A Methodology for Extrinsic Evaluation of Text Summarization: Does ROUGE Correlate?

Bonnie J. Dorr and **Christof Monz** and **Stacy President** and **Richard Schwartz**[†] and **David Zajic**

Department of Computer Science and UMIACS
University of Maryland
College Park, MD 20742
{bonnie,christof,stacypre,dmzajic}@umiacs.umd.edu

[†]BBN Technologies
9861 Broken Land Parkway
Columbia, Maryland 21046
schwartz@bbn.com

Abstract

This paper demonstrates the usefulness of summaries in an extrinsic task of relevance judgment based on a new method for measuring agreement, *Relevance-Prediction*, which compares subjects' judgments on summaries with their own judgments on full text documents. We demonstrate that, because this measure is more reliable than previous gold-standard measures, we are able to make stronger statistical statements about the benefits of summarization. We found positive correlations between ROUGE scores and two different summary types, where only weak or negative correlations were found using other agreement measures. However, we show that ROUGE may be sensitive to the choice of summarization style. We discuss the importance of these results and the implications for future summarization evaluations.

1 Introduction

People often prefer to read a summary of a text document, e.g., news headlines, scientific abstracts, movie previews and reviews, and meeting minutes. Correspondingly, the explosion of online textual material has prompted advanced research in document summarization. Although researchers have demonstrated that users can read summaries faster than full text (Mani et al., 2002) with some loss of accuracy, researchers have found it difficult to draw strong conclusions about the usefulness of summarization due to the low level of interannotator agreement in the gold standards that they have used. Definitive conclusions about the usefulness of summaries would provide justification for continued research and development of new summarization methods.

To investigate the question of whether text summarization is useful in an extrinsic task, we examined human performance in a relevance assessment task using a human text *surrogate* (i.e. text intended to stand in the place of a document). We use single-document English summaries as these are sufficient for investigating task-based usefulness, although more elaborate surrogates are possible, e.g., those that span more than one document (Radev and McKeown, 1998; Mani and Bloedorn, 1998).

The next section motivates the need for developing a new framework for measuring task-based usefulness. Section 3 presents a novel extrinsic measure called *Relevance-Prediction*. Section 4 demonstrates that this is a more reliable measure than that of previous gold standard methods, e.g., the *LDC-Agreement* method used for SUMMAC-style evaluations, and that this reliability allows us to make stronger statistical statements about the benefits of summarization. We expect these findings to be important for future summarization evaluations.

Section 5 presents the results of correlation between task usefulness and the Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric (Lin and Hovy, 2003).[1] While we show that ROUGE correlates with task usefulness (using our Relevance-Prediction measure), we detect a slight difference between informative, *extractive* headlines (containing words from the full document) and less informative, *non-extractive* "eye-catchers" (containing words that might not appear in the full document, and intended to entice a reader to read the entire document).

Section 6 further highlights the importance of this point and discusses the implications for automatic evaluation of non-extractive summaries. To evaluate non-extractive summaries reliably, an automatic measure may require knowledge of sophisticated meaning units.[2] It is our hope that the conclusions drawn herein will prompt investigation into more sophisticated automatic metrics as researchers shift their focus to non-extractive summaries.

[1] ROUGE has been previously used as the primary automatic evaluation metric by NIST in the 2003 and 2004 DUC Evaluations.

[2] The *content units* proposed in recent methods (Nenkova and Passonneau, 2004) are a first step in this direction.

Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 1–8, Ann Arbor, June 2005.

2 Background

In the past, assessments of usefulness involved a wide range of both intrinsic and extrinsic (task-based) measures (Sparck-Jones and Gallier, 1996). Intrinsic evaluations focus on coherence and informativeness (Jing et al., 1998) and often involve quality comparisons between automatic summaries and reference summaries that are pre-determined to be of high quality. Human intrinsic measures determine quality by assessing document accuracy, fluency, and clarity. Automatic intrinsic measures such as ROUGE use n-gram scoring to produce rankings of summarization methods.

Extrinsic evaluations concentrate on the use of summaries in a specific task, e.g., executing instructions, information retrieval, question answering, and relevance assessments (Mani, 2001). In relevance assessments, a user reads a topic or event description and judges relevance of a document to the topic/event based solely on its summary.[3] These have been used in many large-scale extrinsic evaluations, e.g., SUMMAC (Mani et al., 2002) and the Document Understanding Conference (DUC) (Harman and Over, 2004). The task chosen for such evaluations must support a very high degree of interannotator agreement, i.e., consistent relevance decisions across subjects with respect to a predefined *gold standard*.

