TAC 2010 Update Summarization and AESOP 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 guided summarization task of TAC 2010, we submitted two runs using the combination of manifold ranking score and aspects information. The only difference between these two runs is that one computes a score according to aspect information while the other one considers aspect coverage in each sentence. For the update summarization task, we adopt similar methods and simultaneously penalize the information overlap between docset *B* and docset *A*. For the AESOP task, we focused on No Models evaluation case and submitted three runs. The first one uses a simple linear combination of three ROUGE scores, including ROUGE-1, ROUGE-2 and ROUGE-SU4, with four human summaries as models. The other two methods extract term information from four human summaries, such as term frequency and model frequency, based on which we evaluate each automatic summary. The difference between these two runs is that one of them introduces more term information, such as bigram information.

1. Introduction

The TAC[1] 2010 guided summarization task is not quite the same as that in TAC 2009(Hoa 2009). Although they both aim at generating short (no more than 100 words) fluent multi-document summaries of news articles with or without considering earlier articles on the topic, TAC 2010 summarization task aims to encourage a deeper linguistic (semantic) analysis of documents. Then in the TAC 2010 summarization task, topics are classified into five topic categories, and each category has a predefined list of important aspects. Good summaries are expected to cover all these aspects and be readable. We still adopt the sentence-extractive framework, where both the manifold ranking method and the aspect information are used to extract important sentences. When generating update summaries for document set *B*, sentences are penalized for their content overlap with document set *A*.

The TAC 2010 AESOP summarization task is almost the same as that in TAC 2009 and aims to promote research and development of systems that automatically evaluate the quality of summaries. It is the first time that we participate in this task. Focusing mainly on the "No Models" evaluation, we submitted three runs. The reference summaries are used to compute automatic summaries' scores for performance evaluation.

2. Our methods

2.1 Guided summarization task

In the guided summarization task, all the documents for a given topic are separated into two docsets *A* and *B*. The documents in docset *A* are summarized directly and the documents in docset *B* are summarized with the assumption that the documents in docset *A* have been read. For TAC 2010 summarization task, we use both sentence manifold ranking scores and category aspects information to rank the importance of each sentence, while choosing sentences from a set of documents to construct a summary. A sentence's manifold ranking score is computed by manifold ranking method which was used in our submission to TAC 2009 update summarization task (Li 2009, zhou 2003). In the first run, we assign each sentence a score as the first step of our manifold ranking method and then choose highest

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

scored sentences while simultaneously considering their aspect coverage. In the second run, we quantify the category aspects information with an aspect score and then linearly combine the aspect score with the manifold ranking score to form the final score of each sentence. Sentences with highest scores are selected into the summary.

We assume that the aspect information can be encoded by some words or patterns, which we will call aspect words here. Then, we need to select aspect words from the corpus, and define aspect word sets for each category's aspects. Those words which appear in the sentences belonging to a particular aspect are picked out as a signature of the corresponding aspect. In practice, if a sentence contains the word in the aspect word set, it's assumed that the sentence covers this aspect. Aspect word sets can be constructed automatically or manually. In order to guarantee the quality, we pick out aspect word sets manually through the observation of each topic on TAC 2010. We pick out 6 words for each aspect on average, mostly verbs. Aspect words can also be chosen automatically using term information or some useful rules. We will try these methods in the future work.

In the first run, we compute the sentences' manifold scores in the first step, according to which sentences are sorted in descending order. While choosing sentences with highest scores into the topic summary, we record the aspects which are covered by this chosen sentence with a predefined aspect set. This method assumes that if one sentence doesn't contain any additional aspect information, it won't be selected. If all the aspects belonging to a sentence have already been included in the chosen aspect set, this sentence won't be chosen. This method will choose sentences until the word limit of 100 words is reached.

In the second run, the manifold score of each sentence is computed as in the first run. However, since different aspects have different contributions to sentences, we compute a weight for each aspect to measure its coverage globally. Here we think that different aspects may contribute to the meaning of a sentence with different degrees of importance. If an aspect occurs in a lot of sentences, this aspect is not so important to the sentences. With this assumption, we compute the ratio of the sentences coverage for each aspect to all sentences. Similar with *tf*idf* (term frequency * inverse document frequency) methods, we compute an aspect coverage score for each sentence *s* with the following formula:

$$S_{aspect}(s) = -\sum_{a} tf_s(a) \log \frac{N(a)+1}{N} \qquad (1)$$

Where a stands for a particular kind of aspect and $tf_s(a)$ gives the frequency of aspect a in sentence s, $N(a)$ represents the number of sentences where a occurs and N is the number of all the sentences. Now, we multiply the sentence manifold score by the aspect coverage score to get the final sentence score, and rank sentences in descending order. Finally, sentences with highest scores are chosen into the topic summaries. The stop condition is the same as that of the first run.

