IIT Kharagpur at TAC 2009: Statistical and Nugget-based Model for Automatic Summary Evaluation

Sushant Kumar and Abhijeet Kumar
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
India – 721302
sushant3d@gmail.com, abhijeet.cse.kgp@gmail.com

Abstract

In this paper we present our participation at TAC 2009 AESOP (Automatically Evaluating Summaries Of Peers) task. We make use of a statistical model for evaluation metric correlating to Overall Responsiveness and a nugget-based pyramid model for correlating to the Pyramid manual metric of TAC 2009. We also present the performance of our three submitted runs as per the official TAC 2009 evaluation results.

1 Introduction

The TAC 2009 AESOP task was to develop an automatic metric system to evaluate peer summaries obtained from the Update Summarization task of TAC 2009. A number of systems have been developed for automatic summarization of documents like SUMMONS (SUMMarizing Online NewS articles). These summarizers, during their development, needed to be tested from time to time to verify the performance of the system. If we use manual techniques for such system evaluations, then it will prove to be a very time-consuming task. Also, the evaluation metric might not be uniform for every manual evaluation. Hence, this brought the need for developing an automatic system to evaluate the summaries. The automatic summary evaluation systems aided the development of automatic summarizers with lesser effort for system evaluation. Some of the evaluation metrics used commonly for automatic evaluations are ROUGE (Lin and Hovy, 2003) and Pyramid F-score (Lin and Demner-Fushman, 2006).

The TAC 2009 AESOP task was primarily focused on developing automatic metrics that accurately measure summary content. The output of automatic metrics were correlated to two manual metrics: the (Modified) Pyramid score, which measures summary content, and Overall Responsiveness, which measures a

combination of content and linguistic quality.

2 Our Approach

We submitted three runs for the AESOP task. Different metrics were used for the three runs to find the usefulness of different parameters and resources in evaluating summaries. We discuss the three runs separately in sub-sections 2.1, 2.2 and 2.3 respectively.

2.1 Method 1: Statistical Model

The first run was built to correlate with the manual metric for Overall Responsiveness for TAC evaluation. For this system, we used the given model summaries apart from the topic statements. We did not use the original documents for this run. The system overview for evaluating each test summary is presented in Figure 1.

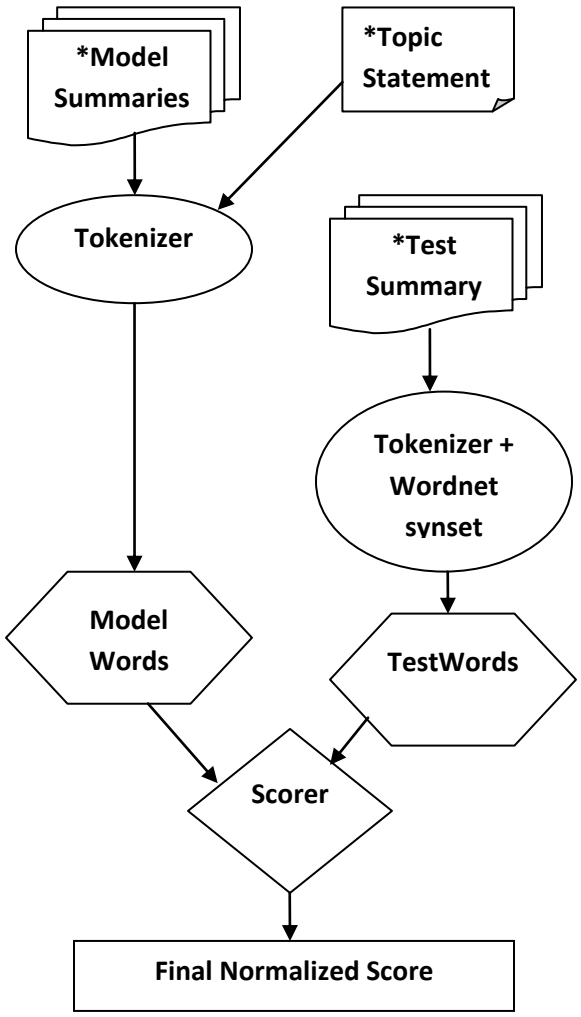

Figure 1: *System Overview for the first run correlated with Overall Responsiveness metric.* (* → *provided by TAC*)

The system description for evaluating one test summary is as follows:

1. We first extracted the words from the topic statement and extended the list further by adding its synsets from the WordNet after removing the stopwords. Let this wordlist be the TopicWords.
2. We then tokenized the words from the model summaries for the given topic statement. The stopwords were removed from this list. This gave us a list of ModelWords.
3. We generated a frequency distribution of each word in the ModelWords list by its occurrence in the Model summaries. Let it be denoted by CountM(m) = Number of occurrences of a word m in the Model summaries (where m ∈ ModelWords).

