WHUSUM Participation at TAC 2011 Guided Summarization Track

Po Hu and Donghong Ji
Computer School, Wuhan University, China
phu@mail.ccnu.edu.cn, donghong_ji2000@yahoo.com.cn

Abstract

In this report, we present details about the participation of WHUSUM in the guided summarization track at TAC 2011. Guided summarization task requires participants to produce short, coherent summaries of news articles with the guidance of predefined categories and aspects for each category. This year, we extended our query-focused update summarization system with aspect related information. In our system, a graph-based ranking approach was adopted to facilitate the extraction of aspect-related sentences with diversity and novelty.

1 Introduction

Similar to that in TAC 2010, the guided summarization task of TAC 2011 is 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. All participating systems are given a list of important aspects for each category, and the generated summaries must cover all these aspects or other information relevant to the topic. Additionally, an "update" component of the guided summarization task is to write a 100-word "update" summary of the subsequent 10 newswire articles for the topic. The summary for the subsequent documents should be created under the assumption that the user has already known the contents in the earlier batch of documents and should avoid repeating old information and inform the user of new information about the topic.

In contrast to previous query focused summarization tasks, the guided summarization task is guided by a list of important aspects whose information should be contained in the generated summary. The documents for this year's summarization task come from the newswire portion of the TAC 2010 KBP Source Data. Each topic has 20 relevant documents which have been divided into 2 sets: Document Set A and Document Set B. Each set has 10 documents, and all the documents in Set A chronologically precede the documents in Set B. No topic narrative is provided since the category and its aspects have clearly defined what information the user is looking for.

Given a topic, the task is to write 2 summaries (one for Document Set A and one for Document Set B) according to the list of aspects given for the topic category. The summary for Document Set A should be a straightforward aspect-focused summary, and the update summary for Document Set B is also aspect-focused but should be produced with the requirement for non-redundancy taking priority over the requirement to address all category aspects.

The task naturally led to the consideration of combining the aspect related information with existing query-focused summarization methods since the information extracted from all the aspects for a category can be regarded as a pseudo-query for the documents belonging to this category. We submitted two runs altogether: the first uses both topic title and aspect-related words for this topical category as query words to guide the summarization process, and the second ignores the aspect information and uses only topic title as query terms. In the following sections we will describe our runs and discuss the results attained.

The rest of the paper is organized as follows. In Section 2 we give a detail description of the WHUSUM system. Section 3 presents the evaluation results. Finally, we conclude in Section 4.

2 Our System

WHUSUM 2011 retains a similar structure to previous version with the major change made for this year's task: aspect-related sentence selection.

1. Data preprocessing
2. Aspect-related sentence selection
3. Graph-based sentence ranking
4. (For update summaries) post-processing for redundancy removal and novel sentence's extraction.

In preprocessing component, all documents are segmented into sentences and stop-words are removed. The remaining words are stemmed by Porter Stemmer.

In the aspect-related sentence selection process, we need to retrieve out the aspect-related information from a given document set based on the aspect-related terms. The defined categories and its aspects are the following:

1. Accidents and Natural Disasters:
   1.1 WHAT: what happened
   1.2 WHEN: date, time, other temporal placement markers
   1.3 WHERE: physical location
   1.4 WHY: reasons for accident/disaster
   1.5 WHO_AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the accident/disaster
   1.6 DAMAGES: damages caused by the accident/disaster
   1.7 COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the accident/disaster
2. Attacks (Criminal/Terrorist):
2.1 WHAT: what happened
   2.2 WHEN: date, time, other temporal placement markers
   2.3 WHERE: physical location
   2.4 PERPETRATORS: individuals or groups responsible for the attack
   2.5 WHY: reasons for the attack
   2.6 WHO_AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the attack
   2.7 DAMAGES: damages caused by the attack
   2.8 COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the attack (e.g. police investigations)
3. Health and Safety:
   3.1 WHAT: what is the issue
   3.2 WHO_AFFECTED: who is affected by the health/safety issue
   3.3 HOW: how they are affected
   3.4 WHY: why the health/safety issue occurs
   3.5 COUNTERMEASURES: countermeasures, prevention efforts
4. Endangered Resources:
   4.1 WHAT: description of resource
   4.2 IMPORTANCE: importance of resource
   4.3 THREATS: threats to the resource
   4.4 COUNTERMEASURES: countermeasures, prevention efforts
5. Investigations and Trials (Criminal/Legal/Other):
   5.1 WHO: who is a defendant or under investigation
   5.2 WHO_INV: who is investigating, prosecuting, or judging
   5.3 WHY: general reasons for the investigation/trial
   5.4 CHARGES: specific charges to the defendant

