Extractive Summarization Based on Event Term Clustering

Maofu Liu[1,2], Wenjie Li[1], Mingli Wu[1] and Qin Lu[1]

[1]Department of Computing
The Hong Kong Polytechnic University
{csmfliu, cswjli, csmlwu, csluqin}@comp.polyu.edu.hk

[2]College of Computer Science and Technology
Wuhan University of Science and Technology
mfliu_china@hotmail.com

Abstract

Event-based summarization extracts and organizes summary sentences in terms of the events that the sentences describe. In this work, we focus on semantic relations among event terms. By connecting terms with relations, we build up event term graph, upon which relevant terms are grouped into clusters. We assume that each cluster represents a topic of documents. Then two summarization strategies are investigated, i.e. selecting one term as the representative of each topic so as to cover all the topics, or selecting all terms in one most significant topic so as to highlight the relevant information related to this topic. The selected terms are then responsible to pick out the most appropriate sentences describing them. The evaluation of clustering-based summarization on DUC 2001 document sets shows encouraging improvement over the well-known PageRank-based summarization.

1 Introduction

Event-based extractive summarization has emerged recently (Filatova and Hatzivassiloglou, 2004). It extracts and organizes summary sentences in terms of the events that sentences describe.

We follow the common agreement that event can be formulated as "[Who] did [What] to [Whom] [When] and [Where]" and "did [What]" denotes the key element of an event, i.e. the action within the formulation. We approximately define the verbs and action nouns as the event terms which can characterize or partially characterize the event occurrences.

Most existing event-based summarization approaches rely on the statistical features derived from documents and generally associated with single events, but they neglect the relations among events. However, events are commonly related with one another especially when the documents to be summarized are about the same or very similar topics. Li et al (2006) report that the improved performance can be achieved by taking into account of event distributional similarities, but it does not benefit much from semantic similarities. This motivated us to further investigate whether event-based summarization can take advantage of the semantic relations of event terms, and most importantly, how to make use of those relations. Our idea is grouping the terms connected by the relations into the clusters, which are assumed to represent some topics described in documents.

In the past, various clustering approaches have been investigated in document summarization. Hatzivassiloglou et al (2001) apply clustering method to organize the highly similar paragraphs into tight clusters based on primitive or composite features. Then one paragraph per cluster is selected to form the summary by extraction or by reformulation. Zha (2002) uses spectral graph clustering algorithm to partition sentences into topical groups. Within each cluster, the saliency scores of terms and sentences are calculated using mutual reinforcement principal, which assigns high salience scores to the sentences that contain many terms with high salience scores. The sentences and key phrases are selected by their saliency scores to generate the summary. The similar work based on topic or event is also reported in (Guo and Stylios, 2005).

The granularity of clustering units mentioned above is rather coarse, either sentence or paragraph. In this paper, we define event term as clustering

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 185–188,
Prague, June 2007.

unit and implement a clustering algorithm based on semantic relations. We extract event terms from documents and construct the event term graph by linking terms with the relations. We then regard a group of closely related terms as a topic and make the following two alterative assumptions:

(1) If we could find the most significant topic as the main topic of documents and select all terms in it, we could summarize the documents with this main topic.

(2) If we could find all topics and pick out one term as the representative of each topic, we could obtain the condensed version of topics described in the documents.

Based on these two assumptions, a set of cluster ranking, term selection and ranking and sentence extraction strategies are developed. The remainder of this paper is organized as follows. Section 2 introduces the proposed extractive summarization approach based on event term clustering. Section 3 presents experiments and evaluations. Finally, Section 4 concludes the paper.

2 Summarization Based on Event Term Clustering

2.1 Event Term Graph

We introduce VerbOcean (Chklovski and Pantel, 2004), a broad-coverage repository of semantic verb relations, into event-based summarization. Different from other thesaurus like WordNet, VerbOcean provides five types of semantic verb relations at finer level. This just fits in with our idea to introduce event term relations into summarization. Currently, only the stronger-than relation is explored. When two verbs are similar, one may denote a more intense, thorough, comprehensive or absolute action. In the case of change-of-state verbs, one may denote a more complete change. This is identified as the stronger-than relation in (Timothy and Patrick, 2004). In this paper, only stronger-than is taken into account but we consider extending our future work with other applicable relations types.

