Category oriented Extractive Content Selection for Guided Summarization

Ruifang He, Kang Fu, Xinya Zhou
Information System and Software Engineering Lab
School of Computer Science and Technology
Tianjin University, Tianjin, 15001, China
rfhe@tju.edu.cn

Abstract

The guided summarization aims to capture evolving information of a single topic changing over time under the background of much emergency happenings. It delivers salient and novel information to a user who has already read a set of older documents covering the same topic. Topic is category oriented, however, there is no query description. Therefore, guided summarization raises new challenges. In this paper, the category oriented extractive content selection method for guided summarization is proposed, which is completely language independence. Meanwhile, we submitted two systems. Our systems rank top 5 under PYRAMID metrics among 48 running systems. However, we rank middle under ROUGE and BE averagely. Our methods show the great difference for different evaluation metrics. Therefore, we try to think why this happens? What is the best method? What is the best evaluation metrics?

1 Introduction

Guided summarization task is proposed by Text Analysis Conference(TAC) 2010, which originates from the update summarization task in DUC 2007. The main difference is that the topic for guided summarization is emergency category oriented and there is no topic description; while update summarization handles the ordinary news series, and there is topic description. The emergency happens frequently in recent years. Therefore the categories mostly contain the abnormal events, such as accidents and natural disasters, attacks (Criminal/Terrorist), health and safety, endangered resources, investigations and trials (Criminal/Legal/Other)and so on. The event with these categories belongs to a small probability event. Once it happens, it will bring the great loss of life and estate. Therefore, we need to fast grasp the evolving information about an emergency, which essentially belongs to the semantic understanding. Such kind of reality need brings us the new challenges for guided summarization task, and also natural language processing and information retrieval.

Specifically, the guided summarization task was to write a 100-word summary of a set of 10 newswire articles for a given topic, where the topic falls into a predefined category. Participants (and human summarizers) were given a list of aspects for each category, and a summary must include all aspects found for its category.

Additionally, an update component of the guided summarization task was to write a 100-word update summary of a subsequent 10 newswire articles for the topic, under the assumption that the user had already read the earlier articles.

There are three main challenges brought by guided summarization: (1) Semantic understanding of an emergency. Topic is category oriented, yet there is no topic description. This challenge needs us to fully understand the semantic information of the emergency. Thus could we extract the relevantly evolving information guided by a goal. (2) The capture of evolving information. An emergency usually has its life cycle, including birth, growth, decay, and death, reflecting its popularity over time. Therefore, people hope to incrementally care the important and novel information so as to reduce the burden of acquiring information. (3) The balanced coverage of summary content. An emergency has many threads. How could we keep them in a limited summary?

2 System Description

In order to model the new challenges, we submitted two systems based on TAC 2008 (He et al., 2008), designing three groups of experiments: (1) query construction; (2) evolutionary manifold-ranking; (3) spectral clustering based on eigenvector selection. Our system designs are as follows.

RUN1(ID=13): (1) + (2);

RUN2(ID=7): (1) + (2) + (3);

Now we illustrate the steps of framework.

2.1 Semantic Understanding of an Emergency

Since guided summarization is category oriented and there is no topic description, we need to learn the domain knowledge and construct the query. Through analyzing the guided summarization corpus in 2010, we extract the important domain keywords, such as verbs and nouns. For the different aspects of a topic in a category, such as who, what, where, when and so on, the combination of these elements can be understood to be an event in a concrete context. Through analyzing the event extraction task of ACE (Automatic Content Extraction), most noun and verb can trigger the happening of an event, which essentially indicate the 'who' and 'what' of an event. therefore we just only extract the important nouns and verbs to express the pseudo query intent for a topic due to the limited time.

About the importance measure of event trigger, we just use the simple statistical method of word frequency. Moreover, we need to remove the stopping words. Then the most effective trigger words are used for the pseudo query construction in order to rebuild the evolutionary manifold-ranking algorithm.

Here, we do not use any web resources, such as wikipedia and so on. Due to the constraint of the emergency category, domain-specific ontology modeling for semantic understanding will be beneficial to the pseudo query construction.