4. Each word m ∈ ModelWords was given a weight, Weight(m) = 1 + number of occurrences of word m in the TopicWords list.
5. We then extracted the words from the test summary and removed the stopwords from the list. Let this list be TestWords. The frequency distribution of each word in TestWords was obtained. Let CountT(t) = Number of occurrences of a word t in the test summary (where t ∈ TestWords)
6. The scoring module then compared the frequency distribution of the ModelWords and the TestWords to score the test summary. For scoring, we matched each word in ModelWords with every word in TestWords and their synsets. The final score for a summary was obtained as,

   Score =

   ∑ { min(countM(m),countT(m))* Weight(m) },

   ∀m ∈ ModelWords ∩ TopicWords +synsets

The final score for each test summary was normalized to an appropriate range.
For evaluating a Model summary, we only used the other three Model summaries for the above algorithm (Step 2).

2.2 Method 2: Pyramid Nugget-based Model

The second run built to correlate with the manual metric for Pyramid Score used only the model summaries for scoring. The Model summaries were used to identify key nuggets with appropriate weights. These nuggets were then used to evaluate the test summaries. The system overview is presented in Figure 2.

The system description for this run is as follows:

1. We first extracted all the sentences from the given Model summaries for a topic statement. Let this set be S.
2. Then we obtained the frequency distribution of each word (after removing the stopwords) for individual Model summary. Let this set be W.
3. Each sentence was then given a weight based on the frequency distribution of the words present in it. Weight of a sentence i, SW_i = Number of wordmatches for the sentence i/ Total number of wordmatches for each sentence, where a wordmatch = max (number of Model summaries in which the word appears/Total number of Model summaries).

   These sentences were now the nuggets with a normalized weight assigned to each of them.
4. Now, to score a given test summary, we first extracted all the words in the summary and added its synsets after removing the stopwords. Let this be the TestList.
5. We identified the presence of a nugget in the test summary by

finding the number of words in a nugget that were present in the TestList. If this was greater than a particular threshold value then the nugget was assumed to be present else absent.

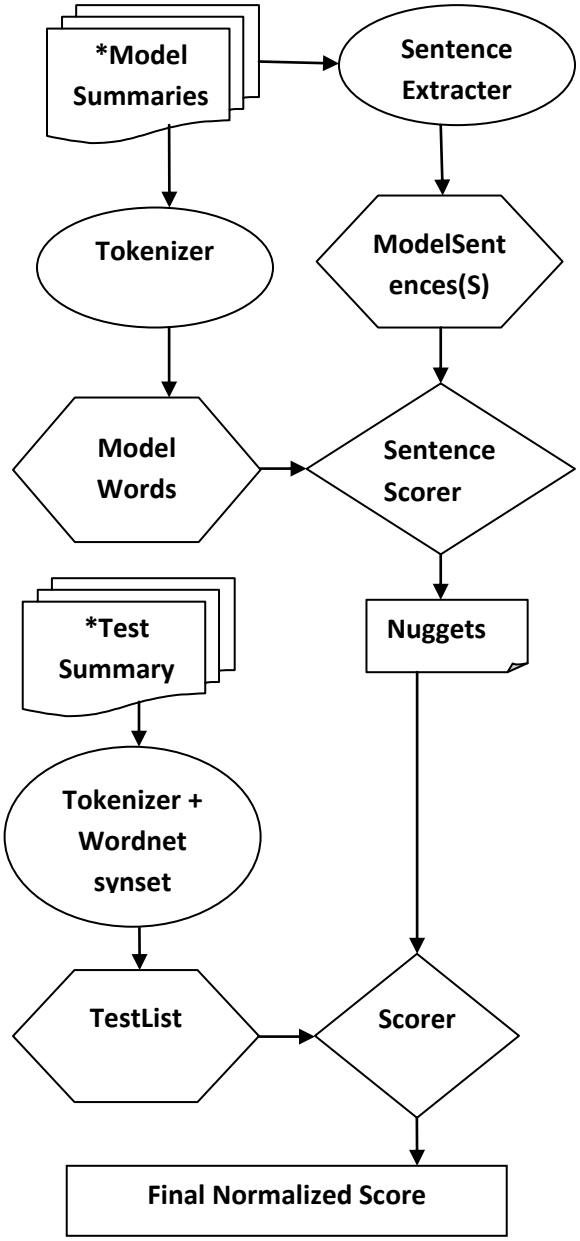

Figure 2: *System Overview for the second run correlated with Pyramid metric.* (* → *provided by TAC*)

6. The final score of the summary was given by,
   Score = ∑(weight of nuggets present) / ∑(weight of all nuggets)

For evaluating a Model summary, we only used the other three Model summaries for the above algorithm (Step 1).