For update summarization, we add penalty to sentences from docset *B* which have information overlap with docset *A*. The key is how to compute the overlap. Intuitively, overlap should be measured by the similarity of a sentence from docset B to docset *A*. We apply a simple formula to measure the overlap:

$$\text{Overlap}Score(s_i) = b * sim(s_i, old\_content) \quad (2)$$

where *b* is a penalty factor between 0 and 1, *old_content* can be docset *A* or summary *A*. As our experiment in the TAC 2009 shows the former one is better, we choose all the corresponding topic documents from docset *A* as *old_content*. Possible reason is that the summary *A* is automatically generated and usually doesn't include information as comprehensive as docset *A*, and using the summary may generate more uncertainty. Here, the cosine

similarity is adopted. Similarity normally can be computed by *maxsim* or *avgsim* methods. *Maxsim* computes all the similarity values between s_i and each sentence in the *old_content*, and returns the highest one, while *avgsim* returns the average of all these values. We choose *maxsim* method for its better performance in our TAC 2009 experiments. Thus the final formula for update summarization sentence score is as follows:

$$S(s) = S_B(s) - b * maxsim(s, docsetA) \qquad (3)$$

where $S_B(s)$ computes scores of sentences in docset *B* using manifold score and aspect coverage information as described above.

2.2 AESOP task

In the AESOP task, No Models case is to produce scores for each peer summary excluding the model summaries, while All Peers case is for each peer and model summary. We focused on No Models evaluation case and submitted three runs. It's our first time to participate AESOP task and we implement some simple methods.

In the first run for No Models evaluation case, we firstly get ROUGE scores while using human summaries as models and automatic summaries as peers, we use a simple linear combination of ROUGE-1 score, ROUGE-2 score and ROUGE-SU4 score (Lin 2004) as peer summary score.

The other two methods extract term information from human summaries of all topics, based on which we evaluate each peer summary. The idea of the pyramid evaluation is adopted, but we use *term* instead of *scu* (summary content unit), because *scu* are labeled manually and difficult to be generated automatically. In the second run for No Models evaluation case, we extract unigram frequency (*uf*) and model frequency (*mf*) from four model summaries, according to which peer summaries' scores are computed. Unigram frequency (*uf*) means the frequency of the unigram appearing in one topic's all four model summaries. Model frequency (mf) means the number of model summaries the unigram appears in one topic. Unigrams are extracted from the stemmed model summaries without stop words. The difference between model frequency and inverse document frequency is how they weight the effect of terms. The former one prefers unigrams which occur in more models, while the other one attenuates the effect if the term occurs in too many documents. High model frequency is taken as an agreement to the term's importance. We assume important summary words are unigrams whose *uf* and *mf* are high. The importance of summary word (unigram) is computed with the following formula:

$$score(u) = \frac{\log(uf(u)+1)}{\log(\frac{4}{mf(u)}+1)} \qquad (4)$$

where u stands for a unigram in model summaries of one topic, *uf* means unigram frequency, *mf* means model frequency. Then we can compute all peers' summaries quality according to the unigrams' scores.

In the third run for No Models evaluation case, the method is similar to the second run, but uses more term information, such as bigram information.

3 Experiments and Evaluation

TAC 2010 guided summarization task test datasets comprises of 46 topics. Each topic belongs to a predefined category and has 20 relevant documents which have been divided evenly into 2 docsets (*A*, *B*). We will introduce our evaluation results on TAC 2010 guided summarization task in subsection 1 and introduce our evaluation results on TAC 2010 AESOP task in subsection 2.

