Deriving Event Relevance from the Ontology Constructed with Formal Concept Analysis

Wei Xu[1,2], Wenjie Li[1], Mingli Wu[1], Wei Li[1], and Chunfa Yuan[2]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{cswxu, cswli, csmlwu, cswli}@comp.polyu.edu.hk
[2] Department of Computer Science and Technology, Tsinghua University, China
{vivian00, cfyuan}@mails.tsinghua.edu.cn

Abstract. In this paper, we present a novel approach to derive event relevance from event ontology constructed with Formal Concept Analysis (FCA), a mathematical approach to data analysis and knowledge representation. The ontology is built from a set of relevant documents and according to the named entities associated to the events. Various relevance measures are explored, from binary to scaled, and from symmetrical to asymmetrical associations. We then apply the derived event relevance to the task of multi-document summarization. The experiments on DUC 2004 data set show that the relevant-event-based approaches outperform the independent-event-based approach.

1 Introduction

Extractive summarization is to select the sentences which contain salient concepts in documents. An important issue with it is what criteria should be used to extract the sentences. Event-based summarization attempts to select and organize the sentences in a summary with respect to the events or the sub-events that the sentences describe [1, 2]. As the relevance of events reveals the significance of events, it helps singling out the sentences with the most core events. However, the event-based summarization techniques reported so far explored the events independently.

In the realm of information retrieval, term relations were commonly derived either from a thesaurus like WordNet or from the corpus where the contexts of the terms were investigated. Likewise, mining event relevance requires taking contexts of event happenings into account. The event contexts in our definition are event arguments, such as participants, locations and occurrence times, etc. They are important in defining events and distinguishing them from one another. By this observation, we make use of the named entities associated with the events as event contexts and characterize the events with the verbs and action-denoting nouns prescribed by the named entities.

In this paper, we present a novel approach to learn event relevance with the event ontology constructed from a set of relevant documents and according to the named entities associated to the events. Formal Concept Analysis (FCA) is employed as an effective learning technique to support the building of the event ontology. Based on the ontology, various relevance measures are explored, from binary to scaled, and from symmetrical to asymmetrical associations. The events are then evaluated with

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 480 – 489, 2006.

their relevance and in turn the sentences are ranked according to the events they describe. Finally, the top-ranked sentences are selected into the summary. The experiments on DUC 2004 data set suggest that the event-relevance-based approaches outperform the independent-event-based approach.

The reminder of this paper is organized as follows. Section 2 discusses the related work. Section 3 explains how event ontology is constructed and introduces different relevance measures. Section 4 introduces event-relevance-based summarization. Section 5 then presents the experiments and evaluations. Finally, Section 6 concludes the paper.

2 Related Work

Event-based summarization has been investigated in previous researches. Daniel Radev and Allison recognized a news topic in multi-document summarization as a series of sub-events according to human understanding of the topic [1]. They determined the degree of the sentence relevance to each sub-event by human judgment and evaluated six extractive approaches. It was concluded in their paper that recognizing the sub-events that comprise a single news event is essential to produce better summaries. However, it is an obstacle to automatically break a news topic into sub-events. Later, in Filatova and Hatzivassiloglou's work [2], they defined atomic events as the links of major constituent parts of the actions (such as participants, locations, and times) through the verbs or action-denoting nouns. They evaluated the sentences based on the co-occurrence statistics of the events and the named entities involved. As a matter of fact, events in the documents are related in some ways. Judging whether the sentences are salient or not and organizing them in a coherent summary can take advantage from event relevance. Unfortunately, it was neglected in their work and most other previous work. On the other hand, Barzilay and Lapata exploited the use of distributional and referential information of discourse entities to improve summary coherence [3]. While they captured text relatedness with entity transition sequences, i.e. entity-based summarization, we will introduce the relevance between events into event-based summarization.

Ontology is described as a hierarchy of concepts related by subsumption relations [4]. It can be seen as a system containing the concepts and their relations, which can be utilized to analyze the relevance between concepts. In addition to its application in machine translation [5], ontlogy was also used as the domain knolwedge to guide infomration extraction and summarization. For instance, Artequakt [6] was a system to generate biographies of artists based on the extracted relations between the entities of interest, by following ontology relation declaration and WordNet. Formal Concept Analysis (FCA) had been used as an effective learning technique for ontology construction. While, Haav constructed ontology with FCA in estate domain presenting taxonomic relations of domain-specific entities [7], Alani et al attempted to build a context-based ontology in clinical domain to help identifying the relevant medical concepts and the types of their relations [8]. Besides, Li also employed FCA to construct IT-domain ontology automatically based on lexicon or corpus [9]. All these work has focused on how to select data sources and attribute sets in FCA for ontology construction. The work presented in this paper is motivated by the successful applica-

tion of FCA in automatic ontology construction and will make use of ontology as a means to evaluate the event relevance for text summarization.

