Non-Expert Evaluation of Summarization Systems is Risky

Dan Gillick
University of California, Berkeley
Computer Science Division
dgillick@cs.berkeley.edu

Yang Liu
University of Texas, Dallas
Department of Computer Science
yangl@hlt.utdallas.edu

Abstract

We provide evidence that intrinsic evaluation of summaries using Amazon's Mechanical Turk is quite difficult. Experiments mirroring evaluation at the Text Analysis Conference's summarization track show that non-expert judges are not able to recover system rankings derived from experts.

1 Introduction

Automatic summarization is a particularly difficult task to evaluate. What makes a good summary? What information is relevant? Is it possible to separate information content from linguistic quality?

Besides subjectivity issues, evaluation is time-consuming. Ideally, a judge would read the original set of documents before deciding how well the important aspects are conveyed by a summary. A typical 10-document problem could reasonably involve 25 minutes of reading or skimming and 5 more minutes for assessing a 100-word summary. Since summary output can be quite variable, at least 30 topics should be evaluated to get a robust estimate of performance. Assuming a single judge evaluates all summaries for a topic (more redundancy would be better), we get a rough time estimate: 17.5 hours to evaluate two systems.

Thus it is of great interest to find ways of speeding up evaluation while minimizing subjectivity. Amazon's Mechanical Turk (MTurk) system has been used for a variety of labeling and annotation tasks (Snow et al., 2008), but such crowd-sourcing has not been tested for summarization.

We describe an experiment to test whether MTurk is able to reproduce system-level rankings that match expert opinion. Unlike the results of other crowd-sourcing annotations for natural language tasks, we find that non-expert judges are unable to provide expert-like scores and tend to disagree significantly with each other.

This paper is organized as follows: Section 2 introduces the particular summarization task and data we use in our experiments; Section 3 describes the design of our Human Intelligence Task (HIT). Section 4 shows experimental results and gives some analysis. Section 5 reviews our main findings and provides suggestions for researchers wishing to conduct their own crowd-sourcing evaluations.

2 TAC Summarization Task

Topic: Peter Jennings
Description: Describe Peter Jennings' lung cancer and its effects.

Reference: Peter Jennings's announcement April 5, 2005, that he had lung cancer left his colleagues at ABC News saddened and dismayed. He had been ”World News Tonight” anchorman since 1983. By the end of the week, ABC had received 3,400 e-mails offering him prayers and good wishes. A former heavy smoker, Jennings had not been well for some time and was unable to travel abroad to cover foreign events. However, his diagnosis came as a surprise to him. ABC announced that Jennings would continue to anchor the news during chemotherapy treatment, but he was unable to do so.

Table 1: An example topic and reference summary from the TAC 2009 summarization task.

Our data comes from the submissions to the Text Analysis Conference (TAC) summarization track in 2009 (Dang, 2009). The main task involved 44 query-focused topics, each requiring a system to produce a 100-word summary of 10 related news documents. Experts provided four reference summaries for each topic. Table 1 shows an example.

	Score Difference				
	0	1	2	3	mean
OQ	119	92	15	0	0.54
LQ	117	82	20	7	0.63

Table 2: Identical summaries often were given different scores by the same expert human judge at TAC 2009. Counts of absolute score differences are shown for Overall Quality (OQ) and Linguistic Quality (LQ).

2.1 Agreement and consistency

In the official TAC evaluation, each summary was judged by one of eight experts for "Overall Quality" and "Linguistic Quality" on a 1 ("very poor") to 10 ("very good") scale. Unfortunately, the lack of redundant judgments means we cannot estimate inter-annotator agreement. However, we note that out of all 4576 submitted summaries, there are 226 pairs that are identical, which allows us to estimate annotator consistency. Table 2 shows that an expert annotator will give the same summary the same score just over half the time.

2.2 Evaluation without source documents

One way to dramatically speed up evaluation is to use the experts' reference summaries as a gold standard, leaving the source documents out entirely. This is the idea behind automatic evaluation with ROUGE (Lin, 2004), which measures ngram overlap with the references, and assisted evaluation with Pyramid (Nenkova and Passonneau, 2004), which measures overlap of facts or "Semantic Content Units" with the references. The same idea has also been employed in various manual evaluations, for example by Haghighi and Vanderwende (2009), to directly compare the summaries of two different systems. The potential bias introduced by such abbreviated evaluation has not been explored.

3 HIT design

The overall structure of the HIT we designed for summary evaluation is as follows: The worker is asked to read the topic and description, and then two reference summaries (there is no mention of the source documents). The candidate summary appears next, followed by instructions to provide scores between 1 (very poor) and 10 (very good) in each category[1]. Mouse-over on the category names provides extra details, copied with slight modifications from Dang (2007).

