Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management xxx (2007) xxx–xxx
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Task-based evaluation of text summarization using Relevance Prediction

Stacy President Hobson [a,*], Bonnie J. Dorr [a], Christof Monz [b], Richard Schwartz [c]

[a] *Department of Computer Science and UMIACS, University of Maryland, College Park, MD 20742, United States*
[b] *Department of Computer Science, Queen Mary, University of London, London E1 4NS, UK*
[c] *BBN Technologies, Columbia, MD 21046, United States*

Received 14 July 2006; received in revised form 3 January 2007; accepted 8 January 2007

Abstract

This article introduces a new task-based evaluation measure called *Relevance Prediction* that is a more intuitive measure of an individual's performance on a real-world task than interannotator agreement. Relevance Prediction parallels what a user does in the real world task of browsing a set of documents using standard search tools, i.e., the user judges relevance based on a short summary and then that *same* user—not an independent user—decides whether to open (and judge) the corresponding document. This measure is shown to be a more reliable measure of task performance than *LDC Agreement*, a current gold-standard based measure used in the summarization evaluation community. Our goal is to provide a stable framework within which developers of new automatic measures may make stronger statistical statements about the effectiveness of their measures in predicting summary usefulness. We demonstrate—as a proof-of-concept methodology for automatic metric developers—that a current automatic evaluation measure has a better correlation with Relevance Prediction than with LDC Agreement and that the significance level for detected differences is higher for the former than for the latter.

PACS: 07.05.Mh; 89.20.Ff; 43.71.Sy

Keywords: Summarization evaluation; Summary usefulness; Relevance prediction

1. Introduction

With the increased usage of the internet, tasks such as browsing and retrieval of information have become commonplace. Users often skim the first few lines of a document or prefer to have information presented in

* Corresponding author.
E-mail addresses: stacypre@cs.umd.edu (S.P. Hobson), bonnie@umiacs.umd.edu (B.J. Dorr), christof@dcs.qmul.ac.uk (C. Monz), schwartz@bbn.com (R. Schwartz).

0306-4573/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2007.01.002

a reduced or summarized form. Examples of this include document abstracts, news headlines, movie previews and document summaries. Human generated summaries are often costly and time consuming to produce. Therefore, many automatic summarization algorithms/techniques have been proposed to solve the task of text summarization.

To measure the impact of summarization techniques, it is important to have a consistent and easy-to-use method for determining the quality of a given summary (how reflective the summary is of the original document's meaning) and for comparing a summary against other automatic and human summaries. Currently, numerous automatic and semi-automatic evaluation metrics have been developed and are becoming more widely used in the text summarization evaluation community. Many of these methods claim to correlate *highly* (Papineni, Roukos, Ward, & Zhu, 2002) or *surprisingly well* (Lin & Hovy, 2003) with human measures of task performance, and a goal of this work is to investigate these claims. Therefore, we have conducted several relevance-assessment experiments where automatic evaluation metrics are compared to judgments of human performance.

In a study pre-dating this work, users were asked to determine the relevance of a particular document to a specified topic or event, based on the presented document summary or entire document text (Zajic, Dorr, Schwartz, & President, 2004). Judgments made by individual users were compared to "gold standard" judgments as provided by the University of Pennsylvania's Linguistic Data Consortium (LDC, 2006); we refer to this evaluation approach as *LDC Agreement*. These gold standards were considered to be the "correct" judgments, yet we will show that they yield very low interannotator agreement rates and inconsistencies in the user's judgments. Thus, it was difficult to make strong statistical statements using the results of these earlier experiments.

This paper introduces a new measure of summary usefulness, called *Relevance Prediction*, that yields better agreement levels than *LDC Agreement*. Our goal is to provide a stable framework within which developers of new automatic measures may verify more reliably—through correlation studies against our new measure—the effectiveness of their measures in predicting summary usefulness. We demonstrate—as a proof-of-concept methodology for automatic metric developers—that a current automatic evaluation measure has a better correlation with Relevance Prediction than with LDC Agreement and that the significance level for detected differences is higher for the former than for the latter. As such, automatic metric developers may use Relevance Prediction to make stronger statistical statements about the effectiveness of their measures in predicting summary usefulness.