Unfortunately, a consistent gold standard has not yet been reported. For example, in two previous studies (Mani, 2001; Tombros and Sanderson, 1998), users' judgments were compared to "gold standard judgments" produced by members of the University of Pennsylvania's Linguistic Data Consortium. Although these judgments were supposed to represent the *correct* relevance judgments for each of the documents associated with an event, both studies reported that annotators' judgments varied greatly and that this was a significant issue for the evaluations. In the SUMMAC experiments, the Kappa score (Carletta, 1996; Eugenio and Glass, 2004) for interannotator agreement was reported to be 0.38 (Mani et al., 2002). In fact, large variations have been found in the initial summary scoring of an individual participant and a subsequent scoring that occurs a few weeks later (Mani, 2001; van Halteren and Teufel, 2003).

This paper attempts to overcome the problem of interannotator inconsistency by measuring summary effectiveness in an extrinsic task using a much more consistent form of user judgment instead of a gold standard. Using Relevance-Prediction increases the confidence in our results and strengthens the statistical statements we can make about the benefits of summarization.

The next section describes an alternative approach to measuring task-based usefulness, where the usage of external judgments as a gold standard is replaced by the user's own decisions on the full text. Following the lead of earlier evaluations (Oka and Ueda, 2000; Mani et al., 2002; Sakai and Sparck-Jones, 2001), we focus on relevance assessment as our extrinsic task.

[3] A topic is an event or activity, along with all directly related events and activities. An event is something that happens at some specific time and place, and the unavoidable consequences.

3 Evaluation of Usefulness of Summaries

We define a new extrinsic measure of task-based usefulness called *Relevance-Prediction*, where we compare a summary-based decision to the subject's own full-text decision rather than to a different subject's decision. Our findings differ from that of the SUMMAC results (Mani et al., 2002) in that using Relevance-Prediction as an alternative to comparision to a gold standard is a more realistic agreement measure for assessing usefulness in a relevance assessment task. For example, users performing browsing tasks must examine document surrogates, but open the full-text only if they expect the document to be interesting to them. They are not trying to decide if the document will be interesting to someone else.

To determine the usefulness of summarization, we focus on two questions:

- Can users make judgments on summaries that are consistent with their full-text judgments?
- Can users make judgments on summaries more quickly than on full document text?

First we describe the Relevance-Prediction measure for determining whether users can make accurate judgments with a summary. Following this, we describe our experiments and results using this measure, including the timing results of summaries compared to full documents.

3.1 Relevance-Prediction Measure

To answer the first question above, we define a measure called *Relevance-Prediction*, where subjects build their own "gold standard" based on the full-text documents. Agreement is measured by comparing subjects' surrogate-based judgments against their own judgments on the corresponding texts. The subject's judgment is assigned a value of *1* if his/her surrogate judgment is the same as the corresponding full-text judgment, and *0* otherwise. These values were summed over all judgments for a surrogate type and were divided by the total number of judgments for that surrogate type to determine the effectiveness of the associated summary method.

Formally, given a summary/document pair (s, d), if subjects make the same judgment on s that they did on d, we say $j(s, d) = 1$. If subjects change their judgment between s and d, we say $j(s, d) = 0$. Given a set of summary/document pairs DS_i associated with event i, the Relevance-Prediction score is computed as follows:

$$Relevance\text{-}Prediction(i) = \frac{\sum_{s,d \in DS_i} j(s, d)}{|DS_i|}$$

This approach provides a more reliable comparison mechanism than gold standard judgments provided by

other individuals. Specifically, Relevance-Prediction is more helpful in illuminating the usefulness of summaries for a real-world scenario, e.g., a browsing environment, where credit is given when an individual subject would choose (or reject) a document under both conditions. To our knowledge, this subject-driven approach to testing usefulness has never before been used.

3.2 Experiment Design

Ten human subjects were recruited to evaluate full-text documents and two summary types.[4] The original text documents were taken from the Topic Detection and Tracking 3 (TDT-3) corpus (Allan et al., 1999) which contains news stories and headlines, topic and event descriptions, and a mapping between news stories and their related topic and/or events. Although the TDT-3 collection contains transcribed speech documents, our investigation was restricted to documents that were originally text, i.e., newspaper or newswire, not broadcast news.

