WHUSUM: Wuhan University at the Update Summarization Task of TAC 2009

Po Hu and Donghong Ji
Computer School, Wuhan University, Wuhan 430079, China
phu@mail.ccnu.edu.cn, donghong_ji2000@yahoo.com.cn

Abstract

This paper describes the system WHUSUM we developed to participate in the update summarization task of TAC 2009. Given a topic and corresponding topic statement, this year's task is to write 2 summaries (one for Document Set A and one for Document Set B) that meet the information need expressed in the topic statement. In order to generate a topic-oriented summary for Set A, We present a co-training based strategy to select the topic relevant sentences from two abundant views and adopt a graph-based ranking algorithm (i.e. GRASSHOPPER) to achieve both information richness and content diversity in the generated summary. Furthermore, to capture the novel information in Set B and remove the possible redundant information in historical Document Set A, we propose two approaches to encourage novelty. One is to incorporate similarity between sentences in historical set and current set in the prior ranking of GRASSHOPPER. Another is to directly rank sentences for Document Set B first, and then to adjust their ranking scores based on the content comparison between the relevant sentence sets in A and B. The official evaluation results show that our system gets competitive performance in general topic-oriented summarization task and ranks in the middle among 52 submitted systems in update summarization task, which demonstrate that there is still large room to improve the novelty detection mechanism of the system.

1 Introduction

The Text Analysis Conference (TAC) is one of the most well-known series of workshops that provides the infrastructure necessary for large-scale evaluation of natural language processing methodologies and technologies. This is the first time that we attended the TAC evaluations and we participated in the update task of summarization track. The update summarization task of TAC 2009 is similar to that in TAC 2008, which aims to generate two short and fluent summaries respectively for two chronologically ordered document sets to meet the topic-relevant information need. The summary of the second document set should be written under the assumption that the user has already read the earlier documents and should avoid repeating old information and inform the user of novel information about the specific topic.

Given a topic, corresponding topic statement and two document sets (i.e. Document Set A and Document Set B) with all the documents in Set A chronologically preceding the documents in Set B, the update summarization task in TAC 2009 can be divided into two subtasks. Task 1 is a topic-focused multi-document summarization task for Document Set A, and task 2 is a novelty-oriented and topic-focused multi-document summarization task for Document Set B. The design motivation behind our system WHUSUM is that a good topic-focused multi-document summarizer should preserve the information biased to the topic description as much as possible, and remain the richest and diverse information at the same time. In addition, the novelty-oriented update summarizer should put more novel contents in the generated summary with least duplicate information from the historical documents.

In our system, we use a co-training based strategy to better select the topic relevant sentences from different angles and adopt a graph-based ranking algorithm to encourage both information richness and content diversity in a unified framework. We also try two approaches to detect novelty based on the content similarity comparison between the historical set and current set.

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 for TAC 2009

2.1 Overview

WHUSUM system uses sentence extraction strategy for this year's task. In this system, informative sentences with the characteristics of topic relevance, information richness and content diversity or content novelty are automatically extracted from the document set and are concatenated to form a summary. The whole system consists of four major modules: preprocessing, topic-relevant sentence selection, sentence ranking and post-processing. In preprocessing module, topic description and all documents are segmented into sentences, and the headlines of documents as well as stop-words are removed. The remaining words are stemmed by Porter Stemmer. The function of the second module in WHUSUM system is to select a small number of sentences from the original sentence space that can meet the needs in topic description. Sentence ranking module will evaluate each topic-relevant sentence and ranking all of them by combining information richness with content diversity. For summarizing Document Set A, the first three modules are employed. However, to summarizing Document Set B and avoid redundant information from historical document Set A, a post-processing module is also adopted to re-rank topic-relevant sentences in Set B.

2.2 Topic-Relevant Sentence Selection

One of the critical components for a topic-focused multi-document summarizer is to retrieve out the topic-related information from a document set. Many methods incorporate the topic information into generic summarizers by only computing the similarity value between each sentence and the topic description (Saggion et al., 2003). More related work can be found on TAC publications (Hoa Trang Dang and Karolina Owczarzak, 2008). In our system, the co-training algorithm (Blum and Mitchell, 1998) is adopted to choose topic-relevant sentences from two abundant views, which can incorporate multi-dimensional complementary information in the process.

Co-training can make use of a large number of unlabeled examples to boost learning performance. It has already been successfully used in many natural language processing applications (Muller et al., 2002; Sarkar, 2001; Wong et al., 2008). To our knowledge, there is little research in applying co-training algorithm to topic-relevant sentence selection especially when labeled relevant and irrelevant sentences are absent.