Our initial HIT design asked workers to perform a head-to-head comparison of two candidate summaries, but we found this unsatisfactory for a number of reasons. First, many of the resulting scores did not obey the transitive property: given summaries x, y, and z, a single worker showed a preference for $y > x$ and $z > y$, but also $x > z$. Second, while this kind of head-to-head evaluation may be useful for system development, we are specifically interested here in comparing non-expert MTurk evaluation with expert TAC evaluation.

We went through a few rounds of revisions to the language in the HIT after observing worker feedback. Specifically, we found it was important to emphasize that a good summary not only responds to the topic and description, but also conveys the information in the references.

3.1 Quality control

Only workers with at least a 96% HIT approval rating[2] were allowed access to this task. We monitored results manually and blocked workers (rejecting their work) if they completed a HIT in under 25 seconds. Such suspect work typically showed uniform scores (usually all 10s). Nearly 30% of HITs were rejected for this reason.

To encourage careful work, we included this note in our HITs: "High annotator consistency is important. If the scores you provide deviate from the average scores of other annotators on the same HIT, your work will be rejected. We will award bonuses for particularly good work." We gave a few small bonuses ($0.50) to workers who left thoughtful comments.

3.2 Compensation

We experimented with a few different compensation levels and observed a somewhat counter-intuitive result. Higher compensation ($.10 per HIT) yielded lower quality work than lower compensation ($.07 per HIT), judging by the number of HITs we rejected. It seems that lower compensation attracts workers who are less interested in making money, and thus willing to spend more time and effort. There is a trade-off, though, as there are fewer workers willing to do the task for less money.

[1] Besides Overall Quality and Linguistic Quality, we include Information Content, to encourage judges to distinguish between content and readability.

[2] MTurk approval ratings calculated as the fraction of HITs approved by requesters.

Sys	TAC		MTurk		
	OQ	LQ	OQ	LQ	C
A	5.16	5.64	7.03	7.27	7.27
B	4.84	5.27	6.78	6.97	6.78
C	4.50	4.93	6.51	6.85	6.49
D	4.20	4.09	6.15	6.59	6.50
E	3.91	4.70	6.19	6.54	6.58
F	3.64	6.70	7.06	7.78	6.56
G	3.57	3.43	5.82	6.33	6.28
H	3.20	5.23	5.75	6.06	5.62

Table 3: Comparison of Overall Quality (OQ) and Linguistic Quality (LQ) scores between the TAC and MTurk evaluations. Content (C) is evaluated by MTurk workers as well. Note that system F is the lead baseline.

4 Experiments and Analysis

To assess how well MTurk workers are able to emulate the work of expert judges employed by TAC, we chose a subset of systems and analyze the results of the two evaluations. The systems were chosen to represent the entire range of average Overall Quality scores. System F is a simple lead baseline, which generates a summary by selecting the first sentences up to 100 words of the most recent document. The rest of the systems were submitted by various track participants. The MTurk evaluation included two-times redundancy. That is, each summary was evaluated by two different people. The cost for the full evaluation, including 44 topics, 8 systems, and $2x$ redundancy, at $.07 per HIT, plus 10% commission for Amazon, was $55.

Table 3 shows average scores for the two evaluations. The data suggest that the MTurk judges are better at evaluating Linguistic Quality than Content or Overall Quality. In particular, the MTurk judges appear to have difficulty distinguishing Linguistic Quality from Content. We will defend these claims with more analysis, below.

4.1 Worker variability

The first important question to address involves the consistency of the workers. We cannot compare agreement between TAC and MTurk evaluations, but the MTurk agreement statistics suggest considerable variability. In Overall Quality, the mean score difference between two workers for the same HIT is 2.4 (the standard deviation is 2.0). The mean is 2.2 for Linguistic Quality (the standard deviation is 1.5).

In addition, the TAC judges show more similarity with each other—as if they are roughly in agreement about what makes a good summary. We compute each judge's average score and look at the standard deviation of these averages for the two groups. The TAC standard deviation is 1.0 (ranging from 3.0 to 6.1), whereas the MTurk standard deviation is 2.3 (ranging from 1.0 to 9.5). Note that the average number of HITs performed by each MTurk worker was just over 5.

Finally, we can use regression analysis to show what fraction of the total score variance is captured by judges, topics, and systems. We fit linear models in **R** using binary indicators for each judge, topic, and system. Redundant evaluations in the MTurk set are removed for unbiased comparison with the TAC set. Table 4 shows that the differences between the TAC and MTurk evaluations are quite striking: Taking the TAC data alone, the topics are the major source of variance, whereas the judges are the major source of variance in the MTurk data. The systems account for only a small fraction of the variance in the MTurk evaluation, which makes system ranking more difficult.