5.5 PLEAD: defendant's reaction to charges, including admission of guilt, denial of charges, or explanations
5.6 SENTENCE: sentence or other consequences to defendant

We begin with the terms in the topic title as our query terms. Then, they are expanded with category terms and aspect terms where appropriate. For instance, as topic "D1101A" belongs to the second category called “Attacks (Criminal/Terrorist)”, we select both the topic title “Amish Shooting” and the topical related aspect terms as our query terms, which is shown as follows: “Amish Shooting attack criminal terrorist happen perpetrator reason casualty death injury affect damage cause countermeasure rescue efforts prevention effort investigation”. Then, the co-training algorithm (Blum and Mitchell, 1998) used in our previous system (Hu and Ji, 2009) is adopted to choose aspect-related sentences from two abundant views, which can incorporate multi-dimensional complementary information in the process and leverage both the individual information in each sentence and the relationship information among sentences. After the procedure, we can select a number of aspect-related sentences from documents and consider them as the candidates for summary sentence extraction.

To extract a certain sentence set from the candidates into the summary, we adopts a graph-based sentence ranking algorithm (Zhu et al., 2007) to achieve both informativeness and diversity in a unified framework. The underlying idea is that the items and inter-item relationships can be encoded by a graph. A random walk can be defined on the graph correspondingly and the importance of an item can be determined by stationary distribution of random walk. If a node is most similar to many other nodes, it will first become a highly ranked one and at the same time be adjusted into the absorbing state, which will cut down the significance of similar unranked nodes and encourage diversity. In this component, sentence ranking is carried out by the following procedure.

1) Construct an undirected affinity graph G_r over the candidate aspect-related sentence set, where each sentence is considered as a node and edges are created between two sentences if their pair-wise similarity exceeds 0.01.

2) Define an adjacency matrix M_r to represent G_r with each entry corresponding to the cosine similarity of two corresponding sentence vectors.

3) Normalize matrix M_r to matrix $\widetilde{M_r}$ by dividing each element in M_r by the corresponding row sum.

4) Use $\widetilde{M_r}$ to form a stochastic matrix M_s by integrating a prior ranking distribution r on these sentences according to formula $M_s = \lambda \widetilde{M_r} + (1-\lambda) 1 r^T$.

5) Compute M_s’s stationary distribution and take the sentence with the largest stationary probability to be the top one for the final ranking.

6) Turn ranked sentences into absorbing states and compute the expected number of visits for all the rest sentences. Then pick the next higher ranked sentence with the maximum expected number of visits. Repeat step 6 until all the sentences are ranked.

In the above procedure, M_s can be considered as the transition matrix of a Markov chain with the entry $M_s(i,j)$ specifying the transition probability from state i (i.e. sentence s_i) to state j (i.e. sentence s_j) in the corresponding Markov chain. $\lambda \in [0,1]$ is a damping factor, 1 is an all-1 vector, and $1r^T$ denotes the prior ranking that is represented as a probability distribution. The teleporting random walks based on M_s act in such a way that moving to an adjacent state according to the entry in $\widetilde{M_r}$ with probability λ or jumping to a random state according to the prior ranking distribution with probability $1-\lambda$ at each step.

In this way, we can produce the aspect-focused summary for Document Set A with both informativeness and diversity in accordance with the length limit. For Document Set B, the summary should also capture the novel information and avoid redundant information from Document Set A, so a

post-processing module used in our previous system is used to re-rank aspect-related sentences selected from Document Set B by setting the parameter λ 0.95 to avoid possible redundant information and encourage diversified novelty.

3 Evaluation Results

The test dataset used in TAC 2011 guided summarization task is composed of 44 topics, divided into five categories: Accidents and Natural Disasters, Attacks, Health and Safety, Endangered Resources, Investigations and Trials. Each topic has a topic ID, category, title, and 2 docsets (A, B). The category and its aspects define what information the reader is looking for. For each docset, different NIST assessors wrote 4 100-word model summaries covering all the aspects listed for the topic category.

This year, 48 runs from 22 participants were submitted for the guided summarization task. The participants each submitted up to two runs, and their summarizer IDs are 3-50. Our submitted system IDs are 10 and 18. All summaries were truncated to 100 words before being evaluated. NIST evaluated all summaries manually for overall responsiveness and for content according to the Pyramid method. All summaries were also automatically evaluated using ROUGE/BE. ROUGE-2 and ROUGE-SU4 scores were computed by running ROUGE-1.5.5 with stemming but no removal of stopwords.

