Evaluation of automatic summaries: Metrics under varying data conditions

Karolina Owczarzak and **Hoa Trang Dang**
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899
karolina.owczarzak@nist.gov hoa.dang@nist.gov

Abstract

In evaluation of automatic summaries, it is necessary to employ multiple topics and human-produced models in order for the assessment to be stable and reliable. However, providing multiple topics and models is costly and time-consuming. This paper examines the relation between the number of available models and topics and the correlations with human judgment obtained by automatic metrics ROUGE and BE, as well as the manual Pyramid method. Testing all these methods on the same data set, taken from the TAC 2008 Summarization track, allows us to compare and contrast the methods under different conditions.

1 Introduction

Appropriate evaluation of results is an important aspect of any research. In areas such as automatic summarization, the problem is especially complex because of the inherent subjectivity in the task itself and its evaluation. There is no single objective standard for a good quality summary; rather, its value depends on the summary's purpose, focus, and particular requirements of the reader (Spärck Jones, 2007). While the purpose and focus can be set as constant for a specific task, the variability of human judgment is more difficult to control. Therefore, in attempts to produce stable evaluations, it has become standard to use multiple judges, not necessarily for parallel evaluation, but in such a way that each judge evaluates a different subset of the many summaries on which the final system assessment is based. The incorporation of multiple points of view is also reflected in automatic evaluation, where it takes the form of employing multiple model summaries to which a candidate summary is compared.

Since these measures to neutralize judgment variation involve the production of multiple model summaries, as well as multiple topics, evaluation can become quite costly. Therefore, it is interesting to examine how many models and topics are necessary to obtain a relatively stable evaluation, and whether this number is different for manual and automatic metrics. In their examination of summary evaluations, van Halteren and Teufel (2003) suggest that it is necessary to use at least 30 to 40 model summaries for a stable evaluation; however, Harman and Over (2004) argue that a stable evaluation can be conducted even with a single model, as long as there is an adequate number of topics. This view is supported by Lin (2004a), who concludes that "correlations to human judgments were increased by using multiple references but using single reference summary with enough number of samples was a valid alternative". Interestingly, similar conclusions were also reached in the area of Machine Translation evaluation; in their experiments, Zhang and Vogel (2004) show that adding an additional reference translation compensates the effects of removing 10–15% of the testing data, and state that, therefore, "it seems more cost effective to have more test sentences but fewer reference translations".

In this paper, we look at how various metrics behave with respect to a variable number of topics and models used in the evaluation. This lets us determine the stability of individual metrics, and helps to illuminate the trade-offs inherent in designing a good evaluation. For our experiments, we used data from the Summarization track at the Text Analysis Conference (TAC) 2008, where participating systems were assessed on their summarization of 48 topics, and the automatic metrics ROUGE and BE, as well as the manual Pyramid evaluation method, had access to 4 human models. TAC 2008 was the first task of the TAC/DUC (Document Understanding Conference) series in which the Pyramid method was used on all evaluated data, making it possible to conduct a full com-

Proceedings of the 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLP 2009, pages 23–30,
Suntec, Singapore, 6 August 2009.

parison among the manual and automatic methods. Despite the lack of full Pyramid evaluation in DUC 2007, we look at the remaining metrics applied that year (ROUGE, BE, and Content Responsiveness), in order to see whether they confirm the insights gained from the TAC 2008 data.

2 Summary evaluation

The main evaluation at TAC 2008 was performed manually, assessing the automatic candidate summaries with respect to Overall Responsiveness, Overall Readability, and content coverage according to the Pyramid framework (Nenkova and Passonneau, 2004; Passonneau et al., 2005). Task participants were asked to produce two summaries for each of the 48 topics; the first (initial summary) was a straightforward summary of 10 documents in response to a topic statement, which is a request for information about a subject or event; the second was an update summary, generated on the basis of another set of 10 documents, which followed the first set in temporal order and described further developments in the given topic. The idea behind the update summary was to avoid repeating all the information included in the first set of documents, on the assumption that the reader is familiar with that information already.

