TAC 2011 Guided Summarization of ICL

Sujian Li, Tao Song, Xun Wang

Key Laboratory of Computational Linguistics, Ministry of Education, Peking University

{lisujian, songtao, wangxun}@pku.edu.cn

Abstract

For the update summarization task of TAC 2011, we submitted two runs applying a improved graph-based sentence ranking method. The difference between these two runs is that the second one aims to acquire the category words and use them for extracting information-rich sentences. For the update summarization task, we adopt similar methods used in previous evaluations and simultaneously penalize the information overlap between docset *B* and docset *A*.

1. Introduction

The TAC 2011 and 2010 guided summarization task is not quite the same to that in TAC 2009 and before (Hoa 2009, Owczarzak 2010). Although they both aim at generating short (no more than 100 words) fluent multi-document summaries of news articles with or without some related earlier articles considered, TAC 2010 and 2011 summarization tasks aim to make a deeper linguistic (semantic) analysis of documents. TAC 2011 clearly presents redundancy and non-readability problem in multi-document summarization. Same as 2010, in the TAC 2011 summarization task, the source documents are all from five categories and each category is predefined by a list of important aspects. Good summaries are expected to cover all these aspects. In this task, we still adopt the sentence extraction framework, where both a graph based method inspired by the manifold ranking method and the category information are used to extract important sentences. When generating update summaries for document set *B*, sentences are penalized by their content overlap with document set *A*.

The rest of the paper is organized as follows. Section 2 briefly describes our method adopted. Section 3 presents our evaluation results in TAC 2011. Section 4 shows the future work and concludes the paper.

2. Our Method

Inspired by manifold learning method, we present a damped propagation model with the assumption that the query relevance score of the label query should keep its invariance.

Table 1. Our damped propagation model

Algorithm 1: BasicPropagationRank

Input: The document set $\{s_0, s_1, s_2, \cdots s_n\}$, s_0 is the query, d is damping factor($0 \leq d \leq 1$)

Output: The query relevance score vector f

BEGIN

(1) k=0, Initialize f: $f_0^{(0)} = f_0 = 1$, $f_u^{(0)} = (0,0,\cdots,0)^{\mathrm{T}}$;

(2) Update the sentence scores by calculating: $f^{(k+1)} = dTf^{(k)}$;

(3) k=k+1;

(4) Normalize $f^{(k)}$;

(5) Restore the query's score $f_0^{(k)}$ to 1, go to step (2) until the score vector

converge[1];

END

We have proved the convergence of this damped propagation model, which can be seen the generalization of the manifold learning model. Also we experiment this model on previous evaluation data and found that it is more stable than the manifold learning algorithm.

How to efficiently use the aspect information is a problem. In TAC 2010, we design a method which collects aspect information for each sentence and the results are not satisfying. In this evaluation, we change the strategy and first acquire noteworthy category words since we assume these category words reflect the main features of each category and must contain the aspect information. We exemplified some category words as in the following table:

Accidents & Natural Disasters	survivor, dead, danger, traffic, blast
Attacks	foreign, muslim, fight, injur, dismiss, shoot, mass, battl
Health & Safety	Health, diseas, patient, service, suffer, hospit
Endangered Resources	Wildlife, biologist, extinct, agricultur, environmentalist,
Investigations & Trials	Lawyer, accus, charg, defend, comment, wit, alleg

Table 2. Extracted category words*(stemmed)*

To extract category words, we adopt the entity-aspect model as Figure 1, which clusters all the words as background words, aspect words, document words [Li 2010].

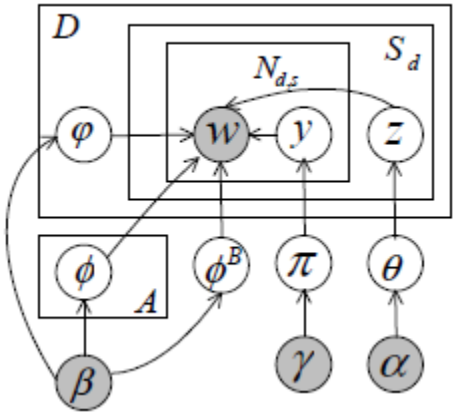

Figure 1: the Entity-aspect model.

We use the entity-aspect model to model each category. With the Gibbs sampling techniques, we can finally get background words, document words, and aspect words for each category. Since background words present some common information of each category, we plan to extract category words through comparing background words across different categories, with the assumption that real category words should rank top in some category background words, but lower in other category. That is, the ranking variance of background words is the foundation to extract category words.

