Assessing the Effect of Inconsistent Assessors on Summarization Evaluation

Karolina Owczarzak
National Institute of Standards and Technology
Gaithersburg, MD 20899
karolina.owczarzak@gmail.com

Peter A. Rankel
University of Maryland
College Park, Maryland
rankel@math.umd.edu

Hoa Trang Dang
National Institute of Standards and Technology
Gaithersburg, MD 20899
hoa.dang@nist.gov

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

Abstract

We investigate the consistency of human assessors involved in summarization evaluation to understand its effect on system ranking and automatic evaluation techniques. Using Text Analysis Conference data, we measure annotator consistency based on human scoring of summaries for Responsiveness, Readability, and Pyramid scoring. We identify inconsistencies in the data and measure to what extent these inconsistencies affect the ranking of automatic summarization systems. Finally, we examine the stability of automatic metrics (ROUGE and CLASSY) with respect to the inconsistent assessments.

1 Introduction

Automatic summarization of documents is a research area that unfortunately depends on human feedback. Although attempts have been made at automating the evaluation of summaries, none is so good as to remove the need for human assessors. Human judgment of summaries, however, is not perfect either. We investigate two ways of measuring evaluation consistency in order to see what effect it has on summarization evaluation and training of automatic evaluation metrics.

2 Assessor consistency

In the Text Analysis Conference (TAC) Summarization track, participants are allowed to submit more than one run (usually two), and this option is often used to test different settings or versions of the same summarization system. In cases when the system versions are not too divergent, they sometimes produce identical summaries for a given topic. Summaries are randomized within each topic before they are evaluated, so the identical copies are usually interspersed with 40-50 other summaries for the same topic and are not evaluated in a row. Given that each topic is evaluated by a single assessor, it then becomes possible to check assessor consistency, i.e., whether the assessor judged the two identical summaries in the same way.

For each summary, assessors conduct content evaluation according to the Pyramid framework (Nenkova and Passonneau, 2004) and assign it Responsiveness and Readability scores[1], so assessor consistency can be checked in these three areas separately. We found between 230 (in 2009) and 430 (in 2011) pairs of identical summaries for the 2008-2011 data (given on average 45 topics, 50 runs, and two summarization conditions: main and update), giving in effect anywhere from around 30 to 60 instances per assessor per year. Using Krippendorff's *alpha* (Freelon, 2004), we calculated assessor consistency within each year, as well as total consistency over all years' data (for those assessors who worked multiple years). Table 1 shows rankings of assessors in 2011, based on their Readability, Responsiveness, and Pyramid judgments for identical summary pairs (around 60 pairs per assessor).

Interestingly, consistency values for Readability are lower overall than those for Responsiveness and Pyramid, even for the most consistent assessors. Given that Readability and Responsiveness are evaluated in the same way, i.e. by assigning a numerical score according to detailed guidelines, this sug-

[1] http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 359–362,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

ID	Read	ID	Resp	ID	Pyr
G	0.867	G	0.931	G	0.975
D	0.866	D	0.875	D	0.970
A	0.801	H	0.808	H	0.935
H	0.783	A	0.750	A	0.931
F	0.647	F	0.720	E	0.909
C	0.641	E	0.711	C	0.886
E	0.519	C	0.490	F	0.872

Table 1: Annotator consistency in assigning Readability and Responsiveness scores and in Pyramid evaluation, as represented by Krippendorff's *alpha* for interval values, on 2011 data.

gests that Readability as a quality of text is inherently more vague and difficult to pinpoint.

On the other hand, Pyramid consistency values are generally the highest, which can be explained by how the Pyramid evaluation is designed. Even if the assessor is inconsistent in selecting Summary Content Units (SCUs) across different summaries, as long as the total summary weight is similar, the summary's final score will be similar, too.[2] Therefore, it would be better to look at whether assessors tend to find the same SCUs (information "nuggets") in different summaries on the same topic, and whether they annotate them consistently. This can be done using the "autoannotate" function of the Pyramid process, where all SCU contributors (selected text strings) from already annotated summaries are matched against the text of a candidate (un-annotated) summary. The autoannotate function works fairly well for matching between extractive summaries, which tend to repeat verbatim whole sentences from source documents.