Eval	Judges	Topics	Systems
TAC	0.28	0.40	0.13
MTurk	0.44	0.13	0.05

Table 4: Linear regression is used to model Overall Quality scores as a function of judges, topics, and systems, respectively, for each data set. The R^2 values, which give the fraction of variance explained by each of the six models, are shown.

4.2 Ranking comparisons

The TAC evaluation, while lacking redundant judgments, was a balanced experiment. That is, each judge scored every system for a single topic. The same is not true for the MTurk evaluation, and as a result, the average per-system scores shown in Table 3 may be biased. As a result, and because we need to test multiple system-level differences simultaneously, a simple t-test is not quite sufficient. We use Tukey's Honestly Significant Differences (HSD), explained in detail by Yandell (1997), to assess statistical significance.

Tukey's HSD test computes significance intervals based on the range of the sample means rather than individual differences, and includes an adjustment to correct for imbalanced experimental designs. The **R** implementation takes as input a linear model, so we

Eval	Ranking							
TAC (OQ)	A	B	C	D^{A}	E^{B}	F^{C}	G^{C}	H^{D}
MTurk (OQ)	F	A	B	C	E^{F}	G^{F}	D^{B}	H^{B}
TAC (LQ)	F	A^{F}	B^{F}	H^{F}	C^{F}	E^{A}	D^{B}	G^{E}
MTurk (LQ)	F	A	B^{F}	C^{F}	D^{F}	E^{F}	H^{C}	G^{C}
MTurk (C)	A	B	E	F	D	C	G^{A}	H^{D}

Table 5: Systems are shown in rank order from highest (left) to lowest (right) for each scoring metric: Overall Quality (OQ), Linguistic Quality (LQ), and Content (C). The superscripts indicate the rightmost system that is significantly different (at 95% confidence) according to Tukey's HSD test.

model scores using binary indicators for (J)udges, (T)opics, and (S)ystems (see equation 1), and measure significance in the differences between system coefficients (δ_k).

$$score = \alpha + \sum_i \beta_i J_i + \sum_j \gamma_j T_j + \sum_k \delta_k S_k \quad (1)$$

Table 5 shows system rankings for the two evaluations. The most obvious discrepancy between the TAC and MTurk rankings is system F, the baseline. Both TAC and MTurk judges gave F the highest scores for Linguistic Quality, a reasonable result given its construction, whereas the other summaries tend to pull sentences out of context. But the MTurk judges also gave F the highest scores in Overall Quality, suggesting that readability is more important to amateur judges than experts, or at least easier to identify. Content appears the most difficult category for the MTurk judges, as few significant score differences emerge. Even with more redundancy, it seems unlikely that MTurk judges could produce a ranking resembling the TAC Overall Quality ranking using this evaluation framework.

5 Discussion

Through parallel evaluations by experts at TAC and non-experts on MTurk, we have shown two main results. First, as expected, MTurk workers produce considerably noisier work than experts. That is, more redundancy is required to achieve statistical significance on par with expert judgments. This finding matches prior work with MTurk. Second, MTurk workers are unlikely to produce a score ranking that matches expert rankings for Overall Quality. This seems to be the result of some confusion in separating content from readability.

What does this mean for future evaluations? If we want to assess overall summary quality—that is, balancing content and linguistic quality like expert judges do—we will need to redesign the task for non-experts. Perhaps MTurk workers will be better able to understand Nenkova's Pyramid evaluation (2004), which is designed to isolate content. Extrinsic evaluation, where judges use the summary to answer questions derived from the source documents or the references, as done by Callison-Burch for evaluation of Machine Translation systems (2009), is another possibility.

Finally, our results suggest that anyone conducting an evaluation of summarization systems using non-experts should calibrate their results by asking their judges to score summaries that have already been evaluated by experts.

Acknowledgments

Thanks to Benoit Favre for discussing the evaluation format and to the anonymous reviewers for helpful, detailed feedback.

References

C. Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazons Mechanical Turk. *Proceedings of EMNLP*.

H.T. Dang. 2007. Overview of DUC 2007. In *Proceedings of the Document Understanding Conference*.

H.T. Dang. 2009. Overview of the TAC 2009 opinion question answering and summarization tasks. In *Proceedings of Text Analysis Conference (TAC 2009)*.

A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of HLT-NAACL*.

C.Y. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop: Text Summarization Branches Out*.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*.

B.S. Yandell. 1997. *Practical data analysis for designed experiments*. Chapman & Hall/CRC.