Relevance Prediction is a more intuitive measure of an individual's performance on a real-world task than interannotator agreement. Specifically, Relevance Prediction parallels what a user does in the real world task of browsing a set of documents using standard search tools, i.e., the user judges relevance based on a short summary and then that *same* user—not an independent user—decides whether to open (and judge) the corresponding document. This method eliminates the need for an externally induced "gold standard" by making use of the same user's relevance judgment on both the summary and the corresponding full text.

The next section provides the background and motivation for our work on task-based evaluation of summarization techniques. Following this, Section 3 describes the *LDC Agreement* evaluation approach and introduces the new *Relevance Prediction* measure. Sections 4 and 5 describe experiments that use these measures to verify that it is possible to save time using summaries for relevance assessments without greatly impacting the degree of accuracy that is achieved with full documents. Our results and analyses indicate that Relevance Prediction more reliably predicts task performance than LDC Agreement. Section 6 describes a study that examined various document presentation orderings, to confirm that the order in which documents and summaries were presented in the preceding sections did not affect user's judgments. Finally, we present our conclusions and future work. 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.

2. Motivation

Text summarization evaluation is an area wrought with many challenges. Human evaluations of summary quality are very expensive, labor intensive and time consuming. Participants are usually compensated financially or assigned assessment tasks as part of their normal daily job requirements. Tasks can last from one to a few hours per participant depending upon the number of documents and summaries to be judged.

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Participants' judgments vary greatly and generally do not match gold standard judgments. Very low agreement rates have been reported by Mani (2001), Tombros and Sanderson (1998) in studies that use such standards. At least four total participants are usually needed to produce representative results, although more participants are needed for the most reliable results.

These and other challenges have led researchers to investigate the use of automatic summarization evaluation methods. Such methods are fast, inexpensive, easy to use, and reusable; moreover, they allow developers to continuously check for improvements based on small changes to their summarization system. An example of an automatic intrinsic measure is ROUGE (Lin, 2004; Lin & Hovy, 2003), a modified *n*-gram recall-based metric.[1] However, a previous study has shown only minimal (if any) correlations between automatic summarization measures of human task performance (Zajic et al., 2004).

One issue with these prior studies is that they adopted evaluation designs that were *intrinsic* in nature, i.e., assessments of summary quality are made without reference to a particular task. Of these, *human* intrinsic evaluations have been used to assess the summarization system itself, based on factors such as clarity, coherence, fluency and informativeness (Jing, Barzilay, McKeown, & Elhadad, 1998). Alternatively, *automatic* intrinsic evaluation measures have been used to compare a candidate summary (output of a summarizer) against an 'ideal' or model human summary (Mani, Klein, House, & Hirschman, 2002).

While important, intrinsic measures do not address an *extrinsic* question that is central to the work reported in this paper: *how is text summarization useful*? Summarization has previously been shown to reduce cognitive load (Tombros & Sanderson, 1998). Our focus, however, is on two other possible benefits of using a summary over the full text: (1) Summaries should reduce the reading and judgment time for relevance assessments or other tasks; and (2) Summaries should provide enough information for a reader to get the general meaning of a document so that he/she can make judgments that are as accurate as the judgments on full texts in a relevance assessment task.

Previous work—in the Tipster SUMMAC studies (Mani et al., 2002)—de-mon-stra-ted that users can read summaries faster than the full text, with some loss of accuracy; however, researchers have found it difficult to draw strong conclusions about the usefulness of summarization due to the low level of interannotator consistency in the gold standards that they have used. Moreover, these studies focused on extrinsic task-based evaluations rather than on correlations between intrinsic measures and extrinsic measures of human task performance. As we will see in the next section, our new extrinsic measure—*Relevance Prediction*—is demonstrated to predict task performance more reliably than gold-standard approaches and, as such, allows developers of automatic intrinsic measures to make stronger statistical statements about the effectiveness of their measures in predicting summary usefulness (through correlation studies against this measure).

