Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality

John M. Conroy
IDA/Center for Computing Sciences
Bowie, Maryland, USA
conroy@super.org

Hoa Trang Dang
Information Access Division
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
hoa.dang@nist.gov

Abstract

In this paper, we analyze the state of current human and automatic evaluation of topic-focused summarization in the Document Understanding Conference main task for 2005-2007. The analyses show that while ROUGE has very strong correlation with responsiveness for both human and automatic summaries, there is a significant gap in responsiveness between humans and systems which is not accounted for by the ROUGE metrics. In addition to teasing out gaps in the current automatic evaluation, we propose a method to maximize the strength of current automatic evaluations by using the method of canonical correlation. We apply this new evaluation method, which we call ROSE (ROUGE Optimal Summarization Evaluation), to find the optimal linear combination of ROUGE scores to maximize correlation with human responsiveness.

1 Introduction

ROUGE (Lin, 2004) and its linguistically-motivated descendent, Basic Elements (BE) (Hovy et al., 2005), evaluate a summary by computing its overlap with a set of model (human) summaries; ROUGE considers lexical n-grams as the unit for comparing the overlap between summaries, while Basic Elements uses larger units of comparison based on the output of syntactic parsers. The ROUGE/BE toolkit has become the standard automatic method for evaluating the content of machine-generated summaries, but the correlation of these automatic scores with human evaluation metrics has not always been consistent.

In this paper, we analyze the state of current human and automatic evaluation of topic-focused summarization. Using the results of the Document Understanding Conference main task for 2005-2007 we explore the correlation between variants of ROUGE and the human metrics of responsiveness and linguistic quality. The analyses expose a number of challenges and several surprising results. In particular, while ROUGE has very strong correlation with responsiveness for both human and system summaries, there is a significant gap in responsiveness between humans and systems which is not accounted for by the ROUGE metrics. One cause of the gap is that many automatic summarizers truncate the last sentence of their summary, which shows significant reduction in the responsiveness score but does not result in a statistically significant drop in ROUGE scores. In addition to teasing out gaps in the current automatic evaluation, we propose a method to maximize the strength of current automatic evaluations by using the method of canonical correlation. We apply this new evaluation method, which we call ROSE (ROUGE Optimal Summarization Evaluation), to find the optimal linear combination of ROUGE metrics to maximize correlation with human responsiveness.

2 DUC 2005-2007 Task and Evaluation

The main task for DUC 2005-2007 was a complex question-focused summarization task that required summarizers to piece together information from multiple documents to answer a question or set of questions as posed in a DUC topic statement. The topic statement was a request for infor-

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 145–152
Manchester, August 2008

mation that could not be met by just stating a name, date, quantity, etc. The summarization task was the same for both human and automatic summarizers: Given a topic statement and a set of 25-50 relevant newswire documents, the summarization task was to create from the documents a brief, well-organized, fluent summary that answered the need for information expressed in the topic statement. The summary could be no longer than 250 words. Summaries over the size limit were truncated, and no bonus was given for creating a shorter summary.

NIST Assessors developed the DUC topics used as test data. There were 50 DUC topics each year in 2005-2006, and 45 topics in DUC 2007. Each year, 10 NIST assessors produced a total of 4 human summaries for each of the topics. The assessor who developed a particular topic always wrote one of the 4 summaries for that topic.

NIST manually assessed each summary for both content and readability. Readability was assessed using a set of linguistic quality questions; summary content was assessed using the pseudo-extrinsic measure of *content responsiveness*.

All summaries for a given topic were judged by a single assessor who was usually the same as the topic developer. In all cases, the assessor was one of the summarizers for the topic. Assessors first judged each summary for a topic for *readability*, assigning a separate score for each of 5 linguistic qualities; each summary for the topic was then judged for *content responsiveness*. Each of these manual evaluations was based on a five-point scale (1=very poor, 5=very good), resulting in 6 scores for each summary.