The participating teams submitted up to three runs each; however, only the first and second runs were evaluated manually due to limited resources. For each summary under evaluation, assessors rated the summary from 1 (very poor) to 5 (very good) in terms of Overall Responsiveness, which measures how well the summary responds to the need for information expressed in the topic statement and whether its linguistic quality is adequate. Linguistic qualities such as grammaticality, coreference, and focus were also evaluated as Overall Readability, also on the scale from 1 to 5. Content coverage of each summary was evaluated using the Pyramid framework, where assessors create a list of information nuggets (called Summary Content Units, or SCUs) from the set of human-produced summaries on a given topic, then decide whether any of these nuggets are present in the candidate summary. All submitted runs were evaluated with the automatic metrics: ROUGE (Lin, 2004b), which calculates the proportion of *n*-grams shared between the candidate summary and the reference summaries, and Basic Elements (Hovy et al., 2005), which compares the candidate to the models in terms of head-modifier pairs.

2.1 Manual metrics

Evaluating Overall Responsiveness and Overall Readability is a rather straightforward procedure, as most of the complex work is done in the mind of the human assessor. Each candidate summary is given a single score, and the final score for the summarization system is the average of all its summary-level scores. The only economic factor here is the number of topics, i.e. summaries per system, that need to be judged in order to neutralize both intra- and inter-annotator variability and obtain a reliable assessment of the summarization system.

When it comes to the Pyramid method, which measures content coverage of candidate summaries, the need for multiple topics is accompanied by the need for multiple human model summaries. First, independent human assessors produce summaries for each topic, guided by the topic statement. Next, in the Pyramid creation stage, an assessor reads all human-produced summaries for a given topic and extracts all "information nuggets", called Summary Content Units (SCUs), which are short, atomic statements of facts contained in the text. Each SCU has a weight which is directly proportional to the number of model summaries in which it appears, on the assumption that the fact's importance is reflected in how many human summarizers decide to include it as relevant in their summary. Once all SCUs have been harvested from the model summaries, an assessor then examines each candidate summary to see how many of the SCUs from the list it contains. The final Pyramid score for a candidate summary is its total SCU weight divided by the maximum SCU weight available to a summary of average length (where the average length is determined by the mean SCU count of the model summaries for this topic). The final score for a summarization system is the average score of all its summaries. In TAC 2008, the evaluation was conducted with 48 topics and 4 human models for each topic.

We examined to what extent the number of models and topics used in the evaluation can influence the Pyramid score and its stability. The stability, similarly to the method employed by Voorhees and Buckley (2002) for Information Retrieval, is determined by how well a system ranking based on a small number of models/topics cor-

Models	Pyramid	ROUGE-2	ROUGE-SU4	BE
1	0.8839	0.8032	0.7842	0.7680
2	0.8943	0.8200	0.7957	0.7983
3	0.8974*	0.8258	0.7999*	0.8098
4 (bootstr)	0.8972*	0.8310	0.8023*	0.8152
4 (actual)	0.8997	0.8302	0.8033	0.8171

Table 1: Mean correlations of Responsiveness and other metrics using 1, 2, 3, or 4 models for TAC 2008 initial summaries. Values in each row are significantly different from each other at 95% level.

Models	Pyramid	ROUGE-2	ROUGE-SU4	BE
1	0.9315	0.8861	0.8874	0.8716
2	0.9432	0.9013	0.8961	0.8978
3	0.9474*	0.9068*	0.8994	0.9076
4 (bootstr)	0.9481*	0.9079*	0.9023	0.9114
4 (actual)	0.9492	0.9103	0.9020	0.9132

Table 2: Mean correlations of Responsiveness and other metrics using 1, 2, 3, or 4 models for TAC 2008 update summaries. Values in each row are significantly different from each other at 95% level except ROUGE-2 and ROUGE-SU4 in 1-model category.

Models	ROUGE-2	ROUGE-SU4	BE
1	0.8789	0.8671	0.8553
2	0.8972	0.8803	0.8917
3	0.9036	0.8845	0.9048
4 (bootstr)	0.9082	0.8874	0.9107
4 (actual)	0.9077	0.8877	0.9123

Table 3: Mean correlations of 4-model Pyramid score and other metrics using 1, 2, 3, or 4 models for TAC 2008 initial summaries. Values in each row are significantly different from each other at 95% level except ROUGE-2 and BE in 4-model category.