3 Deriving Event Relevance

The event arguments are usually realized as named entities. Based on this observation, we represent an event approximately with a set of event terms prescribed by the associated name entities. An *event*, denoted by *E*, is defined as $E = \{t_i \mid (n_m, t_i, n_n)\}$ in our work, where t_i is the event term, either a verb or an action-denoting noun according to Word-Net's noun hierarchy [10], between the two successive name entities n_m, n_n in a sentence. The assumption behind this definition is that events are delegated by event terms and discriminated and interrelated by the associated name entities. Four types of named entities are currently under the consideration. They are <Person>, <Organization>, <Location> and <Date>.

3.1 Building Event Ontology with FCA

The events in triple patterns consisting of an event term and two name entities, $< n_m, t_i, n_n >$, are extracted from documents. For instance, we can extract two event terms, spoke and attacking from the following illustrative sentence. They are both associated with the <Person> James Clark and <Organization> Microsoft.

> <Organization> Netscape </Organization> chairman <Person> James Clark </Person> spoke boldly of attacking <Organization> Microsoft </Organization> head-on.

The hierarchical structure of event terms, which is deemed as event ontology, is constructed with Formal Concept Analysis (FCA). FCA takes two sets of data, one is called the *object* set and the other is called the *attribute* set, to find a binary relationship between the data of the two sets, and further constructs a so-called formal ontology. Attributes allow more complex relations to be modelled using the ontology.

The associated name entities of event terms conceal the relations between events. We believe that if two events are concerned with the same person or same location, or occurred at the same time, these two events are probably interrelated with each other. To construct event ontology with FCA, event terms are mapped into objects and name entities into attributes. The binary relationship between the event term t_i and the name entities n_j is determined to be 1, if t_i and n_j are associated in a triple pattern. It is 0 otherwise.

A FCA tool, called ConExp[1] (Concept Explorer) can be used to visualize the ontology by lattice, as illustrated in Figure 1. To further interpret the relationships of any two objects, we here define two kinds of relations. Objects are *equivalent* when they are associated with exactly the same attributes (such as t_3 and t_4). The object with

[1] Free downloadable from http://sourceforge.net/projects/conexp.

subset of attributes is considered as a *super-class* of the object with superset of attributes (such as t_1 and t_4). Otherwise, they are not directly related.

obj.\att.	n_1	n_2	n_3	n_4
t_1	1	0	0	1
t_2	0	1	1	0
t_3	1	0	1	1
t_4	1	0	1	1

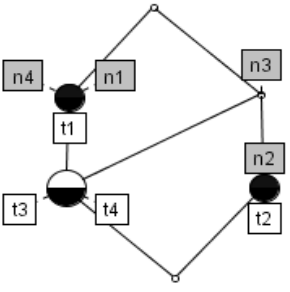

Fig. 1. Example of event ontology (event terms as objects, associated name entities as attributes)

The relations of objects are explicitly indicated in the lattice. As shown in Figure 1, the equivalent objects are denoted by a same node. The nodes in the upper levels are actually the super-classes of those in the lower levels.

3.2 Measuring Event Relevance

We propose the following event relevance measures by exploring the previously constructed ontology. The relevance between t_i and t_j is denoted by $R(t_i, t_j)$.

We first start from clusters (i.e. the nodes in ontology) provided by ontology. Event terms are assumed to be relevant only if they are in the same node (i.e. they are equivalent as ontology specifies). In such a way, the relevance is symmetrical in nature. This is where the idea of the approach Binary and Symmetrical Measure 1 (BSM1) comes from.

As the super/sub class relations are taken into the consideration, the unbalanced relations exhibit. As illustrated in Figure 1, if t_1 is the super-class of t_4, all its attributes, n_1 and n_4, are included in t_4's attribute set. This relation is not hold for t_4, because it has one more attribute n_3. Therefore when t_i is the super-class of t_j, the relation from t_i to t_j is assumed to be stronger than from t_j to t_i, i.e. $R(t_i, t_j) > R(t_j, t_i)$. The approach Binary and Asymmetrical Measure (BAM) are therefore introduced to cope with these unbalanced relations.

To go further, we consider not only the nodes directly related but also those indirectly related, such as t_2 and t_4. They are neither equivalent in one node and nor related by super/sub class relation. But they are indirectly liked by some common super-class, which is a virtual node in Figure 1. The indirect relevance is measured with the approach Binary and Symmetrical Measure 2 (BSM2).