The event term graph connected by term semantic relations is defined formally as $G = (V, E)$, where V is a set of event terms and E is a set of relation links connecting the event terms in V. The graph is directed if the semantic relation has the characteristic of the asymmetric. Otherwise, it is undirected. Figure 1 shows a sample of event term graph built from one DUC 2001 document set. It is a directed graph as the *stronger-than* relation in VerbOcean exhibits the conspicuous asymmetric characteristic. For example, "fight" means to attempt to harm by blows or with weapons, while "resist" means to keep from giving in. Therefore, a directed link from "fight" to "resist" is shown in the following Figure 1.

Relations link terms together and form the event term graph. Based upon it, term significance is evaluated and in turn sentence is judged whether to be extracted in the summary.

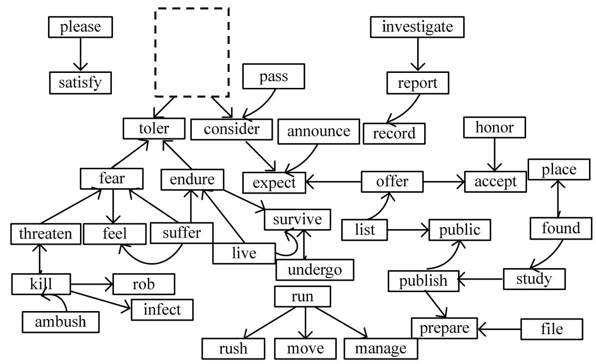

Figure 1. Terms connected by semantic relations

2.2 Event Term Clustering

Note that in Figure 1, some linked event terms, such as "kill", "rob", "threaten" and "infect", are semantically closely related. They may describe the same or similar topic somehow. In contrast, "toler", "resist" and "fight" are clearly involved in another topic; although they are also reachable from "kill". Based on this observation, a clustering algorithm is required to group the similar and related event terms into the cluster of the topic.

In this work, event terms are clustered by the DBSCAN, a density-based clustering algorithm proposed in (Easter et al, 1996). The key idea behind it is that for each term of a cluster the neighborhood of a given radius has to contain at least a minimum number of terms, i.e. the density in the neighborhood has to exceed some threshold. By using this algorithm, we need to figure out appropriate values for two basic parameters, namely, *Eps* (denoting the searching radius from each term) and *MinPts* (denoting the minimum number of terms in the neighborhood of the term). We assign one semantic relation step to *Eps* since there is no clear distance concept in the event term graph. The value of *Eps* is experimentally set in our experiments. We also make some modification on Easter's DBSCAN in order to accommodate to our task.

Figure 2 shows the seven term clusters generated by the modified DBSCAN clustering algorithm from the graph in Figure 1. We represent each cluster by the starting event term in bold font.

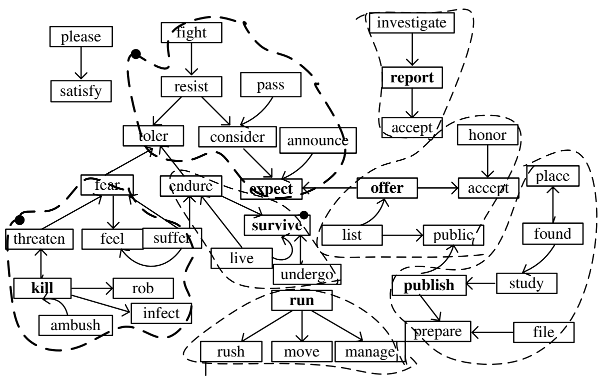

Figure 2. Term clusters generated from Figure 1

2.3 Cluster Ranking

The significance of the cluster is calculated by

$$sc(C_i) = \sum_{t \in C_i} d_t \,/ \sum_{C_i \in C} \sum_{t \in C_i} d_t$$

where d_t is the degree of the term t in the term graph. C is the set of term clusters obtained by the modified DBSCAN clustering algorithm and C_i is the *ith* one. Obviously, the significance of the cluster is calculated from global point of view, i.e. the sum of the degree of all terms in the same cluster is divided by the total degree of the terms in all clusters.

2.4 Term Selection and Ranking

Representative terms are selected according to the significance of the event terms calculated within each cluster (i.e. from local point of view) or in all clusters (i.e. from global point of view) by

LOCAL: $st(t) = d_t \,/ \sum_{t \in c_i} d_t$ or

GLOBAL: $st(t) = d_t \,/ \sum_{c_i \in C} \sum_{t \in c_i} d_t$

Then two strategies are developed to select the representative terms from the clusters.

(1) One Cluster All Terms (**OCAT**) selects all terms within the first rank cluster. The selected terms are then ranked according to their significance.