Models	ROUGE-2	ROUGE-SU4	BE
1	0.9179	0.9110	0.9016
2	0.9336	0.9199	0.9284
3	0.9392	0.9233	0.9383
4 (bootstr)	0.9443	0.9277	0.9436
4 (actual)	0.9429	0.9263	0.9446

Table 4: Mean correlations of 4-model Pyramid score and other metrics using 1, 2, 3, or 4 models for TAC 2008 update summaries. Values in each row are significantly different from each other at 95% level except ROUGE-2 and BE in 4-model category.

relates with the ranking based on another set of models/topics, where the two sets are randomly selected and mutually exclusive. This methodology allows us to check the correlations based on up to half of the actual number of models/topics only (because of the non-overlap requirement), but it gives an indication of the general tendency. We also look at the correlation between the Pyramid score and Overall Responsiveness. We don't expect a perfect correlation between Pyramid and Responsiveness in the best of times, because Pyramid measures content *identity* between the candidate and the model, and Responsiveness measures content *relevance* to topic as well as linguistic quality. However, the degree of variation between the two scores depending on the number of models/topics used for the Pyramid will give us a certain indication of the amount of information lost.

2.2 Automatic metrics

Similarly to the Pyramid method, ROUGE (Lin, 2004b) and Basic Elements (Hovy et al., 2005) require multiple topics and model summaries to produce optimal results. ROUGE is a collection of automatic *n*-gram matching metrics, ranging from unigram to four-gram. It also includes measurements of the longest common subsequence, weighted or unweighted, and the option to compare stemmed versions of words and omit stopwords. There is also the possibility of accepting skip-*n*-grams, that is, counting *n*-grams as matching even if there are some intervening non-matching words. The skip-*n*-grams together with stemming are the only ways ROUGE can accomodate alternative forms of expression and match concepts even though they might differ in terms of their syntactic or lexical form.

These methods are necessarily limited, and so ROUGE relies on using multiple parallel model summaries which serve as a source of lexical/syntactic variation in the comparison process. The fewer models there are, the less reliable the score. Our question here is not only what this relation looks like (as it was examined on the basis of Document Understanding Conference data in Lin (2004a)), but also how it compares to the reliability of other metrics.

Basic Elements (BE), on the other hand, goes beyond simple string matching and parses the syntactic structure of the candidate and model to obtain a set of head-modifier pairs for each, and then compares the sets. A head-modifier pair consist of the head of a syntactic unit (e.g. the noun in a noun phrase), and the word which modifes the head (i.e. a determiner in a noun phrase). It is also possible to include the name of the relation which connects them (i.e. *subject*, *object*, etc.). Since BEs reflect thematic relations in a sentence rather than surface word order, it should be possible to accommodate certain differences of expression that might appear between a candidate summary and a reference, especially as the words can be stemmed. This could, in theory, allow us to use fewer models for the evaluation. In practice, however, it fails to account for the total possible variety, and, what is more,

the additional step of parsing the text can introduce noise into the comparison.

TAC 2008 and DUC 2007 evaluations used ROUGE-2 and ROUGE-SU4, which refer to the recall of bigram and skip-bigram (with up to 4 intervening words) matches on stemmed words, respectively, as well as a BE score calculated on the basis of stemmed head-modifier pairs without relation labels. Therefore, these are the versions we use in our comparisons.

3 Number of models

Since Responsiveness score does not depend on the number of models, it serves as a reference against which we compare the remaining metrics, while we calculate their score with only 1, 2, 3, or all 4 models. Given 48 topics in TAC 2008, and 4-model summaries for each topic, there are 4^{48} possible combinations to derive the final score in the single-model category, so to keep the experiments simple we only selected 1000 random samples from that space. For 1000 repetitions, each time we selected a random combination of model summaries (only one model out of 4 available per topic), against which we evaluated the candidate summaries. Then, for each of the 1000 samples, we calculated the correlation between the resulting score and Responsiveness. We then took the 1000 correlations produced in this manner, and computed their mean. In the same way, we calculated the scores based on 2 and 3 model summaries, randomly selected from the 4 available for each topic. The correlation means for all metrics and categories are given in Table 1 for initial summaries and Table 2 for update summaries. We also ran a one-way analysis of variance (ANOVA) on these correlations to determine whether the correlation means were significantly different from each other. For the 4-model category there was only one possible sample for each metric, so in order to perform ANOVA we bootstrapped this sample to produce 1000 samples. The actual value of the 4-model correlation is given in the tables as **4 (actual)**, and the mean value of the bootstrapped 1000 correlations is given as **4 (bootstr)**.