Two different views X and Y are investigated in our system, which aim to leverage both the individual information in each sentence and the relationship information among sentences. Here X represents the content view which uses content bearing terms to describe a sentence, and Y is the relationship view which represents a sentence by its pair-wise similarity with other sentences. Matrix $[M_{ij}]_{n*m}$ is used to describe the sentence set that is formally represented on X with each entry M_{ij} corresponding to the weight associated with term t_j in sentence s_i, which is calculated by the $TF_{ij}*ISF_i$ formula, where n is the total number of sentences including the topic description, m is the total number of terms in the documents, TF_{ij} denotes the frequency of term t_j appearing in sentence s_i, and ISF_i is the inverse sentence frequency of term t_j, which is calculated by $1+\log(n/n_j)$, where n_j is the number of the sentences that contain term t_j. Matrix $[N_{ij}]_{n*n}$ is used to describe the sentences on view Y with each entry N_{ij} corresponding to the pair-wise cosine similarity between sentence s_i and s_j. The co-training based sentence selection is carried out by the following procedure.

1) Sort the sentences in the whole document set in the descending order of relevance to the topic description.

2) Choose 25 sentences respectively with the highest and lowest relevance scores with the topic as the pseudo-labeled relevant and irrelevant sentences.

3) Create an unlabeled sentence pool U' by selecting 75 sentences from the unlabeled sentences U at random.

4) Loop while there are still some unlabeled sentences in U

Use pseudo-labeled sentences to train a C4.5 decision tree classifier C_x on $[M_{ij}]_{n*m}$.

Use pseudo-labeled sentences to train a C4.5 decision tree classifier C_y on $[N_{ij}]_{n*n}$.

Use C_x to label 1 topic-relevant sentence and 1 topic-irrelevant sentence with the highest classifying confidence from U'.

Use C_y to label 1 topic-relevant sentence and 1 topic-irrelevant sentence with the highest classifying confidence from U'.

Add these labeled sentences to the set of pseudo-labeled sentences and remove them from U.

Randomly choose 4 sentences from U to replenish U'.

After the procedure, the module can automatically select a number of topic-focused sentences with the automatically inferred topic-relevant label.

2.3 Sentence Ranking

This module adopts a graph-based sentence ranking algorithm (i.e. GRASSHOPPER) (Zhu et al., 2007) to achieve both information richness and content diversity in a unified framework.

The underlying idea of GRASSHOPPER 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 module, sentence ranking is carried out by the following procedure.

1) Construct an undirected affinity graph G_r over the topic-relevant sentences that have been selected by the second module, 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 1.

$$M_s = \lambda \widetilde{M_r} + (1-\lambda) 1 r^T \quad (1)$$

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.

5) Compute M_s's stationary distribution and take the sentence (i.e. state) 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. An illustration that the Markov chain with the stochastic matrix M_s will converge to a unique stationary

distribution and the detailed description about how to compute the expected number of visits in an absorbed Markov chain can be found at the reference paper written by Xiaojin Zhu et al., 2007. Repeat step 6 until all the topic-relevant sentences are ranked.

In WHUSUM system, to generate a topic-oriented summary for Document Set A, the first three modules mentioned above are employed to achieve both information richness and content diversity in the generated summary in accordance with the length limit.

2.4 Post-Processing

The summary for Document Set B is also topic-focused, so we can still use the first three modules in WHUSUM to generate the summary. However, the summary for Set B should also capture the novel information and avoid redundant information from Set A, so a post-processing module is added in our system to rank or re-rank topic-relevant sentences selected from Set B. Two approaches are tried in our system. Method 1 is to first compute the similarity of the selected topic relevant sentences between set A and Set B, and then incorporate the normalized average similarity value for those relevant sentences of Set B in the prior ranking $1r^T$ of GRASSHOPPER. Method 2 is to directly rank selected sentences of Set B using GRASSHOPPER algorithm first, and then to adjust their ranking scores based on the content similarity between the relevant sentences in A and B. The intuition of these methods is that the topic-related sentences in B with higher similarity to the topic-related sentences in A should have lower ranking scores. We have submitted two runs with different post-processing modules described here. Run 1 adopted method 1 and the parameter λ in formula 1 was set to 0.6, and run 2 used method 2 and the parameter λ in formula 1 was set to 0.9, and let the prior ranking the uniform probability distribution vector.