2.2 Constructing the Similarity Graph

No matter evolutionary manifold-ranking or spectral clustering, building the similarity graph is necessary. Here, we scatter documents into sentences. Every sentence can be considered to be a node of a graph.

Given a set of documents, let an undirected and weighted similarity graph $G = (V, E)$ to reflect the relationships between sentences in document set. Vertex set $V = \{x_1, ...x_n\}$ denotes the sentence set, each vertex x_i in V is a sentence. E is the set of edges, which is a subset of $V \times V$. Each edge between two vertices x_i and x_j carries a non-negative pair-wise similarity weight $w_{ij}(i \neq j)$ in E.

Here, we just use the standard Cosine measure(Erkan and Radev, 2004) to compute the similarity values, which is denoted as $w_{ij} = sim(x_i, x_j)$. We remove the stop words in each sentence, and stem the remaining words. The weight associated with term t is calculated with the $tf_t * isf_t$ formula, where tf_t is the frequency of term t in the sentence and isf_t is the inverse sentence frequency of term t, i.e. $1 + log(N/n_t)$, N is the total number of sentences and n_t is the number of the sentences containing term t. Then $sim(x_i, x_j)$ is computed according to the normalized inner product of the corresponding term vectors.

We define the weighted affinity matrix $W = \{w_{ij}|i, j = 1, ..., n\}$. If $w_{ij} > 0$, the vertices x_i and x_j are connected, or there is no link. Simultaneously, we let $w_{ii} = 0$ to avoid self transition. Since G is undirected, W is a symmetric matrix. D is the diagonal matrix with (i, i)-element equal to the sum of the i-th row of W.

2.3 The Capture of Evolving Information

The manifold-ranking method (Zhou et al., a; Zhou et al., b; Wan et al., 2007) is a universal ranking algorithm, which ranks the data points along their underlying manifold structure according to their relevance to the query. Yet it cannot model the temporally evolving characteristic of dynamic news series. We propose a new evolutionary manifold-ranking frame based on iterative feedback mechanism for guided summarization, which has the temporally adaptive characteristic.

We assume that the data points evolving over time have the long and narrow manifold structure. However, there is no topic description for guided summarization track. Thus the first thing of our method is to construct the topic description, say the query. While the common topic for dynamic document collection is a static query, which cannot represent the dynamically evolving information. Therefore, we use the iterative feedback mechanism to extend the topic description by using the summarization sentence of previous time slices and the first sentences of documents in current times lice. We assume this topic extension could represent the relay propagation of information in temporally evolving data points and help to capture the changes of a single topic over time.

This approach employs iterative feedback based evolutionary manifold-ranking process to compute the ranking score for each sentence, and then the sentences highly overlapping with other informative ones are penalized by the greedy algorithm. The summary is produced by choosing the sentences with highest overall scores, which are considered to be informative, novel and evolving.

Based on the semantic understanding of an emergency, we could build a pseudo query aiming at the specific topic. Consequently, we use the evolutionary manifold-ranking with the initial pseudo queries to model the importance and novelty of the dynamic information.

2.4 The Balanced Coverage of Summary Content

We also found the coverage about summary could not be better resolved. Since documents can be represented as the structure of sub-topics, which helps to understand the topic from different aspects. Considering the limitations of the traditional clustering methods (Boros et al., 2001; Zelnik-Manor and Perona, 2004; Brand and Huang, 2003; Wan and Yang, 2008), therefore we adopt the spectral clustering (von Luxburg, 2007) to partition the subtopics. Spectral clustering works with the structure of eigenvalue and eigenvector of a similarity matrix. Because of space limitation, we cannot introduce the detail of spectral clustering (von Luxburg, 2007).

In this step, we combine the evolutionary manifold-ranking with spectral clustering to design the new redundancy removal algorithm. During the spectral clustering, not all eigenvectors are essential to clustering, therefore we do the eigenvector selection (Zhao et al., 2010).