For our experiment we selected three distinct events and related document sets[5] from TDT-3. For each event, the subjects were given a description of the event (written by LDC) and then asked to judge relevance of a set of 20 documents associated with that event (using three different presentation types to be discussed below).

The events used from the TDT data set were events from world news occurring in 1998. It is possible that the subjects had some prior knowledge about the events, yet we believe that this would not affect their ability to complete the task. Subjects' background knowledge of an event can also make this task more similar to real-world browsing tasks, in which subjects are often familiar with the event or topic they are searching for.

The 20 documents were retrieved by a search engine. We used a constrained subset where exactly half (10) were judged relevant by the LDC annotators. Because all 20 documents were somewhat similar to the event, this approach ensured that our task would be more difficult than it would be if we had chosen documents from completely unrelated events (where the choice of relevance would be obvious even from a poorly written summary).

Each document was pre-annotated with the headline associated with the original newswire source. These headlines were used as the first summary type. We refer to them as HEAD (*Headline Surrogate*). The average length of the HEAD surrogates was 53 characters. In addition, we commissioned human-generated summaries[6] of each document as the second summary type; we refer to this as HUM (*Human Surrogate*). The average length of the HUM surrogates was 72 characters. Although neither of these summaries was produced automatically, our experiment allowed us to focus on the question of summary usefulness and to learn about the differences in presentation style as a first step toward experimentation with the output of automatic summarization systems.

Two main factors were measured: (1) differences in judgments for the three presentation types (HEAD, HUM, and the full-text document) and (2) judgment time. Each subject made a total of 60 judgments for each presentation type since there were 3 distinct events and 20 documents per event. To facilitate the analysis of the data, the subjects' judgments were constrained to two possibilities, *relevant* or *not relevant*.[7]

Although the HEAD and HUM surrogates were both produced by humans, they differed in style. The HEAD surrogates were shorter than the HUM surrogates by 26%. Many of these were "eye-catchers" designed to entice the reader to examine the entire document (i.e., purchase the newspaper); that is, the HEAD surrogates were not intended to stand in the place of the full document. By contrast, the writers of the HUM surrogates were instructed to write text that conveyed what happened in the full document. We observed that the HUM surrogates used more words and phrases extracted from the full documents than the HEAD surrogates.

Experiments were conducted using a web browser (Internet Explorer) on a PC in the presence of the experimenter. Subjects were given written and verbal instructions for completing their task and were asked to make relevance judgments on a practice event set. The judgments from the practice event set were not included in our experimental results or used in our analyses. The written instructions were given to aid subjects in determining requirements for relevance. For example, in an Election event documents describing new people in office, new public officials, change in governments or parliaments were suggested as evidence for relevance.

Each of the ten subjects made judgments on 20 documents for each of three different events. After reading each document or summary, the subjects clicked on a radio button corresponding to their judgment and clicked a *submit* button to move to the next document description. Subjects were not allowed to move to the next summary/document until a valid selection was made. No backing up was allowed. Judgment time was computed as the number of seconds it took the subject to read the full text document or surrogate, comprehend it, compare it to the event description, and make a judgment (timed up until the subject clicked the *submit* button).

[4] We required all human subjects to be native-English speakers to ensure that the accuracy of judgments was not degraded by language barriers.

[5] The three event and related document sets contained enough data points to achieve statistically significant results.

[6] The human summarizers were instructed to create a summary no greater than 75 characters for each specified full text document. The summaries were not compared for writing style or quality.

[7] If we allowed subjects to make additional judgments such as *somewhat relevant*, this could possibly encourage subjects to always choose this when they were the least bit unsure. Previous experiments indicate that this additional selection method may increase the level of variability in judgments (Zajic et al., 2004).

3.3 Order of Document/Surrogate Presentation

One concern with our evaluation methodology was the issue of possible memory effects or priming: if the same subjects saw a summary and a full document about the same event, their answers might be tainted. Thus, prior to the full experiment, we conducted pre-experiments (using 4 participants) with an extreme form of influence: we presented the summary and full text in immediate succession. In these experiments, we compared two document presentation approaches, termed "Drill Down" and "Complete Set." In the "Drill Down" document presentation approach all three presentation types were shown for each document, in sequence: first a single HEAD surrogate, followed by the corresponding HUM surrogate, followed by the full text document. This process was repeated 10 times.