2.1 Evaluation of Readability

The readability of the summaries was assessed using five linguistic quality questions which measured qualities of the summary that *did not* involve comparison with a reference summary or DUC topic. The linguistic qualities measured were *Q1: Grammaticality, Q2: Non-redundancy, Q3: Referential clarity, Q4: Focus, and Q5: Structure and coherence*.

2.2 Evaluation of Content

NIST performed manual pseudo-extrinsic evaluation of peer summaries in the form of assessment of responsiveness. Responsiveness differs from other measures of summary content such as SEE coverage (Lin and Hovy, 2002) and Pyramid scores (Nenkova and Passonneau, 2004) in that it does not compare a peer summary against a set of known human summaries. Rather, the assessor is given a list of randomly ordered, unlabeled summaries (both human and system-generated) for a topic, and must assign a responsiveness score to each summary (after having read all the summaries first).

In DUC 2005-2007, NIST assessors assigned a *content* responsiveness score to each summary; content responsiveness indicated the amount of information in the summary that helped to satisfy the information need expressed in the topic statement. For content responsiveness, the linguistic quality of the summary was to play a role in the assessment only insofar as it interfered with the expression of information and reduced the amount of information that was conveyed.

In DUC 2006, assessors assigned an additional *overall* responsiveness score, which was based on both information content and readability. Assessors judged overall responsiveness only *after* judging all their topics for readability and content responsiveness; however, they were not given direct access to these previously assigned scores, but were told to give their "gut" reaction to the overall responsiveness of each summary.

The content responsiveness score provides a coarse manual measure of information coverage; overall responsiveness reflects a combination of readability and content. Content responsiveness was largely responsible for determining how assessors perceived the overall quality of a summary, but readability also played an important role. While poor readability could downgrade the overall responsiveness of a summary that had very good content responsiveness, very good readability could sometimes bolster the overall responsiveness score of a less information-laden summary (Dang, 2006). Attempts at greater readability in 2006 paid off among the peers with the best overall responsiveness scores. However, the automatic peers generally had poor readability, and the average overall responsiveness for each peer was generally much lower than its average content responsiveness.

In addition to the human assessment of responsiveness, NIST computed three "official" automatic scores using ROUGE and Basic Elements: ROUGE-2, ROUGE-SU4, and ROUGE-BE recall. For the BE evaluation, summaries were parsed

with Minipar (Lin, 2005), and BE-F were extracted and matched using the Head-Modifier criterion. Jackknifing was used for each $[peer, topic]$ pair so that human and automatic peers could be compared.

3 An Analysis of the Metrics

Figure 1 shows the average scores for each summarizer for DUC 2005, 2006, and 2007. For each year we report the Pearson correlation coefficient for ROUGE-2, ROUGE-SU4, and ROUGE-BE (denoted ρ_{R2}, ρ_{SU4} and ρ_{BE}), against content responsiveness. This correlation is computed including just the systems as the human summarizers are clearly distributed differently.[1] To highlight the trend in the correlation we fit the systems data using robust linear regression. This line could be used to extrapolate the system performance if ROUGE scores were to increase.

As seen in Figure 1, while both the manual and the automatic ROUGE scores of the human summarizers remained relatively constant over the years, the systems made significant progress in their automatic scores, with the top systems performing within statistical confidence of the human summarizers in the ROUGE metrics as reported by Conroy et al. (2007). While the content responsiveness scores of the systems also increased as a group over the years, all systems performed significantly worse than humans in content responsiveness as measured by Tukey's honestly significant difference criterion (Conroy et al., 2007). Thus, there is not only a gap in performance between humans and systems on this task as measured manually by content responsiveness, but there is also a "metric gap" in using any single variant of ROUGE to predict content responsiveness. This metric gap becomes more pronounced as system performance improves to the point where ROUGE is unable to distinguish between systems and humans.