2.5 Ordering Sub-topics and Selecting Sentences

Based on the results of the evolutionary manifold-ranking score $RScore(x_i)$=$f_i^*(i = 1, 2, ..., n)$, and sub-topics partition $C_1, ..., C_k$, we designed a new optimization algorithm for sentence selection shown in Algorithm 1.

Algorithm 1 Ordering sub-topics and selecting sentences

```
Input: Sub-topics $C_1, ..., C_k$, sentences set $S$=$\{x_i|i = 1, 2, ..., n\}$, and $RScore(x_i)$=$f_i^*(i = 1, 2, ..., n)$;
Output: $GS$;
1: Sort the sub-topics in descending order according to the highest sentence rank score in the corresponding sub-topic;
2: Let $C_i$ be the top 1 sub-topic;
3: For $C_i$, suppose $x_i$ is the highest ranked sentence. Sentence $x_i$ is moved from $S$ to the guided summary $GS$, and then the redundancy penalty is imposed to the overall rank score of each sentence linked with $x_i \in C_i$ as follows: for each sentence $x_j \in C_i$, its rank score $RScore(x_j)$ is computed by equation 1, where $S_{ji}$ is the weight of the similarity matrix through Laplacian transformation and $t > 0$ is the exponent decay factor. The larger $t$ is, the greater penalty is imposed to the overall rank score. If $t = 0$, no diversity penalty is imposed at all;
4: Go to step 1 and iterate until $S = \phi$ or exceed the summary length limit;
```

The Algorithm 1 is based on the idea that extracting the summary sentences from the different subtopics helps to understand the topic from different aspects; the overall rank score of less informative sentences overlapping with the sentences in update summary is decreased.

Here, redundancy removal is also the key step of content selection. The basic redundancy removal method (Zhang et al., 2005) reflects a linear decay of redundancy, which can not express the temporal decay characteristics of redundancy. We think that a news topic usually has a life cycle evolving through birth, growth, decay, and death, reflecting its popularity over time. The importance of sentences changes spontaneously, whose decay could not be simply formalized by linear style. Consequently, we try to explore a new redundancy removal strategy with exponent decay shown in equation (1).

$$RScore(x_j) = RScore(x_j) * (1 - S_{ji})^t \qquad (1)$$

Finally, the sentence with the highest rank score in the most important sub-topic is chosen to produce the summary until satisfying the summary length limit.

If there is no sub-topics partition, we can substitute the C_i in the step 3 for S and penalize less informative ones.

3 Evaluation Results

3.1 Data Set and Evaluation Metrics

3.1.1 Basic Settings

In our experiments, the TAC 2010 corpus is used for parameter tuning. For testing, we took part in the guided summarization track of TAC 2011. There are 48 runs from 22 participants for the guided summarization task, and each participant could submit two systems at best, ranked by priority.

For expert summarization, eight NIST assessors selected and wrote summaries for the 44 topics in the TAC 2011 guided summarization task. Each topic had 2 docsets (A, B), and NIST assessors wrote 4 model summaries for each docset. The NIST human summarizer IDs are A-H.

The participants' summarizer IDs are 3-50. In addition, two baseline runs were included in the evaluation, and their summarizer IDs are 1-2:

Baseline 1 (summarizer ID=1): returns all the leading sentences (up to 100 words) in the most recent document. Baseline 1 provides a lower bound on what can be achieved with a simple fully automatic extractive summarizer.

Baseline 2 (summarizer ID=2): output of MEAD automatic summarizer [1] with all default settings, set to producing 80-word summaries (MEAD selects full sentences from the source text and does not strictly adhere to word limit; the setting of 80 words was necessary to create uncut summaries under the 100-word TAC limit, similarly to Baseline 1).

3.1.2 Evaluation Metrics

The official metrics comprise as follows. NIST evaluated all summaries manually for overall responsiveness and for content according to the Pyramid method (Nenkova et al., 2007). All summaries were also automatically evaluated using ROUGE (Lin and Hovy, 2003)/BE (Hovy et al., 2006)[2]. All summaries were truncated to 100 words before being evaluated.