Values for initial summaries are significantly different from their counterparts for update summaries at the 95% level. Pairwise testing of values for statistically significant differences is shown with symbols: in each column, the first value marked with a particular symbol is not significantly different from any subsequent value marked with the same symbol.

We also examined the correlations of the metrics with the 4-model Pyramid score. Table 3 presents the correlation means for the initial summaries, and Table 4 shows the correlation means for the update summaries.

Since the Pyramid, contrary to Responsiveness, makes use of multiple model summaries, we examine its stability given a decreased number of models to rely on. For this purpose, we correlated the Pyramid score based on randomly selected 2 models (half of the model pool) for each topic with the score based on the remaining 2 models, and repeated this 1000 times. We also looked at the 1-model category, where the Pyramid score calculated on the basis of one model per topic was correlated with the Pyramid score calculated on the basis on another randomly selected model. In both case we witness a very high mean correlation: 0.994 and 0.995 for the 2-model category, 0.982 and 0.985 for the 1-model category for TAC initial and update summaries, respectively. As an illustration, Figure1 shows the variance of correlations for the initial summaries.

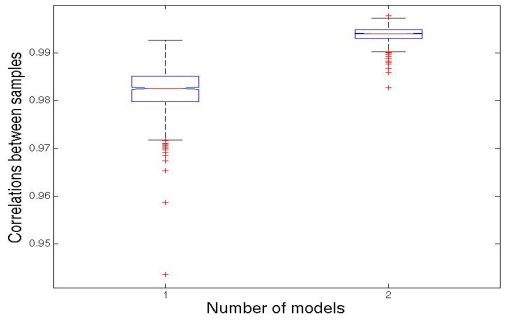

Figure 1: Correlations between Pyramid scores based on 1 or 2 model summaries for TAC 2008 initial summaries.

The variation in correlation levels between other metrics and Pyramid and Responsiveness, presented in Tables 3–4, is more visible in the graph form. Figures 2-3 illustrate the mean correlation values for TAC 2008 initial summaries. While all the metrics record the steepest increase in correlation values with the addition of the second model, adding the third and fourth model provides the metrics with smaller but steady improvement, with the exception of Pyramid-Responsiveness correlation in Figure 2. The increase in correlation mean is most dramatic for BE, which in all cases starts as the lowest-

correlating metric in the single-model category, but by the 4-model point it outperforms one or both versions of ROUGE. The Pyramid metric achieves significantly higher correlations than any other metric, independent of the number of models, which is perhaps unsurprising given that it is a manual evaluation method. Of the two ROUGE versions, ROUGE-2 seems consistently a better predictor of both Responsiveness and the "full" 4-model Pyramid score than ROUGE-SU4.

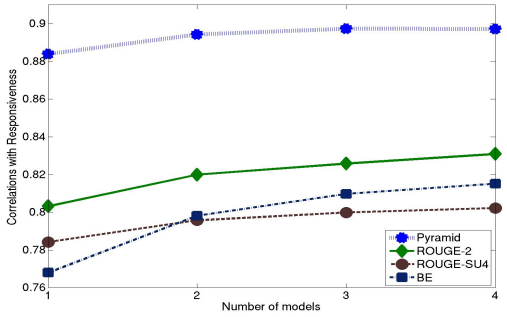

Figure 2: Responsiveness vs. other metrics with 1, 2, 3, or 4 models for TAC 2008 initial summaries.

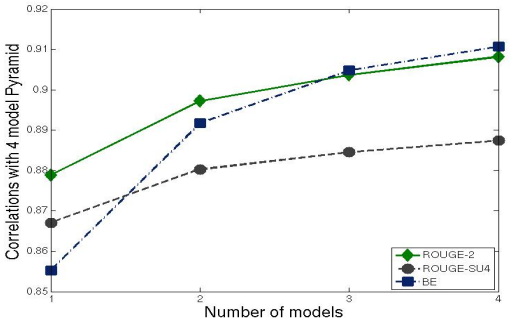

Figure 3: 4-model Pyramid vs. other metrics with 1, 2, 3, or 4 models for TAC 2008 initial summaries.