In this work, we concentrate on short, 75 character single document summaries.[2] Our future work will investigate other areas of summarization, including longer non-headline like summaries, and multi-document topic-focused summaries, as discussed in Section 7. This work yields a usable framework for drawing definitive conclusions about summary usefulness and for justifying continued research and development of new summarization methods.

3. Toward a new extrinsic measure: Relevance Prediction

To investigate the question of whether summaries are useful for a particular extrinsic task, we must first choose a task that is appropriate—one where summaries may serve as a *surrogate*, i.e., a brief snippet that represents the content of one or more full-text documents. We must then determine how to measure summary usefulness with respect to that task.

[1] Although ROUGE is the intrinsic measure used in our own studies, several other metrics that have been proposed will be examined in our future studies, e.g., basic elements (BE) (Hovy, Lin, & Zhou, 2005), the Pyramid Method (Nenkova & Passonneau, 2004), and the Pourpre method (Lin & Demner-Fushman, 2005).

[2] Many search engines use longer summaries but other resources—including news headlines, Google News, and really simple syndication (RSS) feeds—use summaries that are approximately this length. The 75 character length is also consistent with the summary limit used in the document understanding conference (DUC) evaluations of single document summarization.

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

It is important that the extrinsic task be unambiguous enough that it can be performed with a high level of agreement among humans. If the task is so difficult that humans cannot perform it with a high level of agreement—even when they are shown the entire document—it will not be possible to detect significant differences among summarization methods because the amount of variation due to noise will overshadow the variation due to the summarization method.

Common human extrinsic tasks are question-answering, instruction execution, information retrieval, and relevance assessments. For the purpose of the experiments described below, we have selected relevance assessment because of its closeness to a real-world task performed daily by many people, i.e., the task of web searching and information retrieval. Relevance assessment tasks measure the impact of summarization on determining the relevance of a document to a topic (Brandow, Mitze, & Rau, 1995; Jing et al., 1998; Tombros & Sanderson, 1998); these have been used in many large-scale extrinsic evaluations, e.g., the Tipster SUMMAC evaluation (Mani et al., 2002) and the document understanding conference (DUC) (Harman & Over, 2004).

As for the measure of "summary usefulness" we first examine a *gold standard* approach that has been used in past studies. Because relevance assessment is our selected task, the gold standard consists of human relevance judgments—*relevant* or *not relevant*—that are thought to reflect the true relevance level of the document. Agreement is measured by comparing participants' relevance judgments on a summary to the gold standard judgment for the full text represented by that summary. Higher agreement percentages are intended to denote a better quality summary. One variant of a gold-standard measure, *LDC Agreement*, is described in the next section.

Next, we introduce a new measure called *Relevance Prediction* that compares human judgments on a summary with his or her own judgment on the full text document instead of relying on external gold-standard judgments. This approach addresses some of the shortcomings of the SUMMAC studies in that the use of user-centric judgments—rather than an external gold-standard—yields higher agreement rates. In addition, our goals are broader than those of the SUMMAC studies, where the focus was on extrinsic evaluations: we explore both extrinsic and intrinsic measures to determine whether there is a correlation between them.

3.1. LDC Agreement

The University of Pennsylvania's Linguistic Data Consortium (LDC) has used trained annotators to produce gold-standard based judgments for the Topic Detection and Tracking version 3 (TDT-3) corpus (Over & Yen, 2003). We use the term *LDC Agreement* to refer to these judgments as the basis of an extrinsic measure for evaluating summaries. In this approach, individual participants' judgments are compared to gold-standard judgments produced by the LDC annotators. Because the LDC judgments are considered "correct," it is thought that if a summary gives a participant enough information to make the "correct" judgment (the judgment consistent with the gold-standard), then it is a good summary. Likewise, if the summary does not give enough information for the participant to make the "correct" judgment, then it is considered a bad summary.