PYRAMID: It contains six metrics:

(1) average modified score (abbr. AMS);

(2) average numSCUs (abbr. ANSCU);

(3) average numrepetitions (abbr. ANP);

(4) macroaverage modified score with 3 models (abbr. MMS3M);

(5) average linguistic quality (abbr. ALQ);

(6) overall responsiveness (abbr. OR);

(1) (4) are the evaluation metrics of summary content selection. (5) is to assess its readability, (6) is to measure a combination of content and readability.

ROUGE and BE: NIST automatically evaluated all systems using ROUGE and BE, which are evaluation metrics measuring summary content selection. ROUGE-1.5.5 toolkit[3] measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the system summary and the reference summary. BE-1.1 is realized by parsing the evaluated sentences and then using the ROUGE toolkit to compare with word pairs including their dependency relation.

3.2 Our Results and Comparisons

3.2.1 The Results under PYRAMID metrics

There is much inconsistency shown in Table 1. Not all metrics of our systems could rank top 5. System 13 ranks top 3 on OR for time slice A, which shows the better overall responsiveness. System 7 ranks top 2 on ANP for time slice B, which shows the average numrepetitions is relatively high. The ANP score reflects the paraphrase capability of our system besides the number of unique contributors in the peer summary that match an SCU in the model pyramid (means ANSCU).

Seen from Table 1, there is one top 5, and several top 10 for time slice B. The performance of time

[1] http://www.summarization.com/mead/

[2] http://www.nist.gov/tac/

[3] http://haydn.isi.edu/ROUGE/latest.html

Table 1: The Evaluation Results of System 13,7 under PYRAMID metrics

Metrics	Our A		Best A	Model A	Rank	
	Score(ID=13)	Score(ID=7)	Score(ID)	Low,High(ID)	ID=13	ID=7
AMS	0.413	0.392	0.477 (22)		22/50	30/50
ANSCU	5.523	5.091	6.227(22,43)	9.182,11.455(F,G)	14/50	30/50
ANP	1.409	1.841	2(33)		11/50	2/50
MMS3M	0.409	0.387	0.471(22)	0.705,0.888(F,G)	22/50	30/50
ALQ	2.75	2.614	3.75(32)	4.591, 5(F,(D,C,E))	32/50	36/50
OR	3.114	2.773	3.159(25)	4.682,4.955(F,(D,C))	3/50	31/50
Metrics	Our B		Best B	Model B	Rank	
	Score(ID=13)	Score(ID=7)	Score(ID)	Low,High(ID)	ID=13	ID=7
AMS	0.33	0.338	0.353 (9)		10/50	6/50
ANSCU	3.614	3.75	4.023(9)	5.409,8.091(B,D)	11/50	7/50
ANP	0.795	0.841	1(43)		8/50	5/50
MMS3M	0.326	0.333	0.346(12)	0.554,0.823(B,D)	10/50	7/50
ALQ	2.773	2.727	3.455(1)	4.727,5(F,H)	27/50	29/50
OR	2.364	2.477	2.591(35)	4.318,4.909(F,G)	21/50	11/50

slice B is averagely better than the one of A. This shows our evolutionary manifold-ranking is competitive on capturing the novel information. However, the performance of our basic summary is not good enough. We need to further analyze the reasons. Maybe not fully understanding the semantic contained in topic is one reason. Say the constructed pseudo topic description cannot appropriately reflect the query intent.

From the intuitive point of view and the previous work (Boros et al., 2001; Zelnik-Manor and Perona, 2004; Brand and Huang, 2003; Wan and Yang, 2008), clustering should be beneficial to the summary content selection. However, the incremental performance of our spectral clustering on time slice A and B is not consistent, which just shows the better results on B. Therefore, we need to verify the relevant assumptions in the future.