In the "Complete Set" document-presentation approach we presented the complete set of documents using one surrogate type, followed by the complete set using another surrogate type, and so on. That is, the 10 HEAD surrogates were displayed all at once, followed by the corresponding 10 HUM surrogates, followed by the corresponding 10 full-text documents.

The results indicated that there was almost no effect between the two document-presentation approaches. The performance varied only slightly and neither approach consistently allowed subjects to perform better than the other. Therefore, we determined that the subjects were not associating a given summary with its corresponding full-text documents. This may be due, in part, to the fact that all 20 documents were related to the event—and according to the LDC relevance judgments half of these were actually about the same event.

Given that the variations were insignificant in these pre-experiments, we selected only the Complete-Set approach (no Drill-Down) for the full experiment. However, we still needed to vary the ordering for the two surrogate presentation types associated with each full-text document. Thus, each 20-document set was divided in half for each subject. In the first half, the subject saw the first 10 documents as: (1) HEAD surrogates, then HUM surrogates and then the full-text document; or (2) HUM surrogates, then HEAD surrogates, and then the full-text document. In the second half, the subject saw the alternative ordering, e.g., if a subject saw HEAD surrogates before HUM surrogates in the first half, he/she saw the HUM surrogates before HEAD surrogates for the second half. Either way, the full-text document was always shown last so as not to introduce judgment effects associated with reading the entire document before either surrogate type.

In addition to varying the ordering for the surrogate type, the ordering of the surrogates and full documents within the events were also varied. The subjects were grouped in pairs, and each pair viewed the surrogates and documents in a different order than the other pairs.

3.4 Experimental Hypotheses

We hypothesized that the summaries would allow subjects to achieve a Relevance-Prediction rate of 70–90%. Since these summaries were significantly shorter than the original document text, we expected that the rate would not be 100% compared to the judgments made on the full document text. However, we expected higher than a 50% ratio, i.e., higher than that of random judgments on all of the surrogates. We also expected high performance because the meaning of the original document text is best preserved when written by a human (Mani, 2001).

A second hypothesis is that the HEAD surrogates would yield a significantly lower agreement rate than that of the HUM surrogates. Our commissioned HUM surrogates were written to stand in place of the full document, whereas the HEAD surrogates were written to catch a reader's interest. This suggests that the HEAD surrogates might not provide as informative a description of the original documents as the HUM surrogates.

We also tested a third hypothesis: that our Relevance-Prediction measure would be more reliable than that of the *LDC-Agreement* method used for SUMMAC-style evaluations (thus providing a more stable framework for evaluating summarization techniques). LDC-Agreement compares a subject's judgment on a surrogate or full text against the "correct" judgments as assigned by the TDT corpus annotators (Linguistic Data Consortium 2001).

Finally, we tested the hypothesis that using a text summary for judging relevance would take considerably less time than using the corresponding full-text document.

4 Experimental Results

Table 1 shows the subjects' judgments using both Relevance-Prediction and LDC-Agreement for each of three events. Using our Relevance-Prediction measure, the HUM surrogates yield averages between 79% and 86%, with an overall average of 81%, thus confirming our first hypothesis.

However, we failed to confirm our second hypothesis. The HEAD Relevance-Prediction rates were between 71% and 82%, with an overall average of 76%, which was lower than the rates for HUM, but the difference was not statistically significant. It appeared that subjects were able to make consistent relevance decisions from the non-extractive HEAD surrogates, even though these were shorter and less informative than the HUM surrogates.

A closer look reveals that the HEAD summaries sometimes contained enough information to judge relevance, yielding almost the same number of true positives (and true negatives) as the HUM summaries. For example, a document about the formation of a coalition government to avoid violence in Cambodia has the HEAD surrogate *Cambodians hope new government can avoid past mistakes*. By contrast, the HUM surrogate for this same event was *Rival parties to form a coalition government to avoid violence in Cambodia*. Although the HEAD surrogate

Surrogate	EVENT 1		EVENT 2		EVENT 3		Overall Avg		Avg Time
	LDC	RP	LDC	RP	LDC	RP	LDC	RP	(seconds)
HEAD	67%	76%	66%	71%	70%	82%	67%	76%	4.60
HUM	69%	80%	73%	86%	62%	79%	68%	81%	4.57
DOC	—	—	—	—	—	—	—	—	13.38

Table 1: Relevance-Prediction (RP) and LDC-Agreement (LDC) Rates for HEAD and HUM Surrogates for each Event

uses words that do not appear in the original document (*hope* and *mistakes*), the subject may infer the relevance of this surrogate by relating *hope* to the notion of forming a coalition government and *mistakes* to violence.