We turn next to an analysis of sources of the performance gap and "metric gap". Responsiveness is a subjective measure, and because NIST uses the same humans both to generate abstracts and to evaluate the abstracts, there is the possibility that humans may give high scores to their own abstract

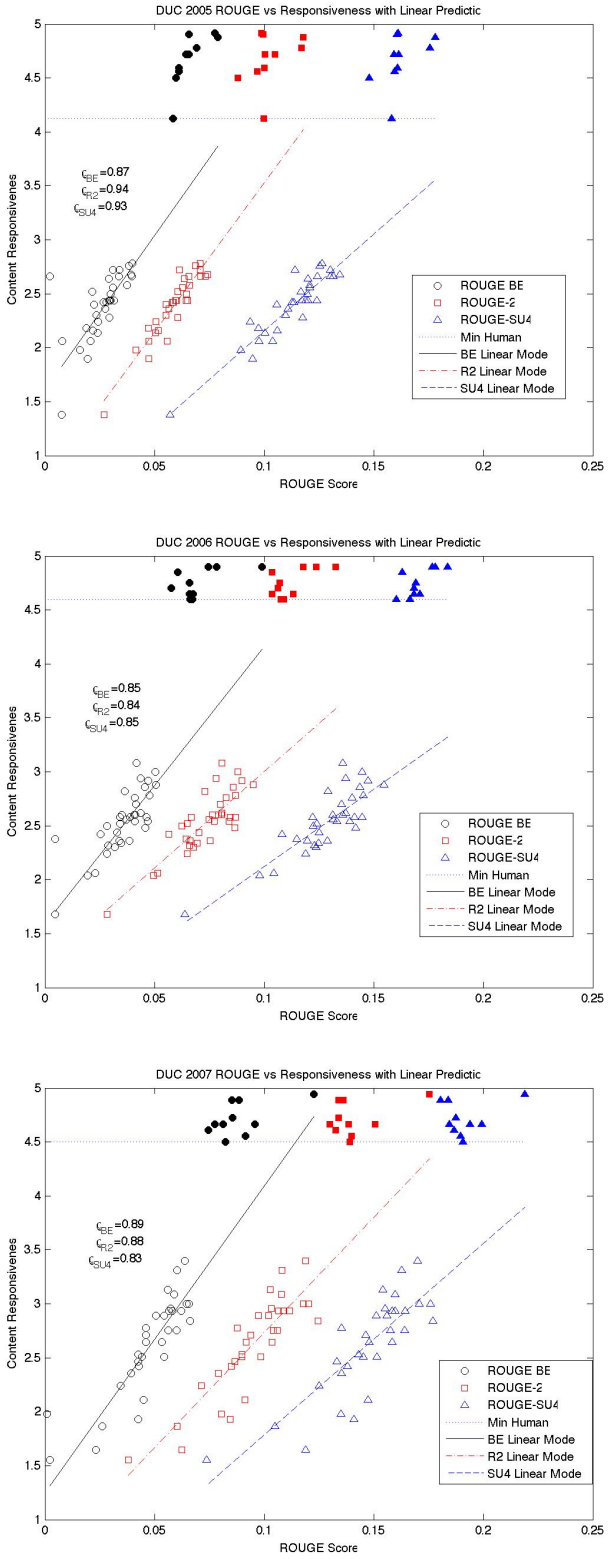

Figure 1: Scatter plot of average manual content responsiveness vs. automatic ROUGE scores (ROUGE-BE, ROUGE-2, and ROUGE-SU4) for humans (filled points) and systems (unfilled points), for DUC 2005-2007.