(2) One Term All Cluster (**OTAC**) selects one most significant term from each cluster. Notice that because terms compete with each other within clusters, it is not surprising to see $st(t_1) < st(t_2)$ even when $sc(c_1) > sc(c_2)$, $(t_1 \in c_1, t_2 \in c_2)$. To address this problem, the representative terms are ranked according to the significance of the clusters they belong to.

2.5 Sentence Evaluation and Extraction

A representative event term may associate to more than one sentence. We extract only one of them as the description of the event. To this end, sentences are compared according to the significance of the terms in them. **MAX** compares the maximum significance scores, while **SUM** compares the sum of the significance scores. The sentence with either higher MAX or SUM wins the competition and is picked up as a candidate summary sentence. If the sentence in the first place has been selected by another term, the one in the second place is chosen. The ranks of these candidates are the same as the ranks of the terms they are selected for. Finally, candidate sentences are selected in the summary until the length limitation is reached.

3 Experiments

We evaluate the proposed approaches on DUC 2001 corpus which contains 30 English document sets. There are 431 event terms on average in each document set. The automatic evaluation tool, ROUGE (Lin and Hovy, 2003), is run to evaluate the quality of the generated summaries (200 words in length). The tool presents three values including unigram-based ROUGE-1, bigram-based ROUGE-2 and ROUGE-W which is based on longest common subsequence weighted by the length.

Google's PageRank (Page and Brin, 1998) is one of the most popular ranking algorithms. It is also graph-based and has been successfully applied in summarization. Table 1 lists the result of our implementation of PageRank based on event terms. We then compare it with the results of the event term clustering-based approaches illustrated in Table 2.

	PageRank
ROUGE-1	0.32749

ROUGE-2	0.05670
ROUGE-W	0.11500

Table 1. Evaluations of PageRank-based Summarization

LOCAL+OTAC	MAX	SUM
ROUGE-1	0.32771	0.33243
ROUGE-2	0.05334	0.05569
ROUGE-W	0.11633	0.11718
GLOBAL+OTAC	MAX	SUM
ROUGE-1	0.32549	0.32966
ROUGE-2	0.05254	0.05257
ROUGE-W	0.11670	0.11641
LOCAL+OCAT	MAX	SUM
ROUGE-1	0.33519	0.33397
ROUGE-2	0.05662	0.05869
ROUGE-W	0.11917	0.11849
GLOBAL+OCAT	MAX	SUM
ROUGE-1	0.33568	0.33872
ROUGE-2	0.05506	0.05933
ROUGE-W	0.11795	0.12011

Table 2. Evaluations of Clustering-based Summarization

The experiments show that both assumptions are reasonable. It is encouraging to find that our event term clustering-based approaches could outperform the PageRank-based approach. The results based on the second assumption are even better. This suggests indeed there is a main topic in a DUC 2001 document set.

4 Conclusion

In this paper, we put forward to apply clustering algorithm on the event term graph connected by semantic relations derived from external linguistic resource. The experiment results based on our two assumptions are encouraging. Event term clustering-based approaches perform better than PageRank-based approach. Current approaches simply utilize the degrees of event terms in the graph. In the future, we would like to further explore and integrate more information derived from documents in order to achieve more significant results using the event term clustering-based approaches.

Acknowledgments

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU5181/03E).

References

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries using *N*-gram Cooccurrence Statistics. In Proceedings of HLT/ NAACL 2003, pp71-78.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based Extractive Summarization. In Proceedings of ACL 2004 Workshop on Summarization, pp104-111.

Hongyuan Zha. 2002. Generic Summarization and keyphrase Extraction using Mutual Reinforcement Principle and Sentence Clustering. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002. pp113-120.

Lawrence Page and Sergey Brin, Motwani Rajeev and Winograd Terry. 1998. The PageRank CitationRanking: Bring Order to the Web. Technical Report,Stanford University.

Martin Easter, Hans-Peter Kriegel, Jörg Sander, et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA, 1996. 226-231.

Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. 1998. The PageRank CitationRanking: Bring Order to the Web. Technical Report,Stanford University.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In Proceedings of Conference on Empirical Methods in Natural Language Processing, 2004.

Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, et al. 2001. Simfinder: A Flexible Clustering Tool for Summarization. In Workshop on Automatic Summarization, NAACL, 2001.

Wenjie Li, Wei Xu, Mingli Wu, et al. 2006. Extractive Summarization using Inter- and Intra-Event Relevance. In Proceedings of ACL 2006, pp369-376.

Yi Guo and George Stylios. 2005. An intelligent summarization system based on cognitive psychology. Journal of Information Sciences, Volume 174, Issue 1-2, Jun. 2005, pp1-36.