event level

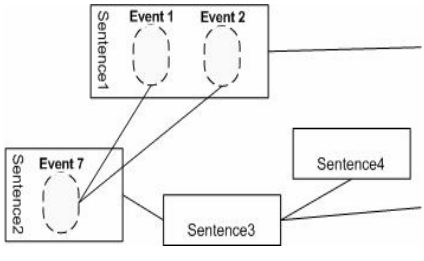

Fig. 5. Sketch Map of Document Graph, on sentence level

3.3 Node Scoring with PageRank for Summarization

To score the significance of nodes in a document graph, our system uses the PageRank algorithm [1]. The thrust of PageRank is that when a node links to more other nodes or links to another "important" node, it becomes more "important". A ranking process starts by assigning arbitrary values to each node in graph and followed by several iterations until convergence.

The formula for calculating Pagerank of a certain node *n* is given as follows:

$$\mathrm{PR}(node_n) = (1-d) + d \sum_{node_i \in L} \frac{\mathrm{PR}(node_n)}{\mathrm{R}(node_i, node_n)} \tag{E3}$$

where L is the set of nodes linking into node n
d is a dampening factor, set to 0.85 experimentally

For different granularity of document graph, the significances of event elements, events and sentences are then scored according to the linking structure and edge weights respectively. After that, the significance of each sentence is obtained by simply summing the significance of the event elements or events it contains. Sentences are extracted for summaries by static greedy algorithm [7], if and only if they cover the most of concepts, removing all duplicate sentences.

With ranking algorithm for graph, process of extractive summarization can be fully unsupervised without training on corpora. Moreover, we can further realize information fusion, sentence compression and sentence generation in the future.

4 Experiments and Discussions

We test our event-based graphical approach by the task of multi-document summarization in DUC 2001(task 2) and DUC 2004(task 2). The documents are pre-processed with GATE to recognize named entities, verbs and nouns.

In order to evaluate the quality of the generated summaries, we use the automatic summary evaluation metric, ROUGE [10]. This metric is found to be highly correlated with human judgments.

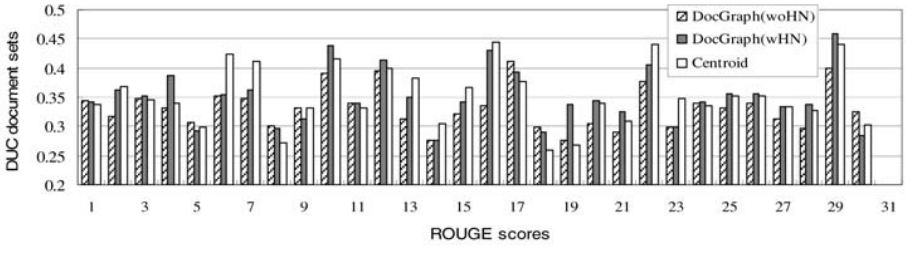

Fig. 6. ROUGE scores, Document Graph (with and without high frequency noun) vs. Centroid

In our first experiment our approach is evaluated on 200-words summaries of DUC 2001. We determine the salient concept by document graph on event element level. We compare the ROUGE scores of adding frequent nouns or not to the set of named entities to our system. A baseline is also included as Centroid-based summarization, which is a widely used and very challenging baseline in the text summarization community [11]. ROUGE scores are reported for each document set rather than average score because ROUGE scores depend on each particular document set (Figure 6). Finally, for 18 sets (60%) out of the 30 document sets, the summary created according to document graph with frequent nouns receives higher ROUGE score than Centroid-based approach. By taking high frequent nouns into the consideration, great improvement is achieved in 20 sets (66.7%) and 5% increase of ROUGE score is gained on average. The advantage of graph-based approach over Centroid is that it indicates redundant information by link weight and prevents improper high idf scores from rare words that are unrelated to the topic.

Next, we compare two methods to measure the strength of relationship between event elements, one is proposed in previous work by times of co-occurrence in events, the other is new in this paper splitting the weight in same named entity pair. As shown in Table 1, a slight improvement is achieved by the new approach. Besides we evaluate this adjustment on different strategies on deriving event relevance by graph-based ranking algorithm in [9], and prove that improvement is slight but constant.