Similar patterns appear in DUC 2007 data (Table 5), despite the fact that the Overall Responsiveness of TAC 2008 is replaced with Content Responsiveness (ignoring linguistic quality), against which we calculate all the correlations. Although the increase in correlation means from 1- to 4-models for the three automatic metrics is smaller than for TAC 2008, the clearest rise occurs with the addition of a second model, especially for BE, and the subsequent additions change little. As in the case of initial summaries 2008, ROUGE-2 outperforms the remaining two metrics independently of the number of models. However, most of the increases are too small to be significant.

This comparison suggests diminishing returns

Models	ROUGE-2	ROUGE-SU4	BE
1	0.8681	0.8254	0.8486
2	0.8747*	0.8291*	0.8577*
3	0.8766*†	0.8299*†	0.8599*
4 (bootstr)	0.8761*†	0.8305*†	0.8633
4 (actual)	0.8795	0.8301	0.8609

Table 5: Mean correlations of Content Responsiveness and other metrics using 1, 2, 3, or 4 models for DUC 2007 summaries. Values in each row are significantly different from each other at 95% level.

with the addition of more models, as well as different reactions among the metrics to the presence or absence of additional models. When correlating with Responsiveness, the manual Pyramid metric benefits very little from the fourth model, but automatic BE benefits most from almost every addition. ROUGE is situated somewhere between the two, noting small but often significant increases. On the whole, the use of multiple models (at least two) seems supported, especially if we use automatic metrics in our evaluation.

4 Number of topics

For the second set of experiments we kept all four models, but varied the number of topics which went into the final average system score. To determine the stability of Responsiveness and Pyramid we looked at the correlations between the scores based on smaller sets of topics. For 1000 repetitions, we calculated Pyramid/Responsiveness score based on a set of 1, 3, 6, 12, or 24 topics randomly chosen from the pool of 48, and compared the system ranking thus created with the ranking based on another, equally sized set, such that the sets did not contain common topics. Table 6 shows the mean correlation for each case. Although such comparison was only possible up to 24 topics (half of the whole available topic pool), the numbers suggest that at the level of 48 topics both Responsiveness and Pyramid are stable enough to serve as reference for the automatic metrics.

	Responsiveness		Pyramid	
Topics	Initial	Update	Initial	Update
1	0.182	0.196	0.333	0.267
3	0.405	0.404	0.439	0.520
6	0.581	0.586	0.608	0.690
12	0.738	0.738	0.761	0.816
24	0.849	0.866	0.851	0.901

Table 6: Mean correlations between Responsiveness/Pyramid scores based on 1, 3, 6, 12, and 24 topic samples for TAC 2008 initial and update summaries.

In a process which mirrored that described in Section 3, we created 1000 random samples in each of the *n*-topics category: 1, 3, 6, 12, 24, 36,

Topics	Pyramid	ROUGE-2	ROUGE-SU4	BE
1	0.4219	0.4276	0.4375	0.3506
3	0.6204	0.5980	0.9016	0.5108
6	0.7274	0.6901	0.6836	0.6233
12	0.8159	0.7618	0.7456	0.7117
24	0.8679	0.8040	0.7809	0.7762
36	0.8890*	0.8208*	0.7951*	0.8017*
39	0.8927*†	0.8231*†	0.7967*†	0.8063*†
42	0.8954*†‡	0.8258*†‡	0.7958*†‡	0.8102*†‡
45	0.8977*†‡§	0.8274*†‡§	0.8008*†‡§	0.8132†‡§
48 (bootstr)	0.8972*†‡§	0.8302*†‡§	0.8046†‡§	0.8138†‡§
48 (actual)	0.8997	0.8302	0.8033	0.8171

Table 7: Mean correlations of 48 topic Responsiveness and other metrics using from 1 to 48 topics for TAC 2008 initial summaries. Values in each row are significantly different from each other at 95% level except: ROUGE-2, ROUGE-SU4 and BE in 1-topic category, ROUGE-2 and ROUGE-SU4 in 3- and 6-topic category.