On the other hand, we found that the lower degree of informativeness of HEAD surrogates gave rise to over 50% more false negatives than the HUM summaries. This statistically significant difference will be discussed further in Section 6.

As for our third hypothesis, Table 1 illustrates a substantial difference between the two agreement measures. For each of the three events, the Relevance-Prediction rate is at least five percent higher than that of the LDC-Agreement approach, with an average of 8.8% increase for the HEAD summary and a 13.3% average increase for the HUM summary. The average rates across events show a statistically significant difference between LDC-Agreement and Relevance-Prediction for both HUM summaries with $p<0.01$ and HEAD summaries with $p<0.05$. This significance was determined through use of a single factor ANOVA statistical analysis. The higher Relevance-Prediction rate supports our statement that this approach provides a more stable framework for evaluating different summarization techniques.

Finally, the average timing results shown in Table 1 confirm our fourth hypothesis. The subjects took 4-5 seconds (on average) to make judgments on both the HEAD and HUM summaries, as compared to about 13.4 seconds to make judgments on full text documents. This shows that it takes subjects almost 3 times longer to make judgments on full text documents as it took to make judgments on the summaries (HEAD and HUM). This finding is not surprising since text summaries are an order of magnitude shorter than full-text documents.

5 Correlation with Intrinsic Evaluation Metric: ROUGE

We now turn to the task of correlating our extrinsic task performance with scores produced by an intrinsic evaluation measure. We used the Recall Oriented Understudy for Gisting Evaluation (ROUGE) metric version 1.2.1. In previous studies (Dorr et al., 2004) ROUGE was shown to have a very low correlation with the LDC-Agreement measurement results of the extrinsic task. This was attributed to low interannotator agreement in the gold standard. Our goal was to test whether our new Relevance-Prediction technique would allow us to induce higher correlations with ROUGE.

5.1 Extrinsic Agreement Data

To reduce the effect of outliers on the correlation between ROUGE and the human judgments, we averaged over all judgments for each subject (20 judgments $\times$ 3 events) to produce 60 data points. These data points were then partitioned into either 1, 2, or 4 partitions of equal size. (Partitions of size four have 15 data points, partitions of size two have 30 data points, and partitions of size one have 60 data points per subject—or a total of 600 datapoints across all 10 subjects). To ensure that the correlation did not depend on a specific partition, we repeated this same process using 10,000 different (randomly generated) partitions for each of the three partition sizes.

Partitioned data points of size four provided a high degree of noise reduction without compromising the size of the data set (15 points). Larger partition sizes would result in too few data points and compromise the statistical significance of our correlation results. In order to show the variation within a single partition, we used the partitioning of size 4 with the smallest mean square error on the human headline compared to the other partitionings as a representative partition. For this representative partitioning, the individual data points P1–P15 of that partition are shown for each of the two agreement measures in Tables 2 and 3. This shows that, across partitions, the maximum and minimum Relevance-Prediction rates for HEAD (93% and 60%) are higher than the corresponding LDC-Agreement rates (85% and 50%). The same trend is seen with the HUM surrogates: Relevance-Prediction maximum of 98%, minimum of 68%; and LDC-Agreement maximum 88%, minimum of 55%.

5.2 Intrinsic ROUGE Score

To correlate the partitioned agreement scores above with our intrinsic measure, we first ran ROUGE on all 120 surrogates in our experiment (i.e., the HUM and HEAD surrogates for each of the 60 event/document pairs) and then averaged the ROUGE scores for all surrogates belonging to the same partitions (for each of the three partition sizes). These partitioned ROUGE values were then used for detecting correlations with the corresponding partitioned agreement scores described above.

Table 4 shows the ROUGE scores, based on 3 reference summaries per document, for partitions P1–P15 used in the previous tables.[8] For brevity, we include

[8]We commissioned a total of 180 human-generated reference summaries (3 for each of 60 documents) (in addition to the human generated summaries used in the experiment).