[1] One system each year in 2005-2007 had formatting problems in their summaries which resulted in abnormally low ROUGE-BE scores. While these systems are included in the scatter plot, they are not included in the correlation coefficient computation.

just because it not surprisingly "says what they would say." To test this hypothesis we performed one-side Student $T-$test, testing if the group of "Self-Assessed" abstracts had significantly higher responsiveness for DUC 2005-2007. Indeed, as Table 1 shows in each year and for both content and overall responsiveness, humans gave significantly higher scores to their own abstracts than the other human abstracts. This bias adds to the gap in content responsiveness between the human and automatic summarizers. Fortunately, this effect is dampened by the fact that NIST used 10 assessors and on average a human got to assess their own abstract only 25% of the time. It is noteworthy to add that at the Multi-lingual Summarization Evaluation of 2006, the human assessors were not the abstractors. This and other factors, notably an easier task, lead to there being no gap in performance between the human and the top scoring system (Schlesinger et al., 2008).

Table 1: Mean responsiveness assessment by humans for their own (Self) vs. Other abstracts.

Data	Self	Other	Signif
DUC 2005 Content	4.88	4.61	0.00277
DUC 2006 Content	4.96	4.68	0.00052
DUC 2006 Overall	4.94	4.67	0.00326
DUC 2007 Content	4.87	4.65	0.01931

We next examine the correlation between responsiveness and each of the five (manual) metrics for linguistic quality. We divide the correlation into three groups: Human (the group of 10 human summarizers), Systems (the automatic systems entered into DUC), and Combined (the union of these two groups). Table 2 gives the Pearson correlation coefficient and the $p-$value of statistical significance between content responsiveness and each of the five linguistic quality questions for DUC 2005-2007. For DUC 2005 there is no significant correlation between the average score of a human or automatic summarizer on linguistic questions and the content responsiveness score. The fact that there is a significant correlation in the "Combined" case is primarily due to the fact that the human summarizers scored higher as a group than the systems in the content metric as well as the linguistic metrics.

In DUC 2006 and 2007, the linguistic question which rewards summaries for not having redundancy (Q2) has a significant *negative* correlation with content responsiveness in the group of systems. This negative correlation is due largely to the fact that a number of low scoring systems (including the baseline) have no significant redundancy. Rarely does any system have sentences which are near duplicates. However, many systems, even those with relatively high responsiveness scores, still suffer from clause level redundancy, much of it in the form of noun phrases for which a human summarizer would employ pronouns.

Table 3 gives additional correlations between *overall* responsiveness and the linguistic questions for DUC 2006. We contrast the correlations for DUC 2006 in Table 2 vs. those in Table 3. Not surprisingly, overall responsiveness, which intentionally penalizes summaries for linguistic problems, does correlate more strongly with the linguistic questions than content responsiveness. Also, we note that the DUC 2007 correlations for *content* responsiveness appear more like those for DUC 2006 *overall* responsiveness than the corresponding correlations for DUC 2006 content responsiveness. NIST did not have sufficient time in 2007 to perform an overall responsiveness evaluation. We hypothesize that the assessors, many of whom worked on DUC 2006, may have inadvertently taken linguistic quality into account more in 2007 than in 2006 for the content responsiveness, since only one measure was done in 2007.

Finally, it was hypothesized at the DUC 2006 workshop[2] that the human assessors penalize systems in content responsiveness which end with a sentence fragment, more than could be accounted for by the missing content of the sentence fragment. We tested the hypothesis by comparing the average grammaticality (Q1), content responsiveness, and ROUGE scores of the 15 systems in DUC 2007 that ended their summaries with a complete sentence, against the 17 systems whose summaries ended with a sentence fragment. Table 4 gives a summary of the results. As measured by a Student $T-$test, systems that ended their summaries with a complete sentence had significantly higher content responsiveness scores than those that did not; however, there was no significant difference in ROUGE scores. The table lists ROUGE-2 as an example; these results are consistent with both ROUGE-BE and ROUGE-SU4.