In table 1 and 2, we show the automatic evaluation results for our submitted runs on Document Set A and B.

Table 1: automatic evaluation results of our submitted runs on Document Set A

peerID	**average ROUGE-2 recall** [0, 0.13447]	**average ROUGE-SU4 recall** [0, 0.16519]	**average BE recall** [0, 0.08553]
10	0.10937	0.14564	0.06969
18	0.10698	0.14297	0.06850

Table 2: automatic evaluation results of our submitted runs on Document Set B

peerID	**average ROUGE-2 recall** [0.03561, 0.09589]	**average ROUGE-SU4 recall** [0.07222, 0.13086]	**average BE recall** [0.01271, 0.06480]
10	0.08083	0.11958	0.05127
18	0.07768	0.11813	0.05091

In table 3 and 4, we show the manual evaluation results for our submitted runs on Document Set A and B.

Table 3: manual evaluation results of our submitted runs on Document Set A

peerID	**average modified (pyramid) score** [0, 0.477]	**average linguistic quality** [1, 3.75]	**average overall responsiveness** [1, 3.159]
10	0.435	2.705	2.864
18	0.436	2.591	2.886

Table 4: manual evaluation results of our submitted runs on Document Set B

peerID	**average modified (pyramid) score** [0.115, 0.353]	**average linguistic quality** [1.477, 3.455]	**average overall responsiveness** [1.614,2.591]
10	0.300	2.659	2.205
18	0.303	2.568	2.386

In table 5 and 6, we show the manual evaluation results for our submitted runs on Document Set A and B averaged over all topics in a given category.

Table 5: manual evaluation results of our submitted runs on Document Set A averaged over all topics in a given category

peerID	categoryID	average modified (pyramid) score	average linguistic quality	average overall responsiveness
10	1	0.53	2.778	3
	2	0.478	2.667	3.111
	3	0.372	2.8	2.8
	4	0.342	2.625	2.5
	5	0.448	2.625	2.875
18	1	0.587	2.778	2.889
	2	0.474	2.667	3.444
	3	0.356	2.5	2.7
	4	0.333	2.5	2.5
	5	0.429	2.5	2.875

Table 6: manual evaluation results of our submitted runs on Document Set B averaged over all topics in a given category

peerID	categoryID	average modified (pyramid) score	average linguistic quality	average overall responsiveness
10	1	0.314	2.778	2
	2	0.289	2.667	1.556
	3	0.243	2.8	2.4
	4	0.357	2.5	2.625
	5	0.311	2.5	2.5
18	1	0.28	2.667	2.222
	2	0.254	2.556	2.111
	3	0.21	2.6	2.2
	4	0.362	2.375	2.5
	5	0.438	2.625	3

The official evaluation results presented in the above tables show that our systems got competitive performance in the guided summarization task on summary content's evaluation. The linguistic quality of our system's output was not good because we didn't take it into account in the current system's configuration yet, which will be considered in our future work.

Besides, from the experimental result, we found that the performance between our submission expanded with aspect terms and the submission without using them were not significant. The reason could be that we add all the aspect terms for a category to expand the query directly, but many aspect terms can not be directly matched with the entities appeared in documents, more advanced information extraction technique should be adopted to improve the aspect-related information discovery such as named entity extraction, relationship discrimination and event extraction.

4 Conclusion

This paper presented our participation in the guided summarization task of TAC 2011. We try incorporating the aspect related information into the sentence selection process. The official evaluation result shows that there is large room to improve our system by conducting a deeper semantic analysis to select important concepts and find aspect-related information. We also plan to use a combination of

searching, dictionaries and knowledge base to populate each of the aspects for each category and use temporal analysis technique to improve the time-related aspect for improving the update information identification.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 90820005 and No. 61070082).

References

Blum, A. and T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT'98)*, pp.92-100.

Zhu, X.J., A. Goldberg, J.V. Gael and D. Andrzejewski. 2007. Improving Diversity in Ranking Using Absorbing Random Walks. *Proceedings of HLT-NAACL'07*, pp.97–104.

Hu, P., and Ji. D.H. 2009. WHUSUM: Wuhan University at the Update Summarization Task of TAC 2009. *Proceedings of the 2009 Text Analysis Conference (TAC 2009)*.