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
HEAD	80%	80%	85%	70%	73%	60%	80%	75%	60%	75%	88%	68%	80%	93%	83%
HUM	83%	88%	85%	68%	75%	75%	93%	75%	98%	90%	75%	70%	80%	90%	78%

Table 2: Relevance-Prediction Rates for HEAD and HUM Surrogates (Representative Partition of Size 4)

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
HEAD	70%	73%	85%	70%	63%	60%	60%	85%	50%	73%	70%	78%	65%	63%	73%
HUM	68%	75%	58%	68%	75%	70%	68%	80%	88%	58%	63%	55%	55%	60%	78%

Table 3: LDC-Agreement Rates for HEAD and HUM Surrogates (Representative Partition of Size 4)

Surrogate	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Avg
HEAD	.10	.23	.13	.27	.20	.24	.26	.22	.13	.08	.30	.16	.26	.27	.30	.211
HUM	.16	.22	.17	.23	.19	.36	.39	.29	.28	.25	.37	.22	.22	.39	.27	.269

Table 4: Average Rouge-1 Scores for HEAD and HUM Surrogates (Representative Partition of Size 4)

only ROUGE 1-gram measurement (R1).[9] The ROUGE scores for HEAD surrogates were slightly lower than those for HUM surrogates. This is consistent with our statements earlier about the difference between non-extractive "eye-catchers" and informative headlines. Because ROUGE measures whether a particular summary has the same words (or n-grams) as a reference summary, a more constrained choice of words (as found in the extractive HUM surrogates) makes it more likely that the summary would match the reference.

A summary in which the word choice is less constrained—as in the non-extractive HEAD surrogates—is less likely to share n-grams with the reference. Thus, we may see non-extractive summaries that have almost identical meanings, but very different words. This raises the concern that ROUGE may be sensitive to the style of summarization that is used. Section 6 discusses this point further.

5.3 Intrinsic and Extrinsic Correlation

To test whether ROUGE correlates more highly with Relevance-Prediction than with LDC-Agreement, we calculated the correlation for the results of both techniques using Pearson's *r* (Siegel and Castellan, 1988):

$$\frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}}$$

where r_i is the ROUGE score of surrogate i, $\bar{r}$ is the average ROUGE score of all data points, s_i is the agreement score of summary i (using Relevance-Prediction or LDC-Agreement), and $\bar{s}$ is the average agreement score. Pearson's statistics is commonly used in summarization and machine translation evaluation, see e.g. (Lin, 2004; Lin and Och, 2004).

As one might expect, there is some variability in the correlation between ROUGE and human judgments for the different partitions. However, the boxplots for both HEAD and HUM indicate that the first and third quartile were relatively close to the median (see Figure 1).

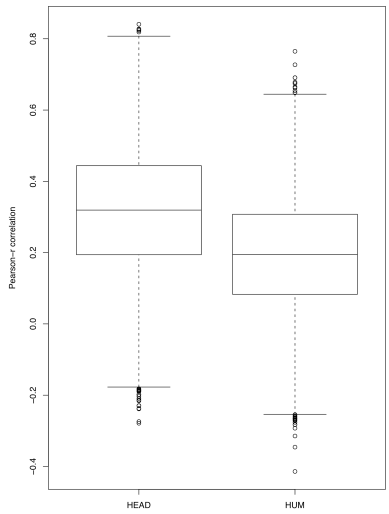

Figure 1: Distribution of the Correlation Variation for Relevance-Prediction on HEAD and HUM

Table 5 shows the Pearson Correlations with ROUGE-1 using Relevance-Prediction and LDC-Agreement. For Relevance-Prediction, we observed a positive correlation for both surrogate types, with a slightly higher correlation for HEAD than HUM. For LDC-Agreement, we observed no correlation (or a minimally negative one) with ROUGE-1 scores, for both the HEAD and HUM surrogates. The highest correlation was observed for Relevance-Prediction on HEAD.

We conclude that ROUGE correlates more highly with the Relevance-Prediction measurement than the LDC-Agreement measurement, although we should add that none of the correlations in Table 5 were statistically significant at $p < 0.05$. The low LDC-Agreement scores are consistent with previous studies where poor correlations

[9] We also computed ROUGE 2-gram, ROUGE L and ROUGE W, but the trend for these did not differ from ROUGE-1.