Because linguistic quality clearly influences content responsiveness, automatic methods of evaluating summary content that try to maximize

[2]Lucy Vanderwende, personal communication

Table 2: Correlation and p-values between Content Responsiveness and Linguistic Quality Questions, DUC 2005-2007

Year	Group	Q1 Grammar	Q2 Non-redund.	Q3 Refer. Clarity	Q4 Focus	Q5 Structure/Coherence
2005	Humans	-0.10(0.78)	0.03(0.94)	0.06(0.87)	0.23(0.53)	0.31(0.39)
2005	Systems	-0.05(0.79)	0.15(0.42)	0.19(0.29)	0.30(0.10)	0.08(0.66)
2005	Combined	0.72(0.00)	0.75(0.00)	0.87(0.00)	0.90(0.00)	0.91(0.00)
2006	Humans	0.26(0.47)	0.15(0.69)	0.04(0.91)	0.64(0.05)	0.40(0.26)
2006	Systems	0.33(0.05)	-0.38(0.03)	0.27(0.11)	0.41(0.01)	0.16(0.35)
2006	Combined	0.74(0.00)	0.68(0.00)	0.86(0.00)	0.87(0.00)	0.89(0.00)
2007	Humans	0.80(0.01)	0.73(0.02)	0.24(0.51)	0.57(0.09)	0.47(0.17)
2007	Systems	0.60(0.00)	-0.43(0.01)	0.59(0.00)	0.71(0.00)	0.49(0.00)
2007	Combined	0.77(0.00)	0.72(0.00)	0.85(0.00)	0.92(0.00)	0.90(0.00)

Table 3: Correlation and p-values between Overall Responsiveness and Linguistic Quality Questions, DUC 2006

Group	Q1 Grammar	Q2 Non-redund.	Q3 Refer. Clarity	Q4 Focus	Q5 Structure/Coherence
Humans	0.60(0.06)	0.27(0.45)	0.39(0.26)	0.74(0.01)	0.82(0.00)
Systems	0.49(0.00)	-0.23(0.19)	0.55(0.00)	0.64(0.00)	0.49(0.00)
Combined	0.77(0.00)	0.72(0.00)	0.89(0.00)	0.89(0.00)	0.93(0.00)

Table 4: Average scores of DUC 2007 systems ending with a complete sentence vs. those ending with a fragment.

Metric	Sentence	Fragment	Signif
Grammaticality	3.88	3.24	0.011
Content Resp.	2.79	2.46	0.021
ROUGE-2	0.098	0.092	0.408

correlation with content responsiveness should attempt to include some measures of linguistic quality. We hypothesize that different variants of ROUGE may capture different qualities of a summary; for example, ROUGE-1 may be a good indicator of the relevance of summary content, but ROUGE variants that take into account larger contexts may capture linguistic qualities of the summary. Hence, a combination of scores (including measures of linguistic quality) would be a better predictor of "content" responsiveness.[3] In the next section, we present a new evaluation metric that finds a linear combination of ROUGE metrics which, in general, has stronger correlation with content responsiveness than any of the current ROUGE metrics.

[3] An additional weakness in the automatic metrics, which we do not attempt to address in our current work, is their inability to adequately handle the generalizations that are often made in model summaries (Dang, 2006), which are abstractive as opposed to the extractive summaries of most systems.

4 ROSE: Un Melange de ROUGEs

We developed an automatic content evaluation model which combines multiple ROUGE scores using canonical correlation (Hotelling, 1935). Canonical correlation finds the linear combination of ROUGE scores that has maximum correlation with human responsiveness on a given data set. As this family of models is a "blend" of ROUGE scores we call this metric ROSE, for ROUGE Optimal Summarization Evaluation. We first apply canonical correlation for each year of DUC using a Monte Carlo method. We then report on preliminary experiments that use ROSE models from one year to predict content responsiveness in subsequent years.

4.1 Blending ROUGE Scoring with a Canonical Correlation Model

Suppose we are given a set of ROUGE scores and the corresponding content responsiveness scores.

We let a_{ij}, for $i = 1, ..., m$ and $j = 1, .., n$, be the ROUGE score of type j for the summarizer i, and b_i the human content evaluation metric. Canonical correlation finds an $n-$long vector x such that

$$x = \text{argmax } \rho(\sum_{j=1}^{n} a_{ij}x_j, b_i), \quad (1)$$

where $\rho(x, y)$ is the Pearson correlation between x and y. A similar approach has been used by Liu and Gildea (2007) in the application of machine translation metrics, where they use a gradient optimization method to solve the maximization problem.