Topics	Pyramid	ROUGE-2	ROUGE-SU4	BE
1	0.5005	0.4882	0.5609	0.4011
3	0.7053	0.6862	0.7340	0.6097
6	0.8080	0.7850	0.8114	0.7274
12	0.8812	0.8498	0.8596	0.8188
24	0.9250	0.8882	0.8859	0.8774
36	0.9408*	0.9023*	0.8960*	0.8999*
39	0.9433*†	0.9045*†	0.8973*†	0.9037*†
42	0.9455*†‡	0.9061*†‡	0.8987*†‡	0.9068*†‡
45	0.9474†‡§	0.9078*†‡§	0.8996*‡‡§	0.9094†‡§
48 (bootstr)	0.9481†‡§	0.9101†‡§	0.9015*†‡§	0.9111†‡§
48 (actual)	0.9492	0.9103	0.9020	0.9132

Table 8: Mean correlations of 48 topic Responsiveness and other metrics using from 1 to 48 topics for TAC 2008 update summaries. Values in each row are significantly different from each other at 95% level except: Pyramid and ROUGE-2 in 1-topic category, Pyramid and ROUGE-SU4 in 6-topic category, ROUGE-2 and BE in 39-, 42-, and 48-topic category.

Topics	ROUGE-2	ROUGE-SU4	BE
1	0.4693	0.4856	0.3888
3	0.6575	0.6684	0.5732
6	0.7577	0.7584	0.6960
12	0.8332	0.8245	0.7938
24	0.8805	0.8642	0.8684
36	0.8980*	0.8792*	0.8966*
39	0.9008*†	0.8812*†	0.9017*†
42	0.9033*†‡	0.8839*†‡	0.9058†‡
45	0.9052*†‡§	0.8853*†‡§	0.9093†‡§
48 (bootstr)	0.9074†‡§	0.8877†‡§	0.9107†‡§
48 (actual)	0.9077	0.8877	0.9123

Table 9: Mean correlations of 48 topic Pyramid score and other metrics using from 1 to 48 topics for TAC 2008 initial summaries. Values in each row are significantly different from each other at 95% level except: ROUGE-2 and ROUGE-SU4 in the 6-topic category, ROUGE-2 and BE in 39- and 48-topic category.

Topics	ROUGE-2	ROUGE-SU4	BE
1	0.5026	0.5729	0.4094
3	0.7106	0.7532	0.6276
6	0.8130	0.8335	0.7512
12	0.8806	0.8834	0.8475
24	0.9196	0.9092	0.9063
36	0.9343*	0.9198*	0.9301*
39	0.9367*†	0.9213*†	0.9341*†
42	0.9386*†‡	0.9227*†‡	0.9376*†‡
45	0.9402*†‡§	0.9236*†‡§	0.9402†‡§
48 (bootstr)	0.9430†‡§	0.9280§	0.9444‡§
48 (actual)	0.9429	0.9263	0.9446

Table 10: Mean correlations of 48 topic Pyramid score and other metrics using from 1 to 48 topics for TAC 2008 update summaries. Values in each row are significantly different from each other at 95% level except: ROUGE-2 and ROUGE-SU4 in 12-topic category, ROUGE-2 and BE in 45-topic category.

39, 42, or 45. Within each of these categories, for a thousand repetitions, we calculated the score for automatic summarizers by averaging over *n* topics randomly selected from the pool of 48 topics available in the evaluation. Again, we examined the correlations between the metrics and the "full" 48-topic Responsiveness and Pyramid. As previously, we then used ANOVA to determine whether the correlation means differed significantly. Because there was only one possible sample with all 48 topics for each metric, we bootstrapped this sample to provide 1000 new samples in the 48-topic category, in order to perfom the ANOVA comparison of variance. Tables 7 and 8, as well as Figures 4 and 5, show the metrics' changing correlations with Responsiveness. Tables 9 and 10, and Figures 6 and 7, show the correlations with the 48-topic Pyramid score. Values for initial summaries are significantly different from their counterparts for update summaries at the 95% level.

In all cases, it becomes clear that the curves flatten out and the correlations stop increasing almost completely beyond the 36-topic mark. This means that the scores for the automatic summarization systems based on 36 topics will be on average practically indistiguishable from the scores based on all 48 topics, showing that beyond a certain minimally necessary number of topics adding or removing a few (or even ten) topics will not influence the system scores much. (However, we cannot conclude that a further considerable increase in the number of topics – well beyond 48 – would not bring more improvement in the correlations, perhaps increasing the stable "correlation window" as well.)