3.2.2 The Results under ROUGE, BE metrics

Table 2 shows the results under ROUGE, BE metrics. Our systems just rank middle among all. It is very surprising that there is great performance difference between ROUGE, BE and PYRAMID metrics. Maybe the basic idea of our method matches the principle of PYRAMID metric well. Therefore, we achieve the better performance under PYRAMID metrics. We observe that the same system shows the different performance under the different metrics and no one system can get top 1 under all metrics. Consequently, we could not directly tell which one is the best or the worst. The more appropriate and objective evaluation method has to be explored further.

4 Conclusion and future work

In this paper, the category oriented extractive content selection method for guided summarization is proposed, which combines the evolutionary manifold-ranking with the spectral clustering based on eigenvector selection to model the important, the novel and the balanced coverage content.

Our contribution mainly includes that (1) the pseudo query is constructed to understand the semantic of emergency for further improving the manifold-ranking; (2) eigenvector selection is done under the process of spectral clustering in order to avoid the eigenvector with less information. Totally, the performance of time slice B is better than that of A, it shows that our evolutionary manifold-ranking is competitive on capturing the novel information.

Our systems rank top 5 under several PYRAMID metrics. However, we rank middle under ROUGE and BE averagely. Our methods show the great difference for different evaluation principles. Therefore, we try to think why this happens? What is the best method? What is the best evaluation principle?

Table 2: The Evaluation Results of System 13,7 under ROUGE,BE metrics

Metrics	Our A		Best A	Model A	Rank	
	Score(ID=13)	Score(ID=7)	Score(ID)	Low,High(ID)	ID=13	ID=7
ROUGE-2	0.10934	0.09687	0.13447(43)	0.1282(D)	16/50	29/50
ROUGE-SU4	0.14340	0.13053	0.16519(43)	0.16412(D)	19/50	30/50
BE	0.06332	0.05707	0.08553(43)	0.09085(D)	29/50	32/50
Metrics	Our B		Best B	Model B	Rank	
	Score(ID=13)	Score(ID=7)	Score(ID)	Low,High(ID)	ID=13	ID=7
ROUGE-2	0.06759	0.06889	0.09589(43)	0.11474(E)	31/50	30/50
ROUGE-SU4	0.10692	0.10753	0.13080(43)	0.14941(E)	30/50	29/50
BE	0.03682	0.04047	0.06480(43)	0.07970(E)	35/50	32/50

We will try to explore the reasons and propose the better resolution. Yet the time is limited, we will supplement the experiments and the relevant analysis in the future.

5 Acknowledge

Thank for the fund support from NSFC No.61100123, and Ministry of Education Fund for the Doctoral.

References

E. Boros, P.B. Kantor, and D.J. Neu. 2001. A Clustering Based Approach to Creating Multi-Document Summaries. In *Proceedings of the SIGIR*.

M. Brand and K. Huang. 2003. A Unifying Theorem for Spectral Embedding and Clustering. In *Proceedings of the Ninth International Workshop on Aritficial Intelligence and Statistics*.

G. Erkan and D.R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

R. He, Y. Liu, B. Qin, T. Liu, and S. Li. 2008. HITIRs update summary at TAC2008: Extractive content selection for language independence. In *Text Analysis Conference*.

E. Hovy, C.Y. Lin, L. Zhou, and J. Fukumoto. 2006. Automated Summarization Evaluation with Basic Elements. In *Proceedings of the LREC*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the NAACL*, pages 71–78.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4–21.

U. von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

X. Wan and J. Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the SIGIR*, pages 299–306.

X. Wan, J. Yang, and J. Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, pages 2903–2908.

L. Zelnik-Manor and P. Perona. 2004. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17(16):1601–1608.

B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.Y. Ma. 2005. Improving web search results using affinity graph. In *Proceedings of the SIGIR*, pages 504–511.

F. Zhao, L. Jiao, H. Liu, X. Gao, and M. Gong. 2010. Spectral clustering with eigenvector selection based on entropy ranking. *Neurocomputing*, 73(10-12):1704–1717.

D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, pages 595–602.

D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf. Ranking on data manifolds. In *Advances in neural information processing systems 16: proceedings of the 2003 conference*, pages 169–176.