After running the entity-aspect model, we get V background words labeled as $w_1, w_2, \ldots, w_V$ with distribution rank in every category. We note the rank of word w_i in category j as $Rank_j(w_i)$ and then the average rank of word w_i in all categories as $\overline{rank}(w_i)$. We calculate a word's global ranking variance in all categories using the following formula:

[1] In our experiment, if $|f_i^{(k)} - f_i^{(k+1)}| \le 0.000001, 1 \le i \le n$,then the iteration process is stopped.

$$Var(w_i) = \frac{\sum_j (Rank_j(w_i) - \overline{rank}(w_i))^2}{M}$$

$$\overline{rank}(P_i) = \frac{\sum_j Rank_j(w_i)}{M}$$

Furthermore, if a word belongs to a certain category, it will have less variance in the other categories. Then the local ranking variance except some category *j* can be calculated as follows:

$$Var_j^{'}(w_i) = \frac{\sum_{k \neq j} (Rank_k(w_i) - \overline{rank}_{\neq j}(w_i))^2}{M-1}$$

$$\overline{rank}_{\neq j}(w_i) = \frac{\sum_{k \neq j} Rank_k(w_i)}{M-1}$$

Finally, with the experience that if a word belongs to the category word of category *j*, it should have a larger global ranking variance and a smaller local variance, we score each word in each category as:

$$Score_j(w_i) = \frac{Var(w_i)}{Var_j^{'}(w_i)}$$

The category words with their scores are then used to calculate the sentence similarity for constructing the sentence graph in the damped propagation model.

3. Experiment and Evaluation

TAC 2011 guided summarization task test datasets comprises of 44 topics. Each topic belongs to a predefined category and has 20 relevant documents which have been divided evenly into 2 docsets (*A*, *B*). NIST assessors wrote 4 model summaries for each document set. All submitted systems are evaluated manually for overall responsiveness and for content according to the Pyramid method. All summaries are also automatically evaluated using ROUGE-2, Rouge-SU4 and BE metrics. We submitted two runs: Run1 adopt the damped propagation model with tf*isf information for graph construction and based on Run 1, Run2 combines the information of category words.

Table 3 illustrates the automatic evaluation results of our systems. The organizer provides Baseline 1 and Baseline 2 (named BASE1 and BASE2 respectively), where BASE1 returns all the leading sentences (up to 100 words) in the most recent document and BASE2 is the output of MEAD automatic summarizer with all default settings. We also list the best peer result as TOP1. The manual evaluation results are listed in Table 4. The suffix "_A" and "_B" represent summarizing for docset A and docset B respectively. The integers in the bracket denote the rank of the corresponding summarizers.

	R-2	R-SU4	BE
TOP1_A	0. 13440 (0. 11940– 0. 14915)	0. 16519 (0. 15090- 0. 17921)	0. 08565 (0.07193 - 0. 10036)
BASE1_A	0. 06410 (0. 05230– 0. 07788) (45)	0. 09934 (0. 08872- 0. 11081) (45)	0. 03403 (0. 02489- 0. 04602) (44)
BASE2_A	0. 08682 (0. 07329– 0. 09973) (32)	0. 11749 (0. 10543- 0. 12950) (36)	0. 05741 (0. 04590- 0. 06925) (31)
Run1_A	0. 11302 (0. 09854– 0. 12641) (**12**)	0. 14873 (0. 13598- 0.12716) (**10**)	0. 07481 (0. 06243- 0. 08738) (**6**)
Run2_A	0. 10838 (0. 09568– 0. 12026) (**19**)	0. 14142 (0. 12960- 0. 15177) (**21**)	0. 06949 (0. 05908- 0. 08055) (**19**)

Table 3: Automatic Evaluation for Docset A in TAC 2011

	Pyramid	Ling. quality	Resp

TOP1_A	0.477	3.75	3.159
BASE1_A	0.304 (45)	3.205 (7)	2.5(37)
BASE2_A	0.362(32)	2.818(30)	2.841(27)
Run1_A	0.431 (16)	2.5(40)	2.864(24)
Run2_A	0.428(18)	2.591(38)	2.909 (20)

Table 4: Manual Evaluation for Docset A in TAC 2011

From Table 3 and Table 4, we can see that our systems perform better in automatic evaluation than manual evaluation. The Pyramid evaluation also demonstrates that our extracted sentences contain enough information. However, since we still adopt the extractive framework and do not conduct post-processing for the extracted sentences, the linguistic quality performs worse. In future work, we should pay more attention on how to fuse the extracted sentences.