3 Evaluation

TAC 2009 provides 44 topics for evaluation. Each topic includes a topic statement and 20 relevant documents which have been divided into 2 sets: Document Set A and Document Set B. Each document set has 10 documents, and all the documents in Set A chronologically precede the documents in Set B. NIST assessors wrote 4 model summaries for each document set. 52 runs from 27 participants are submitted for the update summarization task. All submitted runs are evaluated manually and automatically, including overall responsiveness, pyramid-based evaluation of content, linguistic quality, ROUGE-2, ROUGE-SU4 (Lin and Hovy, 2003) and Basic Elements based evaluation. Our submitted system ID is 55 and 32.

3.1 Manual Evaluation

In table 1, we show the manual evaluation result for our submitted summarization result with system ID 55 on Document Set A and B.

Document Set	average modified pyramid score	average linguistic quality	average overall responsiveness
A	0.339 (3/52) [0.062, 0.383]	4.364 (39/52) [3.432, 5.932]	4.432 (16/52) [2.455, 5.159]
B	0.207 (22/52) [0.05, 0.307]	4.386 (39/52) [3.364, 5.886]	3.523 (37/52) [2.227, 5.023]

Table 1: TAC Manual evaluation result on Document Set A and B over 44 topics

3.2 Automatic Evaluation

In table 2, we show the automatic evaluation result for our submitted summarization result with system ID 55 on Document Set A and B.

Document Set	ROUGE-2	ROUGE-SU4	average Basic Elements recall
A	0.10167 (10/52) [0.02832, 0.12184]	0.13776 (11/52) [0.05925, 0.15131]	0.05304 (11/52) [0.00937, 0.06379]
B	0.07397 (28/52) [0.02625, 0.10417]	0.11694 (25/52) [0.05740, 0.13959]	0.03671 (29/52) [0.00851, 0.06364]

Table 2: TAC Automatic evaluation result on Document Set A and B over 44 topics

The difference between our submitted two runs lies in adopting different post-processing modules. Their automatic evaluation results are shown in Table 3. System 55 used the method 2 to rerank topic-relevant sentences and system 32 used the method 1, which methods have been mentioned in Section 2.4.

System ID	ROUGE-2	ROUGE-SU4	average Basic Elements recall
55	0.07397	0.11694	0.03671
32	0.06261	0.10426	0.02833

Table 3: Our submitted system's automatic evaluation results on Document Set B over 44 topics

3.3 Analysis

The official evaluation results presented in the above tables show that our system gets competitive performance in general topic-oriented summarization task and ranks in the middle among 52 submitted systems in update summarization task, which verify the potential of the integration of co-training based learning algorithm from two abundant views and graph-based ranking algorithm in topic-focused summarization. What's more, in post-processing module of our system for update summarization task, method 2 is more effective than method 1 in reducing redundancy and detecting novelty.

4 Conclusion and Future Work

This paper presented our participation in the update summarization task of TAC 2009 summarization track. In our WHUSUM system for topic-focused summarization task, we presents a co-training based strategy to select the topic relevant sentences and adopt GRASSHOPPER algorithm to ranking sentences in term of information richness and content diversity. This approach got encouraging performance according to the official evaluation results. For topic-focused update summarization task, we try incorporating the similarity knowledge of the selected topic relevant sentences between historical set A and current Set B into the ranking process for the sentences in B. However, the corresponding evaluation result shows that there is still large room to improve the novel information detection mechanism of our system. In addition, our system's evaluation result for average linguistic quality is not good, so we plan to improve it by using sophisticated natural language processing techniques in the future.

Acknowledgements

This work was supported by the Major Research Plan of National Natural Science Foundation of China (90820005), National Natural Science Foundation of China (60773011) and Wuhan University 985 Project (985yk004).

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.

Christoph, M., S. Rapp and M. Strube. 2002. Applying Co-Training to Reference Resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp.352–359.

Hoa T. D. and K. Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task. *Proceedings of TAC 2008*.

Lin C.Y. and H. Eduard. 2003. Automatic Evaluation of Summaries Using N-Gram Co-Occurrence Statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'03)*, pp.71–78.

Saggion, H. K. Bontcheva and H. Cunningham. 2003. Robust generic and query-based summarization. *Proceedings of EACL'2003*.

Sarkar, A. 2001. Applying Co-Training Methods to Statistical Parsing. *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'01)*, pp.175–182.

Wong, K.F., M.L. Wu and W.J. Li. 2008. Extractive Summarization Using Supervised and Semi-Supervised Learning. *Proceedings of the International Conference on Computational Linguistics (COLING'08)*, pp.985-992.

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.