Topics	ROUGE-2	ROUGE-SU4	BE
1	0.6157	0.6378	0.5756
3	0.7597	0.7511	0.7323
6	0.8168	0.7904	0.7957
12	0.8493	0.8123	0.8306
24	0.8690	0.8249*	0.8517*
36	0.8751*	0.8287*†	0.8580*†
39	0.8761*†	0.8295*†‡	0.8592†‡
42	0.8768*†‡	0.8299*†‡§	0.8602†‡§
45 (bootstr)	0.8761*†‡	0.8305†‡§	0.8627†‡§
45 (actual)	0.8795	0.8301	0.8609

Table 11: Mean correlations of 45 topic Content Responsiveness and other metrics using from 1 to 45 topics for DUC 2007 summaries. Values in each row are significantly different from each other at 95% level.

An interesting observation is that if we produce such limited-topic scores for the manual metrics, Responsiveness and Pyramid, and correlate them with their own "full" versions based on

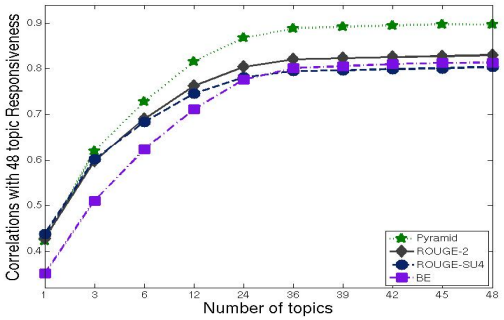

Figure 4: Responsiveness vs. other metrics with 1 to 48 topics for TAC 2008 initial summaries.

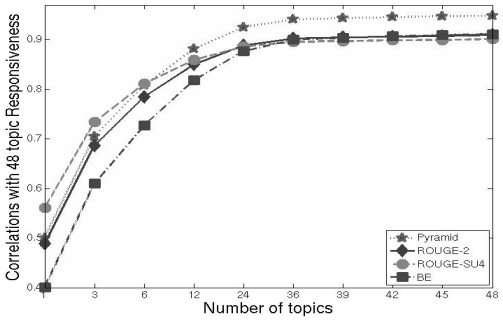

Figure 5: Responsiveness vs. other metrics with 1 to 48 topics for TAC 2008 update summaries.

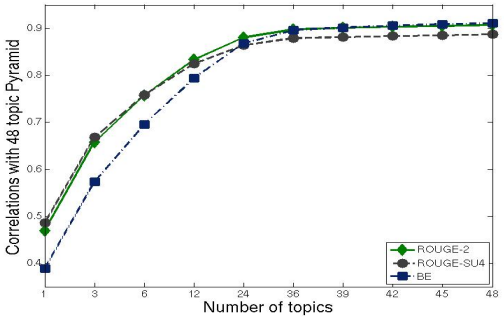

Figure 6: 48-topic Pyramid vs. other metrics with 1 to 48 topics for TAC 2008 initial summaries.

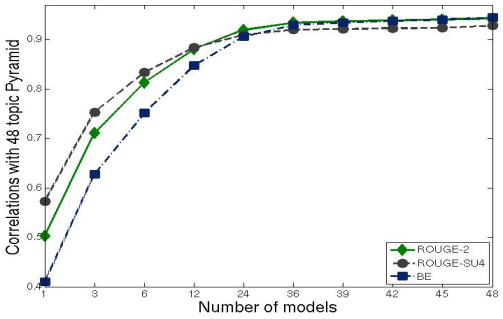

Figure 7: 48-topic Pyramid vs. other metrics with 1 to 48 topics for TAC 2008 update summaries.

all 48 topics, it appears that they are less stable than the automatic metrics, i.e. there is a larger gap between the worst and best correlations they achieve.[1] The mean correlation between the "full" Responsiveness and that based on 1 topic is 0.443 and 0.448 for the initial and update summaries, respectively; for that based on 3 topics, 0.664 and 0.667. Pyramid based on 1 topic achieves 0.467 for initial and 0.525 for update summaries; Pyramid based on 3 topics obtains 0.690 and 0.742, respectively. Some of these values, especially for update summaries, are even lower than those obtained by ROUGE in the same category, despite the fact that 1- and 3-topic Responsiveness or Pyramid is a proper subset of the 48-topic Responsiveness/Pyramid. On the other hand, ROUGE achieves considerably worse correlations with Responsiveness than Pyramid when there are many topics available. ROUGE-SU4 seems to be more stable than ROUGE-2; in all cases ROUGE-2 starts with lower correlations than ROUGE-SU4, but by the 12-topic mark its correlations increase above it.