Canonical correlation actually solves a more general correlation optimization problem, where the goal is to find two linear combinations of variables to maximize the correlation between two sub-spaces. In the application of document summarization, we may wish to consider a matrix B of human evaluation metrics where b_{ij} is the $j-$th human evaluation for the $i-$th summarizer. We could include, for example, content and overall responsiveness or linguistic questions. Here we solve for (x, y) in the equation below:

$$(x, y) = \text{argmax } \rho(\sum_{j=1}^{n} a_{ij}x_j, \sum_{j=1}^{k} b_{ij}y_j). \quad (2)$$

This maximization procedure can be solved via a generalized eigenvalue problem, which we computed in Matlab using a routine distributed by Borga (2000). For the case studied here, as given in Equation (1), the generalized eigenvalue reduces to a linear least squares problem.

To find strong canonical correlations we decided to explore a large space of metrics. To this end, we included in our optimization 7 ROUGE automatic metrics: ROUGE-1,2,3,4,L,SU4, and BE to predict content responsiveness and (for DUC 2006) overall responsiveness. As our analyses of the previous section indicated for DUC 2006 and 2007 there was a significant correlation between the linguistic questions and content responsiveness. We add questions 1 and 4 to our canonical correlation model to see to what extent these questions could improve the correlation with content responsiveness. While the linguistic questions evaluation scores are manually generated we combine them with the automatic methods of ROUGE in an attempt see to what extent these non-content scores can better model both content and overall responsiveness. Thus, in all we consider 9 variables to predict responsiveness. In order to perform an evaluation that would avoid over-fitting the data we used a Monte Carlo method of resampling to evaluate which of the $2^9 - 1 = 511$ combinations of variables (canonical variates) to include in the model. [4]

In each experiment of the Monte Carlo method we randomly held back 1/4 of the data (human and system summarizers) for testing and used 3/4 of the data to build a canonical variate model. We found 4000 random samples sufficient to achieve accuracy within at least 2 digits. For each of the canonical variate models, 4000 trials are performed and then the computed model is applied to the held-back portion of the data and its Pearson correlation and p-value is reported. These 4000 correlations (and p-values) are then used to estimate the median correlation for a canonical variate. The median is computed from the subset of 4000 experiments with statistically significant correlations on the testing data (95% confidence, a p-value less then 0.05). The canonical variate with the highest estimated median correlation is then compared with the best performing ROUGE method. We compare the best of 504=511-7 canonical variates with the best of the 7 ROUGE variants by using the Mann-Whitney U-test, which tests for equal medians.

The procedure is then repeated using only the systems to find the ROSE model that gives the best prediction for just machine summarizers.

Table 5 gives the results of the Monte Carlo experiments. In each case the best canonical variate and the estimated median correlation are reported over the set of ROUGE scores and the ROUGE scores in union with the linguistic questions. As these results are based on 4000 trials they are more reliable than the simple correlation analysis done using the three official DUC automatic metrics, ROUGE-2, SU4, and BE. We note, in particular, that occasionally ROUGE-1 and ROUGE-L were found to be the best predictor even when linguistic questions were allowed in the model. Not surprisingly, the human evaluation of overall responsiveness was harder to predict and the optimal variants included both linguistic questions 1 and 4.

The ROSE models give the best combinations

[4] We also removed one system each year that had a poor ROUGE-BE score due to formatting problems.

Table 5: Monte Carlo Results for Canonical Correlation Model. A * by a variant indicates that it differs significantly from the best single ROUGE correlation with a p-value of 10^{-7} or less as measured by a Mann Whitney U-test.