Finally, the scaled approaches are extended from the binary approaches. Whereas the binary value can only represent whether two event terms are relevant or not, the scaled value indicate how strong the two event terms are related. In conclusion, based on event ontology constructed, several approaches, varied from binary to scaled and symmetrical to asymmetrical, are proposed to measure the event relevance in our work:

- Binary and Symmetrical Measure 1 (BSM1): If two event terms t_i, t_j are equivalent, $R(t_i,t_j) = R(t_j,t_i) = 1$. Otherwise R is 0.
- Binary and Asymmetrical Measure (BAM): BAM is the extension of BSM1. In addition to handle the equivalence terms in the same way as in BSM1, if the event term t_i is the super-class of the event term t_j, $R(t_i,t_j) = 1, R(t_j,t_i) = 0$. Otherwise R is 0.
- Binary and Symmetrical Measure 2 (BSM2): BSM2 is a further extension from BAM. If two event terms t_i, t_j have at least one attribute in common, $R(t_i,t_j) = R(t_j,t_i) = 1$, Otherwise R is 0. On the ontology, these two event terms are either equivalent, directly related by super/sub classes or indirectly related with at least one super-class node in common.
- Scaled and Asymmetrical Measure 1 (SAM1): SAM1 is an extended BAM assessing event relevance by decimal fraction instead of binary value. If the event term t_i is a super-class of the event term t_j, and t_i has k attributes $n_{i1}, n_{i2}, ..., n_{ik}$, t_j has l attributes $n_{j1}, n_{j2}, ..., n_{jl}$ ($k<l$), then $R(t_i,t_j)$ is 1 and $R(t_j,t_i)$ is k/l. Otherwise R is 0.
- Scaled and Asymmetrical Measure 2 (SAM2): Similarly, SAM2 is extended from BSM2. Suppose the event term t_i has k attributes $n_{i1}, n_{i2}, ..., n_{ik}$ and the event term t_j has l attributes $n_{j1}, n_{j2}, ..., n_{jl}$. If t_i and t_j have m attributes in common, then $R(t_i,t_j)$ is m/k and $R(t_j,t_i)$ is m/l. Otherwise R is 0. SAM2 is also an extension from SAM1 in the sense that the nodes with common super-classes are also considered as relevant.

The matrix representation is suitable to formalize the relevance between any two events terms. The value at the cross of column t_i and row t_j is $R(t_i,t_j)$. For instance, the matrix provided by BAM with the data given in Figure 1 is shown in Figure 2.

Relevance	t_1	t_2	t_3	t_4
t_1	-	0	1	1
t_2	0	-	0	0
t_3	0	0	-	1
t_4	0	0	1	-

Fig. 2. Example of relevance measure with BAM with the example data given in Fig.1

4 Summarization with Event Relevance

Given event term relevance, if an event term is relevant with more other event terms, it is assumed to be more significant in representing a salient concept. The event terms relevant to the significant terms are thereby more close to the salient concept than those not. We estimate term significance with *PageRank*, an efficient algorithm to exploit event term maps by linking relevant terms together [11]. It assigns the significance score to each event term according to the number of event terms linking to it as well as the strength of the links. The equation to calculate the page rank (indicated by *PR*) of a certain term *A* is shown as follows:

$$PR(A) = (1-d) + d(\frac{PR(B_1)}{C(B_1)} + \frac{PR(B_2)}{C(B_2)} + ... + \frac{PR(B_t)}{C(B_t)}) \quad . \tag{1}$$

In expression (1), B_1, B_2, … B_t are all terms which link to term *A*. $C(B_i)$ is the number of outgoing links from term B_i. *d* is the factor used to avoid the limitation of loop in the map structure. The significance score of each term can be obtained recursively with this equation. The significance of each sentence to be included in the summary is then calculated from the significance of the event terms it contains.

5 Experiment, Evaluation and Discussion

5.1 Evaluation on Event-Based Summarization

To evaluate the effectiveness of integrating event relevance into multi-documents summarization, we conduct the experiments on the 50 sets of English documents from DUC 2004 multi-document summarization task. The documents are pre-processed with GATE[2] to recognize the previously mentioned four types of name entities[3]. In average, each set contains 10 documents, 149 event terms and 76 name entities.

Figure 3 shows an example of event ontology constructed with FCA based on a paragraph of real news in DUC 2004 data. This paragraph is about the Microsoft Corp.'s firm grip on the personal computer software business. As shown in Figure 3, much important information about this topic is extracted from the news, such as the

[2] Free downloadable from http://gate.ac.uk.

[3] GATE also provides other types of named entities. But only four of them are recognized to fit our application.