Surrogate	P = 1	P = 2	P = 4
HEAD (RP)	0.1270	0.1943	0.3140
HUM (RP)	0.0632	0.1096	0.1391
HEAD (LDC)	-0.0968	-0.0660	-0.0099
HUM (LDC)	-0.0395	-0.0236	-0.0187

Table 5: Pearson Correlations with ROUGE-1 for Relevance-Prediction (RP) and LDC-Agreement (LDC), where Partition size (P) = 1, 2, and 4

were attributed to low interannotator agreement rates.

6 Discussion

Our results suggest that ROUGE may be sensitive to the style of summarization that is used. As we observed above, many of the HEAD surrogates were not actually summaries of the full text, but were eye-catchers. Often, these surrogates did not allow the subject to judge relevance correctly, resulting in lower agreement. In addition, these same surrogates often did not use a high percentage of words that were actually from the story, resulting in low ROUGE scores. (We noticed that most words in the HUM surrogates appeared in the corresponding stories.) There were three consequences of this difference between HEAD and HUM: (1) The rate of agreement was lower for HEAD than for HUM; (2) The average ROUGE score was lower for HEAD than for HUM; and (3) The correlation of ROUGE scores with agreement was higher for HEAD than for HUM.

A further analysis supports the (somewhat counterintuitive) third point above. Although the ROUGE scores of true positives (and true negatives) were significantly lower for HEAD surrogates (0.2127 and 0.2162) than for HUM surrogates (0.2696 and 0.2715), the number of false negatives was substantially higher for HEAD surrogates than for HUM surrogates. These cases corresponded to much lower ROUGE scores for HEAD surrogates (0.1996) than for HUM (0.2586) surrogates.

A summary of this analysis is given in Table 6, where true positives and negatives are indicated by Rel/Rel and NonRel/NonRel, respectively, and false positives and negatives are indicated by Rel/NonRel and NonRel/Rel, respectively.[10] The numbers in parentheses after each ROUGE score refer to the standard deviation for that score. This was computed as follows:

$$Std.\text{-}Dev. = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

where N is the number of surrogates in a particular judgment category (e.g., $N = 245$ for the HEAD-based NonRel/Rel judgments), x_i is the ROUGE score for the i^{th} surrogate, and $\bar{r}$ is the average of all ROUGE scores in that category.

Although there were very few false positives (less than 6% for both HEAD and HUM), the number of false negatives (NonRel/Rel) was particularly high for HEAD (50% higher than for HUM). This difference was statistically significant at $p<0.01$ using the t-test. The large number of false negatives with HEAD may be attributed to the eye-catching nature of these surrogates. A subject may be misled into thinking that this surrogate is not related to an event because the surrogate does not contain words from the event description and is too broad for the subject to extract definitive information (e.g., the surrogate *There he goes again!*). Because the false negatives were associated with the lowest average ROUGE score (0.1996), we speculate that, if a correlation exists between Relevance-Prediction and ROUGE, the false negatives may be a major contributing factor.

Based on this experiment, we conjecture that ROUGE may not be a good method for measuring the usefulness of summaries when the summaries are not extractive. That is, if someone intentionally writes summaries that contain different words than the story, the summaries will also likely contain different words than a reference summary, resulting in low ROUGE scores. However, the summaries, if well-written, could still result in high agreement with the judgments made on the full text.

7 Conclusion

We have shown that two types of human summaries, HEAD and HUM, can be useful for relevance assessment in that they help a user achieve 70-85% agreement in relevance judgments. We observed a 65% reduction in judgment time between full texts and summaries. These findings are important in that they establish the usefulness of summarization and they support research and development of additional summarization methods, including automatic methods.

We introduced a new method for measuring agreement, *Relevance-Prediction*, which takes a subject's full-text judgment as the standard against which the same subject's summary judgment is measured. Because Relevance-Prediction was more reliable than LDC-Agreement judgments, we encourage others to use this measure in future summarization evaluations.

Using this new method, we were able to find positive correlations between relevance assessments and ROUGE scores for HUM and HEAD surrogates, where only

[10] We also included (average) elapsed times for summary judgments in each of the four categories. One might expect a "relevant" judgment to be much quicker than a "non-relevant" judgment (since the latter might require reading the full summary). However, it turned out non-relevant judgments did not always take longer. In fact, the NonRel/NonRel cases took considerably less time than the Rel/Rel and Rel/NonRel cases. On the other hand, the NonRel/Rel cases took considerably more time—almost as much time as reading the full text documents—an indication that the subjects may have re-read the summary a number of times, perhaps vacillating back and forth. Still, the overall time savings was significant, given that the vast majority of the non-relevant judgments were in the NonRel/NonRel category.