Additionally, despite being an automatic metric, BE seems to follow the same pattern as the manual metrics. It is seriously affected by the decreasing number of topics; in fact, if the number of topics drops below 24, BE is the least reliable indicator of either Responsiveness or Pyramid. However, by the 48-topic mark it rises to levels comparable with ROUGE-2.

As in the case of models, DUC 2007 data shows mostly the same pattern as TAC 2008. Again, in this data set, the increase in the correlation mean with the addition of topics for each metric are smaller than for either initial or update summaries in TAC 2008, but the relative rate of increase remains the same: BE gains most from additional topics (+0.28 in DUC vs. +0.47 and +0.51 in TAC), ROUGE-SU4 again shows the smallest increase (+0.19 in DUC vs. +0.36 and +0.34 in TAC), which means it is the most stable of the metrics across the variable number of topics.[2]

[1] For reasons of space, these values are not included in the tables, as they offer little insight besides what is mentioned here.

[2] The smaller total increase might be due to the smaller number of available topics (45 in DUC vs. 48 in TAC), but we have seen the same effect in Section 3 while discussing models, so it might just be an accidental property of a given data set.

5 Discussion and conclusions

As the popularity of shared tasks increases, task organizers face an ever growing problem of providing an adequate evaluation to all participating teams. Often, evaluation of multiple runs from the same team is required, as a way to foster research and development. With more and more system submissions to judge, and the simultaneous need for multiple topics and models in order to provide a stable assessment, difficult decisions of cutting costs and effort might sometimes be necessary. It would be useful then to know where such decisions will have the smallest negative impact, or at least, what might be the trade-offs inherent in such decisions.

From our experiments, it appears that manual metrics such as Pyramid gain less from the addition of more model summaries than the automatic metrics. A Pyramid score based on any two models correlates very highly with the score based on any other two models. For the automatic metrics, the largest gain is recorded with adding the second model; afterwards the returns diminish. BE seems to be the most sensitive metric to changes in the number of models and topics; ROUGE-SU4, on the other hand, is the least sensitive to such changes and the most stable, but it does not obtain the highest correlations when many models and topics are available.

Whatever the number of models, manual Pyramid considerably outperforms automatic metrics, as can be expected, since human understanding is not hampered by the possible differences in surface expression between a candidate and a model. But when it comes to decreased number of topics, the inherent variability of human judgment shows strongly, to the extent that, in extreme cases of very few topics, it might be more prudent to use ROUGE-SU4 than Pyramid or Responsiveness.

Lastly, we observe that, as with models, adding one or two topics to the evaluation plays a great role only if we have very few topics to start with. Our experiments suggest that, as the number of topics available for evaluation increases, so does the number of additional topics necessary to make a difference in the system ranking produced by a metric. It seems that in the case of evaluation based on 48 topics, as in the TAC Summarization track, it would be possible to decrease the number to about 36 without sacrificing much stability.

References

Donna Harman and Paul Over. 2004. The effects of human variation in DUC summarization evaluation. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 10–17, Barcelona, Spain.

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating DUC 2005 using Basic Elements. In *Proceedings of the 5th Document Understanding Conference (DUC)*.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In *Proceedings of NTCIR Workshop 4*, Tokyo, Japan.

Chin-Yew Lin. 2004b. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL 2004 Workshop: Text Summarization Branches Out*, pages 74–81.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The Pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 145–152, Boston, MA.

Rebecca J. Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the Pyramid method in DUC 2005. In *Proceedings of the 5th Document Understanding Conference (DUC)*, Vancouver, Canada.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing and Management*, 43(6):1449–1481.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL DUC Workshop 2003*, pages 57–64, Edmonton, Canada.

Ellen M. Voorhees and Chris Buckley. 2002. Effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 317–323, Tampere, Finland.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the Machine Translation evaluation metrics. In *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 85–94, Baltimore, MD.