3.1 Evaluation of guided summarization task

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 which we call *Run1* and *Run2* as described in section 2.1. Table 1 illustrates the automatic evaluation results of our system. 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 named TOP1. The manual evaluation results are listed in Table 2. The suffix "_A" and "_B" represent summarizer results of *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	0.09574 (0.08421 – 0.10740)	0.13014 (0.12081 - 0.13996)	0.05937 (0.04960 - 0.06992)
BASE1_A	0.05376 (0.04468 – 0.06422) (33)	0.08551 (0.07719 - 0.09453) (34)	0.02713 (0.01985 - 0.03517) (31)
BASE2_A	0.05927 (0.05084 – 0.06821) (28)	0.09112 (0.08242 - 0.09988) (32)	0.03328 (0.02659 - 0.04068) (29)
Run1_A	0.08223 (0.07192 – 0.09329) (17)	0.11865 (0.11070 - 0.12716) (15)	0.04889 (0.04029 - 0.05809) (18)
Run2_A	0.08350 (0.07285 – 0.09302) (14)	0.11892 (0.10997 - 0.12693) (13)	0.04842 (0.03875 - 0.05833) (19)
TOP1_B	0.08024 (0.07109 - 0.08978)	0.12006 (0.11176 - 0.12855)	0.04445 (0.03747 - 0.05203)
BASE1_B	0.05327 (0.04473 - 0.06256) (29)	0.08822 (0.08014 - 0.09636) (32)	0.03170 (0.02369 - 0.04087) (24)
BASE2_B	0.06266 (0.05357 - 0.07206) (19)	0.09673 (0.08773 - 0.10561)) (23)	0.03769 (0.03091 - 0.04477) (12)
Run1_B	0.06221 (0.05350 - 0.07217) (20)	0.10203 (0.09399 - 0.11066) (19)	0.03716 (0.02999 - 0.04501) (13)
Run2_B	0.06664 (0.05821 - 0.07604) (14)	0.10395 (0.09604 - 0.11235) (18)	0.03815 (0.03128 - 0.04517) (10)

Table 1: Automatic Evaluation in TAC 2010

	Pyramid	Ling. quality	Resp
TOP1_A	0.425	3.652	3.174
BASE1_A	0.233(32)	3.652(1)	2.174(32)
BASE2_A	0.296(27)	2.717(30)	2.500(26)
Run1_A	0.398(6)	2.978(20)	3.022(9)
Run2_A	0.381(15)	2.957(21)	2.978(14)
Top 1_B	0.321	3.739	2.717
BASE1_A	0.187(28)	3.739(1)	2.022(28)
BASE2_B	0.262(9)	2.696(30)	2.478(7)
Run1_B	0.261(10)	2.848(23)	2.391(11)
Run2_B	0.280(4)	2.739(29)	2.478(6)

Table 2: Manual Evaluation in TAC 2010

From Table 1 and Table 2, we can see that our results are far better than two baselines and at top 1/3 rank. It can also be seen that our results are better for manual evaluation than for automatic evaluation, especially the overall responsiveness. This is because our methods take the sentences' aspect information into consideration. We think in the manual evaluation, the assessors are likely to pay more attention to the aspect information in the context of the summary. As our summaries contain more aspect information, they tend to have higher scores on manual evaluation. Besides, *Run2* has better performance than *Run1* as a whole. Because we use aspect words to identify whether a sentence covers an aspect, it's possible that aspect coverage score of *Run1* introduces more uncertainty than the method of *Run2* which records the aspect coverage situation.

3.2 Evaluation of AESOP task

Each AESOP run will be evaluated for correlation and discriminative power compared with the manual metric Pyramid and Overall Responsiveness. We submitted three runs, named *Run1*, *Run2* and *Run3* respectively. The organizer provides Baseline 1, Baseline 2 and Baseline 3 (named BASE1, BASE2 and BASE3 respectively), where BASE1 evaluates systems with

ROUGE-2 while BASE2 using ROUGE-SU4 and BASE3 using Basic Elements (BE). The evaluation results for No Models case are listed in the following tables. Table 3 illustrates Pearson's, Spearman's, and Kendall's correlations with Pyramid. We also list the results with the highest correlation as TOP 1. The correlations with Overall Responsiveness are listed in Table 4. Table 5 and Table 6 aim to show the discriminative power. The data is from the DISCRIMINATIVE POWER files, which use the contingency tables to compare the AESOP metric's discriminative power with Pyramid and Overall Responsiveness respectively. Table 5 lists the number of significance agreements, where each AESOP metric (one particular submission) and the manual evaluation (Pyramid or responsiveness method) agree that one summarizer is significantly better than the summarizer. Table 6 lists the number of insignificance agreements, where each AESOP metric and the manual evaluation agree that there is no significant difference between two summarizers. Because the total number of all the summarizers is 43, the number of all summarizer pairs is 903 (43*42/2). The correlation data measures the similarity between AESOP metrics and the Pyramid or Overall Responsiveness, while the discriminative power measures the metric's ability to distinguish between two non-model summarizers. The data listed in the following tables performs better when values are higher.