Judgment (Surr/Doc)	HEAD			HUM		
	Raw	R1-Avg	Avg Time	Raw	R1-Avg	Avg Time
Rel/Rel	211 (35%)	0.2127 (±0.120)	4.6	251 (42%)	0.2696 (±0.130)	4.2
Rel/NonRel	27 (5%)	0.2115 (±0.110)	7.1	35 (6%)	0.2725 (±0.131)	4.6
NonRel/Rel	117 (19%)	0.1996 (±0.127)	8.5	77 (13%)	0.2586 (±0.120)	13.8
NonRel/NonRel	245 (41%)	0.2162 (±0.126)	2.5	237 (39%)	0.2715 (±0.131)	1.9
TOTAL	600 (100%)	0.2115 (±0.124)	4.6	600 (100%)	0.2691 (±0.129)	4.6

Table 6: Subjects' Judgments and Corresponding Average ROUGE 1 Scores

negative correlations were found using LDC-Agreement scores. We found that both the Relevance-Prediction and the ROUGE-1 scores were higher for human-generated summaries than for the original headlines. It appears that most of the difference is induced by surrogates that are eye-catchers (rather than true summaries), where both agreement and ROUGE scores are low.

Our future work will include further experimentation with automatic summarization methods to determine the level of Relevance-Prediction. We aim to determine how well automatic summarizers help users complete tasks, and to investigate which automatic summarizers perform better than others. We also plan to test for correlations between ROUGE and human task performance with automatic summaries, to further investigate whether ROUGE is a good predictor of human task performance.

Acknowledgements

This work was supported in part by DARPA TIDES Cooperative Agreement N66001-00-2-8910.

References

James Allan, Hubert Jin, Martin Rajman, Charles Wayne, Daniel Gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. 1999. Topic-based Novelty Detection. Technical Report 1999 Summer Workshop at CLSP Final Report, Johns Hopkins, Maryland.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Lingusitics*, 22(2):249–254, June.

Bonnie J. Dorr, Christof Monz, Douglas Oard, Stacy President, and David Zajic. 2004. Extrinsic Evaluation of Automatic Metrics for Summarization. Technical report, University of Maryland, College Park, MD. LAMP-TR-115, CAR-TR-999, CS-TR-4610, UMIACS-TR-2004-48.

Barbara Di Eugenio and Michael Glass. 2004. Squibs and Discussions - The Kappa Statistic: A Second Look. *Computational Linguistics*, pages 95–101.

Donna Harman and Paul Over. 2004. *Proceedings of the DUC 2004*. Boston, MA.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI Symposium on Intelligent Summarization*, Stanford University, CA, March 23-25.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of HLT-NAACL 2003 Workshop*, pages 71–78, Edmonton Canada, May-June.

Chin-Yew Lin and Franz Joseph Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, August 23–27.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25–26.

I. Mani and E. Bloedorn. 1998. Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1(1):35–67.

Inderjeet Mani, Gary Klein, David House, and Lynette Hirschman. 2002. SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Inderjeet Mani. 2001. Summarization Evaluation: An Overview. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the NAACL 2004*, Boston, MA.

Mamiko Oka and Yoshihiro Ueda. 2000. Evaluation of Phrase-Representation Summarization Based on an Information Retrieval Task. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*, pages 59–68, New Brunswick, NJ.

Dragomir Radev and Kathleen McKeown. 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, pages 469–500.

Tetsuya Sakai and Karen Sparck-Jones. 2001. Generic Summaries for Indexing in Information Retrieval - Detailed Test Results. Technical Report TR513, Computer Laboratory, University of Cambridge.

Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition.

Karen Sparck-Jones and J.R. Gallier. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, Berlin.

Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–10.

Hans van Halteren and Simone Teufel. 2003. Examining the Consensus Between Human Summaries: Initial Experiments with Factoid Analysis. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*.

David Zajic, Bonnie J. Dorr, Richard Schwartz, and Stacy President. 2004. Headline Evaluation Experiment Results. Technical report, University of Maryland, College Park, MD. UMIACS-TR-2004-18.