Year Metric	Summarizer	Best ROUGE	Corr.	ROSE$_{ROUGE}$	Corr.	ROSE$_{(ROUGE,Q)}$	Corr.
2005 Content	All	BE	0.976	R1,R2,R4,SU4,BE*	0.981	R1,R2,R3,RL,SU4,BE,Q4*	0.986
2005 Content	Systems	R2	0.939	R1,R2,RL	0.940	R2,RL,SU4,Q4	0.941
2006 Content	All	RL	0.928	R1,R2,R3,R4*	0.942	RL,Q1*	0.960
2006 Content	Systems	R1	0.900	R1	0.900	R1	0.900
2007 Content	All	BE	0.937	R1,R4,RL,BE	0.940	BE,Q4*	0.966
2007 Content	Systems	R3	0.906	RL,BE*	0.915	R1,RL,BE,Q1*	0.929
2006 Overall	All	BE	0.893	R1,R2,R3,R4*	0.913	R3,R4,Q1,Q4*	0.946
2006 Overall	Systems	RL	0.854	RL	0.854	R1,R3,SU4,Q1,Q4*	0.894

of ROUGE scores to give maximum correlation with the human judgement of content or overall responsiveness. The ROSE models based on just ROUGE for the automatic summarizers are an appropriate method to use to compare systems that did not compete in DUC with those that did.

4.2 Applying ROSE across the Years

To further evaluate the generality of the ROSE model we apply DUC 2005 canonical correlation models to DUC 2006 and DUC 2007, and similarly apply the DUC 2006 model to the DUC 2007 data. In these experiments we measure the stability of a ROSE model from one year to the next. (Note, we have also computed a model based on the combined data of DUC 2005 and DUC 2006 for use with DUC 2007 and these results are comparable to those presented.) Here, for simplicity, we restrict the ROSE model to use only the "official" ROUGE metrics to build a model based on a given year and then evaluate that model on a subsequent year. Table 6 gives results for ROSE models constructed from only ROUGE-2, ROUGE-SU4, ROUGE-BE, and content responsiveness to create the ROSE model for each year; results are also given for ROSE models (ROSE$_{+Q1,4}$) which also includes the linguistic questions on grammaticality (Q1) and focus (Q4).

The ROSE models built from only ROUGE scores had mixed results, sometimes performing worse than a single ROUGE score (e.g., the ROSE model trained on DUC 2005 and evaluated on DUC 2006), but in other cases performing as well as or better than single ROUGE scores. These preliminary results with ROSE illustrate the difficulty in finding a single canonical variate that can be used from year to year to build ROSE models based on previous years' data. We hypothesize that the task is made more difficult due to humans changing their criteria for judging content responsiveness over the years.

On the other hand, ROSE$_{+Q1,4}$ models that included the linguistic questions Q1 and Q4 always yielded the best correlation with content responsiveness both for the systems and for the group of combined systems and human summarizers.

5 Conclusions

We analyzed the results of the topic-focused summarization task using the data from DUC 2005-2007. Our main concern was to expose causes of the gap that currently exists between automatic and human evaluation of summary content. As the automatic ROUGE scores of system summaries approaches that of human summaries, the disparity between automatic and manual measures of summary content becomes a more important concern. We find that there is a slight bias in the human evaluation: humans give their own summaries significantly higher scores. Furthermore, the responsiveness metric appears to be time varying, i.e., the humans changed their standards for judging responsiveness over the years, making it difficult to use automatic scores from one year to predict responsiveness in another year.

Assessors naturally tend toward taking linguistic quality into account when assessing summaries. The instructions for assessing content responsiveness implicitly acknowledges this; what is surprising is the extent to which linguistic quality does influence content responsiveness. In particular, we demonstrated that content responsiveness in DUC 2006 and 2007 correlated with the linguistic quality questions of grammar (Q1) and focus (Q4), and that systems were significantly penalized in content responsiveness when their summary ended

Table 6: Correlation and p-values between content responsiveness and various metrics for each "Test" year of DUC. ROSE models were constructed using DUC data from "Train" year and evaluated on data from "Test" year.