2.3 Method 3: Pyramid Nugget-based Model without using Model Summaries

Our third run was also designed to correlate with the manual metric for Pyramid Score. Unlike the second run, this run made use of the topic statements and the original documents, but did not use the Model summaries for evaluation. The system overview for this run has been described in Figure 3. The description for this system is as follows:

1. First, we extracted all the sentences from the topic documents. Let this be S.
2. Then, we obtained all the words in the topic statement including their synsets. Let this be T.
3. For each sentence S_i, we calculated its weight as the number of words in that sentence which are also present in T. If this weight was above a particular threshold value then this sentence was used as a nugget and put in the nugget list, say N.

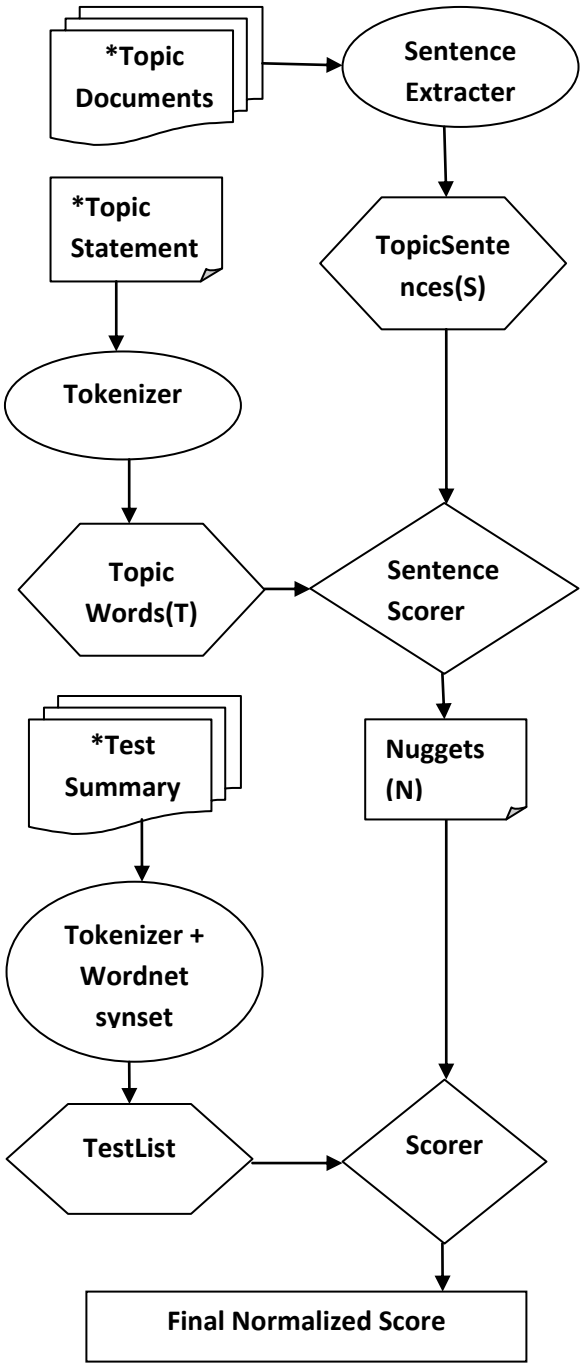

Figure 3: *System Overview for the third run correlated with Pyramid metric.* (* → *provided by TAC*)

4. Now, for scoring a test summary, we first extract the words in the test summary along with its synsets. Let this be the TestList.
5. Next each nugget in N is searched in the test summary. A nugget was said to be present in the test summary if the number of words in that nugget which were also present in the TestList was above a particular threshold value, otherwise the nugget was assumed to be absent.
6. The final score of the summary was given by,
   Score = ∑(weight of nuggets present) / ∑(weight of all nuggets)

3 Evaluation

In the TAC 2009 AESOP task, we were provided with 44 topic statements, 88 document sets. For each topic statement there were 4 human-authored model summaries, 3 baseline summarizer runs and 52 summaries evaluated in the TAC 2009 Update Summarization task used for scoring by AESOP runs. Finally 37 AESOP metrics were evaluated with two baseline automatic metrics - ROUGE-SU4 and Basic Elements (BE). We present below the TAC evaluation results for our three runs and some of the other runs for comparison of performances. In the following tables, our first, second and third runs have been represented as 1, 2 and 3 respectively, while 4 is the best performing run in that category. The categories are Auto-models (comparing

model summary against non-model summary) and No-models (comparing non-model summaries) for both A and B sets of documents. The tables 1 and 2 show the discriminative power of the runs.