	Pearson	Spearman	Kendall
TOP1_A	0.978	0.948	0.836
BASE1_A	0.978	0.917	0.785
BASE2_A	0.968	0.948	0.836
BASE3_A	0.965	0.940	0.813
Run1_A	-0.020	0.063	0.027
Run2_A	0.960	0.902	0.785
Run3_A	0.919	0.851	0.712
TOP1_B	0.964	0.921	0.771
BASE1_B	0.963	0.915	0.762
BASE2_B	0.910	0.884	0.722
BASE3_B	0.953	0.905	0.771
Run1_B	-0.028	0.100	0.061
Run2_B	0.918	0.815	0.671
Run3_B	0.883	0.809	0.676

Table 3: Correlations with Pyramid

	Pearson	Spearman	Kendall
TOP1_A	0.979	0.958	0.835
BASE1_A	0.967	0.923	0.799
BASE2_A	0.955	0.952	0.835
BASE3_A	0.943	0.908	0.751
Run1_A	0.061	0.183	0.110
Run2_A	0.929	0.874	0.759
Run3_A	0.877	0.843	0.732
TOP1_B	0.960	0.910	0.787
BASE1_B	0.953	0.890	0.742
BASE2_B	0.899	0.874	0.729
BASE3_B	0.928	0.868	0.709
Run1_B	0.054	0.183	0.112
Run2_B	0.885	0.773	0.653
Run3_B	0.842	0.762	0.631

Table 4: Correlations with Responsiveness

	A		B	
	Pyramid	Respon-siveness	Pyramid	Respon-siveness
BASE1	278	259	198	185
BASE2	299	279	197	181
BASE3	224	212	177	163
Run1	64	67	35	35
Run2	312	287	218	200
Run3	294	272	208	189

Table 5: Discriminative Power Evaluation 1

	A		B	
	Pyramid	Respon-siveness	Pyramid	Respon-siveness
BASE1	561	566	580	586
BASE2	532	536	586	589
BASE3	573	585	639	644
Run1	353	370	458	467
Run2	484	483	501	502
Run3	488	490	546	546

Table 6: Discriminative Power Evaluation 2

From Table 3 and Table 4, we can see that baselines almost reach the top results, especially for initial summarization evaluation. Our *Run1* use a simple combination of three rouge scores without adjusting parameters, which causes a low correlation and shows that inappropriate combination is harmful to the final performance. Our *Run2* and *Run3* evaluate the summarizers by comparing with the corresponding model summaries with term information. These two runs show better performance. The main reason is that they absorb the idea of the two successful metrics -- ROUGE and the Pyramid. Besides, *Run2* performs better than *Run3*, which means that introducing bigram

information does not causes desirable results. The possible reason is that bigrams are mostly meaningless, and it is easier to cause uncertain effects. The evaluation of AESOP also proves that ROUGE scores have good performance on the summarization evaluation task.

4 Conclusions and Future Work

In this paper, we introduce the manifold ranking method combined with aspect coverage information, which is adopted for the guided summarization task in TAC 2010. We submitted two runs. The experimental results show that they are effective for summarization. For AESOP task, we utilize the automatic evaluation tool ROUGE and term information of model summaries. In the future, we will try more methods on update summarization task. And how to effectively make use of aspects information is another problem.

Acknowledgements

This work is supported by NSFC programs (No: 60875042 and 90920011), National Social Science Foundation (No: 10CYY023).

Refences

Hoa T. D.. 2008, Overview of the TAC 2008 Update Summarization Task. Text Analysis Conference 2008 http://www.nist.gov/tac/, 2008.

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

Li, S.J, Wang W., Wang C.. 2008, TAC 2008 Update Summarization Task of ICL, In Proceedings of TAC 2008.

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

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

Wan X., Yang J. and Xiao J.. 2007. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. IJCAI 2007, pp. 2903-2908.

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