As discussed before, document graph can be constructed by choosing different kinds of nodes. Table 2 shows the result by ranking text units for summarization on different granularity. The advantage of representing with separated actions and entity nodes over simply combining them into event or sentence node is to provide a convenient and effective way for analyzing the relevance between conceptual information. At the same time, the graph on event or sentence level helps people to observe and investigate documents more conveniently.

Table 1. ROUGE scores using different methods to weigh relations in graph

	DUC 2001		DUC 2004	
	co-occurrence times	split weight in same pair	co-occurrence times	split weight in same pair
ROUGE-1	0.35212	0.35250	0.32718	0.33255
ROUGE-2	0.07107	0.07179	0.07027	0.07357
ROUGE-W	0.13603	0.12901	0.12691	0.12949

Table 2. ROUGE scores according to document graph on different level (DUC 2001)

granularity	event elements	event	sentence
ROUGE-1	0.35212	0.33348	0.33957
ROUGE-2	0.07107	0.05886	0.06609
ROUGE-W	0.13603	0.12120	0.12387

5 Conclusion

In this paper, we propose a new approach to present documents by event-based graph and illustrate the application to text summarization. The extraction of event is considered to include basic concepts in news articles as actions and named entities. Document graph makes use of the associations of event elements based on co-occurrence to avoid complex natural language processing techniques. Graph-based ranking algorithm is put forward to determine salience of text units for extractive summarization. The experiment results indicate that this mixed approach of statistics and linguistics is competitive with up-to-date techniques on multiple news articles summarization.

The graph constructed in this way allow further complex processing, such as improving the coherence of summaries by relations and compressing the original

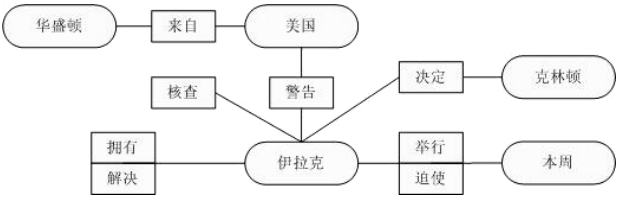

Fig. 7. Document Graph Fragment on Chinese Text

sentences by cutting inessential fragments in the graph. Another advantage of the graph-based document representation and ranking algorithms is that they exclusively rely on the text itself and do not require any training corpora. As a result, our approach can be adapted to other languages. In fact, we have recently attempted to apply the similar method to the texts in Chinese and shown a potential success in summarization (Figure 7).

Acknowledgments. The work presented in this paper is supported partially by National Natural Science Foundation of China (reference number: NSFC 60573186), partially by Research Grants Council on Hong Kong (reference number CERG PolyU5181/03E) and partially by the CUHK strategic grant (# 4410001).

References

1. Page, L., Brin, S.: The Anatomy of a Large-scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 30 (1998) 107-117
2. Dom, B., Eiron, I., Cozzi, A., Shang, Y.: Graph-based ranking algorithms for e-mail expertise analysis. In Proceedings of the 8th ACM SIGMOD workshop on Research Issues in Data Mining and Knowledge Discovery (2003) 42-48
3. Mihalcea, R.: Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (2004) 170-173
4. Erkan, G., Radev D.R.: 2004. LexRank: Graph-based lexical as Salience in Text Summarization. Journal of Artificial Intelligence Research 22 (2004) 457-479
5. Vanderwende, L., Banko, M., Menezes, A.: Event-Centric Summary Generation. In Proceedings of the Document Understanding Conference Workshop (2004)
6. Yoshioka, M., Haraguchi, M.: Multiple News Articles Summarization based on Event Reference Information. In Working Notes of the 4th NTCIR Workshop (2004)
7. Filatova, E., Hatzivassiloglou, V.: Event-based Extractive Summarization. In Proceedings of ACL Workshop on Summarization (2004) 104-111
8. Bradley, J., Rockwell, G.: What Scientific Visualization Teaches Us about Text Analysis. In ALLC/ACH Conference (1994)
9. Li, W., Xu, W., Wu, M., Yuan, C., Lu, Q.: Extractive Summarization using Inter- and Intra- Event Relevance. In Proceedings of COLING-ACL (2006)
10. Lin, C., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceeding of HLT-NAACL (2003) 71-78
11. Radev, D.R., Jing, H., Stys, M., Tam D.: Centroid-based Summarization of Multiple Documents. Information Processing and Management. 40 (2004) 919-938