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	50	27	1179	968
2	437	378	1376	1286
3	253	82	1381	1256
4	440	439	1415	1379

Table 1: *Number of agreements of the runs with TAC's Pyramid metric.*

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	50	21	1029	891
2	429	372	1331	1347
3	253	76	1278	1207
4	440	433	1403	1352

Table 2: *Number of agreements of the runs with TAC's Overall Responsiveness metric.*

The following tables show the Pearson's, Spearman's and Kendall's correlations of our runs (1, 2 and 3) and the best run (4) metrics with the Pyramid and Overall Responsiveness scores for TAC 2009 initial summaries.

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	0.433	-0.038	0.891	0.452
2	0.978	0.978	0.963	0.957
3	0.877	0.740	0.897	0.767
4	0.983	0.978	0.978	0.970

Table 3: *Pearson's correlations with TAC's Pyramid metric.*

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	0.629	0.191	0.820	0.512
2	0.933	0.941	0.902	0.916
3	0.913	0.893	0.873	0.853
4	0.962	0.966	0.950	0.955

Table 4: *Spearman's correlations with TAC's Pyramid metric.*

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	0.455	0.119	0.634	0.352
2	0.796	0.817	0.750	0.781
3	0.766	0.729	0.715	0.683
4	0.835	0.858	0.820	0.841

Table 5: *Kendall's correlations with TAC's Pyramid metric.*

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	0.315	-0.085	0.793	0.506
2	0.938	0.929	0.851	0.833
3	0.819	0.687	0.827	0.761
4	0.968	0.963	0.872	0.833

Table 6: *Pearson's correlations with TAC's Overall Responsiveness metric.*

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	0.535	0.141	0.714	0.466
2	0.833	0.835	0.751	0.756
3	0.832	0.802	0.752	0.717
4	0.913	0.878	0.873	0.826

Table 7: *Spearman's correlations with TAC's Overall Responsiveness metric.*

	Auto-models_A	Auto-models_B	No-models_A	No-models_B
1	0.384	0.090	0.540	0.324
2	0.663	0.675	0.575	0.549
3	0.664	0.627	0.579	0.595
4	0.761	0.728	0.707	0.676

Table 8: *Kendall's correlations with TAC's Overall Responsiveness metric.*

5 Conclusion

We submitted three runs for TAC 2009 AESOP task. Our first run was based on a simple statistical model using the unigram frequency distribution in the model summaries and the test summary to evaluate it. The TAC evaluation scores of this run show a poor performance in both the discriminative power and the correlation scores.

The second run was based on nugget-based pyramid method which used only the model summaries for evaluation making no use of the original documents or the topic statements. The TAC evaluation results for this run have been quite competitive for both the discriminative power and the correlation scores.

The third run was also based on nugget-based pyramid method using the Topic statements and original documents for evaluation. However it did not use the Model summaries for scoring. The TAC evaluation results show that it has a better discriminative power for No-Models evaluation metric than the All Peers scores. The correlation scores for this run have been average compared to the second run.

The TAC evaluations show that the run built on nugget-based model using the Model summaries have performed better than the other two runs. The third run which did not use any Model summary has

also shown some promise and can be an interesting system to develop further.

References

[1] C.Y.Lin and Hovy, E.H. 2003. Automatic Evaluation of Summaries using nGram Cooccurrence Statistics. *Proceedings of the* HLT2003 conference.

[2] Hovy, E.H., C.Y. Lin, and L. Zhou. 2005. Evaluating DUC 2005 using Basic Elements.

[3] Hovy, E.H. 2005. Automated Text Summarization. In R. Mitkov (ed), *The Oxford Handbook of Computational Linguistics*, pp. 583–598. Oxford: Oxford University Press.

[4] Lin, J. and D. DemnerFushman. 2005. Evaluating Summaries and Answers: Two Sides of the Same Coin? *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

[5] A. Louis, A. Nenkova. 2008. Automatic Summary Evaluation without Human Models. *Proceedings of* TAC 2008.

[6] S. Tratz, E.H. Hovy. 2008. Summarization Evaluation Using Transformed Basic Elements. *Proceedings of* TAC 2008.