For each summary in 2008-2011 data, we autoannotated it using all remaining manually-annotated summaries from the same topic, and then we compared the resulting "autoPyramid" score with the score from the original manual annotation for that summary. Ideally, the autoPyramid score should be lower or equal to the manual Pyramid score: it would mean that in this summary, the assessor selected as relevant all the same strings as s/he found in the other summaries on the same topic, plus possibly some more information that did not appear anywhere else. If the autoPyramid score is higher than the manual Pyramid score, it means that either (1) the assessor missed relevant strings in this summary, but found them in other summaries; or (2) the strings selected as relevant elsewhere in the topic were accidental, and as such not repeated in this summary. Either way, if we then average out score differences for all summaries for a given topic, it will give us a good picture of the annotation consistency in this particular topic. Higher average autoPyramid scores suggest that the assessor was missing content, or otherwise making frequent random mistakes in assigning content. Figure 1 shows the macro-average difference between manual Pyramid scores and autoPyramid scores for each assessor in 2011.[3] For the most part, it mirrors the consistency ranking from Table 1, confirming that some assessors are less consistent than others; however, certain differences appear: for instance, Assessor A is one of the most consistent in assigning Readability scores, but is not very good at selecting SCUs consistently. This can be explained by the fact that the Pyramid evaluation and assigning Readability scores are different processes and might require different skills and types of focus.

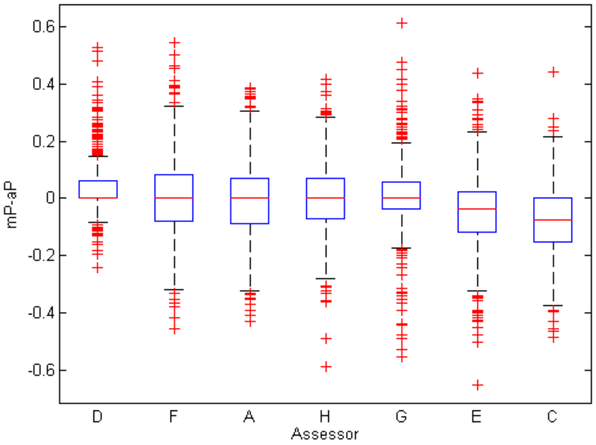

Figure 1: Annotator consistency in selecting SCUs in Pyramid evaluation, as represented by the difference between manual Pyramid and automatic Pyramid scores (mP-aP), on 2011 data.

3 Impact on evaluation

Since human assessment is used to rank participating summarizers in the TAC Summarization track,

[2] The final score is based on total weight of all SCUs found in the summary, so the same weight can be obtained by selecting a larger number of lower-weight SCUs or a smaller number of higher-weight SCUs (or the same number of similar-weight SCUs which nevertheless denote different content).

[3] Due to space constraints, we report figures for only 2011, but the results for other years are similar.

	Pearson's *r*		Spearman's *rho*	
	-1 worst	-2 worst	-1 worst	-2 worst
Readability	0.995	0.993	0.988	0.986
Responsiveness	0.996	0.989	0.986	0.946
Pyramid	0.996	0.992	0.978	0.960
mP-aP	0.996	0.987	0.975	0.943

Table 2: Correlation between the original summarizer ranking and the ranking after excluding topics by one or two worst assessors in each category.

	Readability	Responsiveness	Pyramid	mP-aP
before	0.705	0.930	0.954	0.954
-1 worst	0.718	0.921	0.961	0.942
-2 worst	0.718	0.904	0.952	0.923

Table 3: Correlation between the summarizer rankings according to ROUGE-2 and human metrics, before and after excluding topics by one or two worst assessors in that category.

	Readability	Responsiveness	Pyramid	mP-aP
before	0.579	0.694	0.771	0.771
-1 worst	0.626	0.695	0.828	0.752
-2 worst	0.628	0.721	0.817	0.741

Table 4: Correlation between ROUGE-2 and human metrics on a summary level before and after excluding topics by one or two worst assessors in that category.

we should examine the potential impact of inconsistent assessors on the overall evaluation. Because the final summarizer score is the average over many topics, and the topics are fairly evenly distributed among assessors for annotation, excluding noisy topics/assessors has very little impact on summarizer ranking. As an example, consider the 2011 assessor consistency data in Table 1 and Figure 1. If we exclude topics by the worst performing assessor from each of these categories, recalculate the summarizer rankings, and then check the correlation between the original and newly created rankings, we obtain results in Table 2.

Although the impact on evaluating automatic *summarizers* is small, it could be argued that excluding topics with inconsistent human scoring will have an impact on the performance of automatic *evaluation metrics*, which might be unfairly penalized by their inability to emulate random human mistakes. Table 3 shows ROUGE-2 (Lin, 2004), one of the state-of-the-art automatic metrics used in TAC, and its correlations with human metrics, before and after exclusion of noisy topics from 2011 data. The results are fairly inconclusive: it seems that in most cases, removing topics does more harm than good, suggesting that the signal-to-noise ratio is still tipped in favor of signal. The only exception is Readability, where ROUGE records a slight increase in correlation; this is unsurprising, given that consistency values for Readability are the lowest of all categories, and perhaps here removing noise has more impact. In the case of Pyramid, there is a small gain when we exclude the single worst assessor, but excluding two assessors results in a decreased correlation, perhaps because we remove too much valid information at the same time.