Train/Test	Summarizer	R2	SU4	BE	Q1	Q4	ROSE	ROSE$_{+Q1,4}$
2005/2006	Humans	0.64(0.05)	0.69(0.03)	0.57(0.09)	0.26(0.47)	0.64(0.05)	0.59 (0.07)	0.61(0.06)
2005/2006	Systems	0.83(0.00)	0.85(0.00)	0.85(0.00)	0.33(0.06)	0.41(0.02)	0.83 (0.00)	0.85(0.00)
2005/2006	All	0.90(0.00)	0.88(0.00)	0.90(0.00)	0.74(0.00)	0.87(0.00)	0.90 (0.00)	0.93(0.00)
2005/2007	Humans	0.41(0.24)	0.26(0.47)	0.55(0.10)	0.80(0.01)	0.57(0.09)	0.53 (0.12)	0.57(0.09)
2005/2007	Systems	0.88(0.00)	0.84(0.00)	0.89(0.00)	0.56(0.00)	0.68(0.00)	0.90 (0.00)	0.92(0.00)
2005/2007	All	0.91(0.00)	0.88(0.00)	0.92(0.00)	0.77(0.00)	0.92(0.00)	0.92 (0.00)	0.94(0.00)
2006/2007	Humans	0.41(0.24)	0.26(0.47)	0.55(0.10)	0.80(0.01)	0.57(0.09)	0.52(0.12)	0.67(0.03)
2006/2007	Systems	0.88(0.00)	0.84(0.00)	0.89(0.00)	0.56(0.00)	0.68(0.00)	0.89 (0.00)	0.90(0.00)
2006/2007	All	0.91(0.00)	0.88(0.00)	0.92(0.00)	0.77(0.00)	0.92(0.00)	0.92(0.00)	0.96(0.00)

with a sentence fragment even though the automatic content measures did not show a statistically significant difference. The influence of linguistic quality on "content" responsiveness contributes to the evaluation gap that we see between ROUGE/BE and this coarse human measure of summary content.

Automatic methods of evaluating summary content that try to maximize correlation with content responsiveness should therefore attempt to include some measures of linguistic quality. We found that a blending of ROUGE scores using canonical correlation gave higher correlations with content and overall responsiveness. When the linguistic questions Q1 and Q4 were added to the ROSE model, correlations of up to 0.96 were observed. This result leads to a natural question: What automatic methods could be used to approximate the linguistic questions? The work of Barzilay and Lapata (2005) on local coherence might be a possible candidate for estimating focus (Q4), while an automatic parser could be run on the summaries and the induced score could be used as a surrogate for grammaticality (Q1).

References

Barzilay, Regina and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Borga, Magnus. 2000. Matlab function cca().

Conroy, John M., Judith D. Schlesinger, and Dianne P. O'Leary. 2007. CLASSY 2007 at DUC 2007. In *Proceedings of the Seventh Document Understanding Conference (DUC)*, Rochester, New York.

Dang, Hoa Trang. 2006. Overview of DUC 2006. In *Proceedings of the Sixth Document Understanding Conference (DUC)*, New York City, New York.

Hotelling, H. 1935. The most predictable criterion. *Journal of Educational Psychology*, 26:139–142.

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

Lin, Chin-Yew and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Philadelphia, PA.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Lin, Dekang. 2005. A dependency-base method for evaluating broad-coverage parsers. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland.

Liu, Ding and Daniel Gildea. 2007. Source-language features and maximum correlation training for machine translation evaluation. In *Proceedings of the 2007 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-07)*.

Nenkova, Ani and Rebecca 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.

Schlesinger, Judith D., Dianne P. O'Leary, and John M. Conroy. 2008. Arabic/English multi-document summarization with CLASSY—the past and the future. In *Conference on Intelligent Text Processing and Computational Linguistics 2008*. Lecture Notes in Computer Science, Springer-Verlag. to appear.