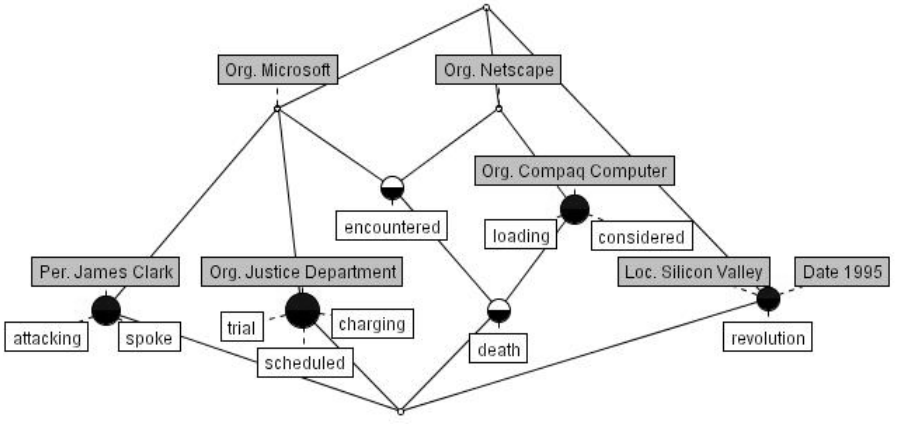

Fig. 3. Event ontology constructed from a paragraph of real news

names of IT companies and Justice Department, the time and location. The node of revolution event is associated with the location Silicon Valley and the year 1995. The efficiency of extracting events, which carry the most important information, from texts was also discussed in [2, 12]. A paragraph of news is shorter comparing to a set of topically related texts in the task of multi-document summarization where many links between events might be presented. But it somehow provides evidence that the relevant events or event terms can be discovered with FCA. For example, the verb trial and charging are equivalent in one node. The action noun death which is a probable consequence of the verb encountered is the super-class of encountered.

To evaluate the quality of summaries, we use an automatic summary evaluation metric ROUGE[4], which has been used in DUCs. ROUGE is a recall-based metric for fixed length summaries. It bases on *N*-gram co-occurrence and compares the system-produced summaries to human judges [13]. For each DUC document set, we create a summary of length less than 665 bytes and present three of the ROUGE metrics: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on longest common subsequence weighed by the length).

Table 1 compares the ROUGE evaluations of relevance-based approaches with the baseline of using event term centroid scheme as sentences selection criteria. As it

Table 1. Evaluations of event-relevance-based and idependent-event-based summarization approaches

	ROUGE-1	ROUGE-2	ROUGE-W
centroid	0.28042	0.04570	0.10858
BSM1	0.28746	0.04339	0.11053
SAM1	0.29062	0.04756	0.11206
BAM	0.29760	0.04662	0.11589
BSM2	0.30166	0.05519	0.11658
SAM2	0.30192	0.05240	0.11774

[4] http://www.isi.edu/~cyl/ROUGE/

indicates, the summaries created by event relevance receive a higher ROUGE score than the baseline summaries created by independent events. Better results are from SAM2 and BSM2. This is natural because they consider the terms related together indirectly. The same can also explain the improvements from BSM1 to BAM to BSM2. When name entity recognition and entity co-reference are not quite successful nowadays, the strict approaches, which consider direct relations only, are more error sensitive. Unfortunately, the asymmetrical measures proposed do not significantly outperform the symmetrical measures right now. In this first set of experiments, we do not merge the attributes. The issue of merging named entities will be discussed in the next subsection.

5.2 Discussion on Named Entity Mention Links

Our work depends on named entities to determine the relevance of events. During experiments, we observe some redundant attributes. Take the set of news about Cambodia as example. Several person names are extracted as follows,

Ranariddh
Prince Norodom Ranariddh
Norodom Sihanouk
Sihanouk
President Prince Norodom Ranariddh
King Norodom Sihanouk

Actually, these six names mentioned above correspond to two person entities, i.e. Prince Norodom Ranariddh and King Norodom Sihanouk. However, they are considered as distinct attributes to differentiate the event terms simply because their surface texts are different. FCA provides the function to merge the redundant attributes. If the named entity mentions that represent the same or similar entities could be linked together (this is hereafter referred to as entity normalization), efficiency and precision of event relevance discovery might be improved. At present, we only consider the person's names as an initial step to investigate the contributions of entity normalization, because of its observable repetitions in texts and its relatively straightforward variations. The clustering algorithm for linking person name mentions is given below:

Step1: For each person name $p_i = w_{i1}w_{i2}...w_{ik}$, w are the words in person name. Its person cluster $C(p_i)$ is initialed by the person name p_i.