A different picture emerges when we examine how well ROUGE-2 can predict human scores on the *summary* level. We pooled together all summaries annotated by each particular assessor and calculated the correlation between ROUGE-2 and this assessor's manual scores for individual summaries. Then we calculated the mean correlation over all assessors. Unsurprisingly, inconsistent assessors tend to correlate poorly with automatic (and therefore always consistent) metrics, so excluding one or two worst assessors from each category increases ROUGE's average per-assessor summary-level correlation, as can be seen in Table 4. The only exception here is when we exclude assessors based on their autoPyramid performance: again, because inconsistent SCU selection doesn't necessarily translate into inconsistent final Pyramid scores, excluding those assessors doesn't do much for ROUGE-2.

4 Impact on training

Another area where excluding noisy topics might be useful is in training new automatic evaluation metrics. To examine this issue we turned to CLASSY (Rankel et al., 2011), an automatic evaluation metric submitted to TAC each year from 2009-2011. CLASSY consists of four different versions, each aimed at predicting a particular human evaluation score. Each version of CLASSY is based on one of three regression methods: robust regression, non-negative least squares, or canonical correlation. The regressions are calculated based on a collection of linguistic and content features, derived from the summary to be scored.

CLASSY requires two years of marked data to score summaries in a new year. In order to predict the human metrics in 2011, for example, CLASSY uses the human ratings from 2009 and 2010. It first considers each subset of the features in turn, and using each of the regression methods, fits a model to the 2009 data. The subset/method combination that best predicts the 2010 scores is then used to predict scores for 2011. However, the model is first retrained on the 2010 data to calculate the coefficients to be used in predicting 2011.

First, we trained all four CLASSY versions on all available 2009-2010 topics, and then trained again excluding topics by the most inconsistent assessor(s). A different subset of topics was excluded depending on whether this particular version of CLASSY was aiming to predict Responsiveness, Readability, or the Pyramid score. Then we tested CLASSY's performance on 2011 data, ranking either automatic summarizers (NoModels case) or human and automatic summarizers together (AllPeers case), separately for main and update summaries, and calculated its correlation with the metrics it was aiming to predict. Table 5 shows the result of this comparison. For Pyramid, (a) indicates that excluded topics were selected based on Krippendorff's *alpha*, and (b) indicates that topics were excluded based on their mean difference between manual and automatic Pyramid scores.

The results are encouraging; it seems that removing noisy topics from training data does improve the correlations with manual metrics in most cases. The greatest increase takes place in CLASSY's correlations with Responsiveness for main summaries in AllPeers case, and for correlations with Readability. While none of the changes are large enough to achieve statistical significance, the pattern of improvement is fairly consistent.

	NoModels		AllPeers	
	main	update	main	update
Pyramid				
CLASSY1_Pyr	0.956	0.898	0.945	0.936
CLASSY1_Pyr_new (a)	0.950	0.895	0.932	**0.955**
CLASSY1_Pyr_new (b)	**0.960**	**0.900**	0.940	**0.955**
Responsiveness				
CLASSY2_Resp	0.951	0.903	0.948	0.963
CLASSY2_Resp_new	**0.954**	**0.907**	**0.973**	0.950
CLASSY4_Resp	0.951	0.927	0.830	0.949
CLASSY4_Resp_new	0.943	**0.928**	**0.887**	0.946
Readability				
CLASSY3_Read	0.768	0.705	0.844	0.907
CLASSY3_Read_new	**0.793**	**0.721**	**0.858**	0.906

Table 5: Correlations between CLASSY and human metrics on 2011 data (main and update summaries), before and after excluding most inconsistent topic from 2009-2010 training data for CLASSY.

5 Conclusions

We investigated the consistency of human assessors in the area of summarization evaluation. We considered two ways of measuring assessor consistency, depending on the metric, and studied the impact of consistent scoring on ranking summarization systems and on the performance of automatic evaluation systems. We found that summarization system ranking, based on scores for multiple topics, was surprisingly stable and didn't change significantly when several topics were removed from consideration. However, on a summary level, removing topics scored by the most inconsistent assessors helped ROUGE-2 increase its correlation with human metrics. In the area of training automatic metrics, we found some encouraging results; removing noise from the training data allowed most CLASSY versions to improve their correlations with the manual metrics that they were aiming to model.

References

Deen G. Freelon. 2010. ReCal: Intercoder Reliability Calculation as a Web Service. *International Journal of Internet Science*, Vol 5(1).

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 78–81. Barcelona, Spain.

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

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

Peter A. Rankel, John M. Conroy, and Judith D. Schlesinger. 2012. Better Metrics to Automatically Predict the Quality of a Text Summary. *Proceedings of the SIAM Data Mining Text Mining Workshop 2012*.