Table 5 illustrates the manual evaluation of docset A for each category. Our system Run2 performs well on category 1 (accident & natural disaster) and category 3 (health and safety). Obviously the category words extracted for category 3 are helpful to extracting appropriate sentences. However, the category words for category 2 and 5 do not work for promoting the system performance and even debase the performance of category 4.

CATE	1	2	3	4	5
TOP1_A	0.631	0.58	0.372	0.405	0.57
BASE1_A	0.437 (36)	0.353 (41)	0.226 (42)	0.223 (46)	0.279 (44)
BASE2_A	0.492 (28)	0.401 (36)	0.309 (23)	0.245 (38)	0.357 (38)
Run1_A	0.607 **(2)**	0.492 (14)	0.321 (18)	0.329 (15)	0.403 (28)
Run2_A	0.631 **(1)**	0.493 (13)	0.365 **(4)**	0.229 (44)	0.403 (29)

Table 5: Modified (pyramid) Score of 5 Categories

Table 6 and Table 7 demonstrate the performance of docset B. Overall, the summarization results of Docset B do not perform well as those of Docset A. The main reason is that the new updating information of the topic is not efficiently identified so that there exist much redundancy in the update summaries.

	R-2	R-SU4	BE
TOP1_B	0. 09581 (0. 08429- 0. 10860)	0.12006 (0.11176 - 0.12855)	0. 06473 (0. 05377- 0. 07701)
BASE1_ B	0. 05685 (0. 04769- 0. 06680) (43)	0. 09449 (0. 08637- 0. 10289) (44)	0. 03483 (0. 02653- 0. 04382) (38)
BASE2_ B	0. 05903 (0. 05037- 0. 06781) (39)	0. 09132 (0. 08444- 0. 09850)) (45)	0. 03704 (0. 03034- 0. 04364) (34)
Run1_B	0. 07467 (0. 06657- 0. 08279) (**21**)	0. 11281 (0. 10552- 0. 12063) (**28**)	0. 04978 (0. 04135- 0. 05852) (**15**)
Run2_B	0. 06335 (0.05821 - 0. 08262) (**27**)	0. 11306 (0. 10408- 0. 12244) (**27**)	0. 04695 (0. 03787- 0. 05692) (**20**)

Table 6: Automatic Evaluation for Docset B

	Pyramid	Ling. quality	Resp
Top 1_B	0.353	3.455	2.591
BASE1_A	0.237(36)	3.455(1)	2.091 (37)
BASE2_B	0.284(24)	2.841 (23)	2.114 (35)
Run1_B	0.273(28)	2.5(42)	2.25(29)
Run2_B	0.323(13)	2.614 (37)	2.295 (26)

Table 7: Manual Evaluation for Docset B

4. Conclusion and Future Work

In this paper, we still use the graph based model to extract summary sentences and

achieve an unsatisfactory result. The extracted sentences indeed contain important information and the Rouge and Pyramid metrics prove this. Now we can see that the extractive framework of summarization task seems to reach the bottleneck. And how to improve the linguistic quality of summaries becomes the key problem. In our future work, we will put more emphasis on how to fuse and organize the important sentences into a summary. Also, how to effectively make use of aspects information will be another consideration.

Acknowledgements

This work is supported by NSFC programs (No: 60875042 and 90920011), National Social Science Foundation (No: 10CYY023) and National Key Technology R&D Program (No: 2011BAH10B04-03)

Reference

Hoa T. D. and Owczarzak K.. 2009. Overview of the TAC 2009 Summarization Task. Text Analysis Conference 2009 proceedings, http://www.nist.gov/tac/, 2009.

Owczarzak K. and Hoa T. D., 2010. Overview of the TAC 2010 Summarization Track: Guided Task and AESOP Task. Text Analysis Conference 2009 proceedings, http://www.nist.gov/tac/, 2010.

Li P., Jiang J., Wang Y.. Generating Templates of Entity Summaries with an Entity-aspect Model and Pattern Mining. In: Proceedings of the 48th ACL, 2010.

Li S.J, Wang W., Zhang Y.W.. 2009, TAC 2009 Update Summarization of ICL, In Proceedings of TAC 2009.

Li S.J, Song T., Wang X., 2010, TAC 2010 Update Summarization and AESOP of ICL, In Proceedings of TAC 2010.

Lin.C.Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization, Barcelona. ACL.

Zhou, D., Weston J., Gretton A., Bousquet O. and SchOlkopf B.. 2003. Ranking on data manifolds. In Proceedings of NIPS’2003.