Step2: For each person name $p_i = w_{i1}w_{i2}...w_{ik}$

For each person name $p_j = w_{j1}w_{j2}...w_{jl}$, if $C(p_i)$ is a substring of $C(p_j)$, then $C(p_i) = C(p_j)$.

Continue Step 2 until no change occurs.

This simple algorithm can avoid merging names overly. For instance, if A= ab, B=a, C=b, then A=B=a in iteration 1, and C can no longer merge with A or B. If A=abc, B=a, C=ab, then A=B=a in iteration 1 and C=A=B=a in next iteration.

Table 2 shows that named entity mention links affect the performance in some extend but not evidently. The most likely reason is that person names are not contained in all events. The results of these experiments also show an interesting phenomenon, i.e. entity mention links improve the performance of BSM1, have no effect on BAM and SAM1 yet cause decreases in BSM2 and SAM2. These results corroborate the previous conclusions. The automatic recognition of name entities unavoidably introduces errors. When the restriction of event relevance is getting less from BSM1 to BAM and then to BSM2, these errors are amplified gradually. When more events are related together in BSM2 and SAM2, the significance of events is indistinct with *PageRank* algorithm. In contrast, the stricter approaches benefit from the entity mention links for the same reason. The experiments suggest that the improvement of name entity recognition can help the event-based summarization.

Table 2. Result: with and without linking entity mentions

		ROUGE-1	ROUGE-2	ROUGE-W
BSM1	Without	0.28746	0.04339	0.11053
	With	0.28790	0.04453	0.11098
SAM1	Without	0.29062	0.04756	0.11206
	With	0.29243	0.04832	0.11286
BAM	Without	0.29760	0.04662	0.11589
	With	0.29760	0.04662	0.11589
BSM2	Without	0.30166	0.05519	0.11658
	With	0.30042	0.05523	0.11654
SAM2	Without	0.30192	0.05240	0.11774
	With	0.29929	0.05198	0.11584

6 Concluding Remark

In this paper, we propose a novel approach for measuring event relevance and integrating event relevance into text summarization. The experimental results indicate that event relevance is effective for extracting the salient concepts in document sets. The discussion on entity mention links shows that the improvement of named entity recognition and entity co-reference can benefit the event-based summarization.

Our approach can be further improved in the following directions. First, we consider refining the definition of event to capture the corresponding name entities more exactly. Second, we are considering prioritizing special name entities to improve the precision of event relevance. Third, we are also looking at extending the name entities to the common entities for associating events.

Acknowledgements

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

References

1. N. Daniel, D. Radev and T. Allison: Sub-event based Multi-document Summarization. In Proceedings of the HLT-NAACL 2003 Workshop on Text Summarization, (2003) pp9-16.
2. E. Filatova and V. Hatzivassiloglou: Event-based Extractive summarization. In Proceedings of ACL 2004 Workshop on Summarization, (2004) pp104-111.
3. R. Barzilay and M. Lapata: Modeling Local Coherence: An Entity-based Approach. In Proceedings of ACL 2005, (2005) pp141-148.
4. N. Guarino: Formal Ontology and Information Systems. In Proceedings of FOIS 98, Amsterdam, (1998) pp3-15.
5. K. Kevin: Building a Large Ontology for Machine Translation. In Proceedings of the ARPA Human Language Technology Workshop, (1993).
6. H. Alani, S. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, and N.R. Shadbolt: Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems, (2003) 8(1):14-21.
7. H.M. Haav: An Application of Inductive Concept Analysis to Construction of Domain-specific Ontologies. In Proceedings of the Workshop of VLDB2003, (2003) pp63-67.
8. G. Jiang, K. Ogasawara, A. Endoh, T. Sakurai: Context-based Ontology Building Support in Clinical Domains using Formal Concept Analysis. International Journal of Medical Informatics, (2003) 71(1):71-81.
9. S.J. Li, Q. Lu, W.J. Li: Experiments of Ontology Construction with Formal Concept Analysis. In Proceedings of IJCNLP 05 Workshop on Ontologies and Lexical Resources, (2005).
10. C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, (1998).
11. L. S. Page, B.R. Motwani, and T. Winograd. The Pagerank Citation Ranking: Bring Order to the Web. Technical Report, Stanford University, (1998).
12. E. Filatova and V. Hatzivassiloglou. Domain-independent Detection, Extraction, and Labeling of Atomic Events. In Proceedings of RANLP, (2003) pp145-152.
13. C.Y. Lin and E. Hovy: Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics. In Proceedings of HLT-NAACL, (2003) pp71-78.