When we compute LDC Agreement, we focus primarily on the extrinsic measure of *accuracy*, i.e., the sum of the "true positives" (those correctly judged relevant) and the "true negatives" (those correctly judged not relevant) over the total number of judgments. The motivation for choosing accuracy as our primary extrinsic measure of human performance is that, unlike the more general task of IR, our experiments enforce a 50% relevant/irrelevant split across our document sets. This balanced split justifies the inclusion of true negatives in the performance assessment (This would not be true in the general case of IR, where the vast majority of documents in the full search space are cases of true negatives.).

Although accuracy is the primary measure for our analysis, other metrics commonly used in the IR literature are imported (following the lead of the SUMMAC experimenters): precision, recall, and *F*-score. The full set of extrinsic measures is given here:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP refers to *true positives*, TN refers to *true negatives*, FP refers to *false positives*, and FN refers to *false negatives*.

An issue with LDC Agreement is that the use of external gold-standard judgments results in low interannotator agreement rates, as seen in an experiment we will describe in Section 4. We maintain that gold-standards are unreliable and, as stated in other work (Edmundson, 1969; Paice, 1990; Hand, 1997; Jing et al., 1998; Ahmad, Vrusias, & de Oliveira, 2003), there is no 'correct' judgment. Rather, judgments of relevance vary and are based on individual user's beliefs.

3.2. Relevance Prediction

We define an alternative to LDC Agreement—an extrinsic measure called *Relevance Prediction*—where each user builds their own "gold standard" based on the full-text documents. Agreement is measured by comparing users' surrogate-based judgments against their own judgments on the corresponding texts. If a user makes a judgment on a summary consistent with the judgment made on the corresponding full-text document, this signifies that the summary provided enough information to make a reliable judgment. Therefore, the summary should receive a high score. If the user makes a judgment on a summary that is inconsistent with the full text judgment, this implies that the summary is lacking in some way; that it did not provide key information to make a reliable judgment, and should receive a low score.

To calculate the Relevance Prediction score, a user'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 are summed over all judgments for a surrogate type and are 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 users make the same judgment on s that they did on d, we say $j(s,d) = 1$. If users 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:

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

In an experiment described in Section 5, users make relevance judgments on a subset of all the summaries produced by a given system and then they make judgments for the corresponding full texts. This ordering ensures that the user does not make a judgment on an individual summary immediately before seeing the corresponding document.

The results of this experiment demonstrate that this approach yields a more reliable comparison mechanism than that of LDC Agreement because it does not rely on gold-standard judgments provided by other individuals. Specifically, Relevance Prediction can be 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 user would choose (or reject) a document under both conditions.

4. Validation of an automatic measure using LDC-Agreement

This experiment—referred to as *LDC Event Tracking*—investigates the question of whether it is possible to find correlations between automatic intrinsic measures and human task performance using the LDC-Agreement method. The task we have chosen is Event Tracking—a more constrained case of relevance assessment—because it has been reported in NIST Topic Detection and Tracking (TDT) evaluations to provide the basis for more reliable results than were obtained in previous studies that used a more general form of

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

relevance assessment.[3] Our goal is to determine if a correlation exists and, moreover, to verify (using statistical significance tests) that this is a reliable method for validating the intrinsic measure.

In our experiment, a user is given a topic or event description and is asked to judge whether or not a document is related to the specified topic/event based solely on the provided summary or the entire text.[4] An example of an event is: "the bombing of the Murrah Federal Building in Oklahoma City." The user sees a detailed description of what information is considered relevant to an event in a given domain. For instance, in the criminal case domain, information about the crime, the investigation, the arrest, the trial and the sentence is considered relevant.

4.1. Hypotheses

One hypothesis for the LDC Event Tracking experiment is that it is possible to save time using summaries for relevance assessment without adversely impacting the degree of accuracy that would be possible with full documents. This is similar to the "summarization condition test" used in SUMMAC (Mani et al., 2002), with the following differences: (1) the lower baseline is fixed to be the first 75 characters (instead of 10% of the original document size); and (2) all other summaries are also fixed-length (no more than 75 characters), following the NIST document understanding conference (DUC) guidelines (Harman & Over, 2004).

A second hypothesis is that this task supports a very high degree of interannotator agreement, i.e., consistent relevance decisions across human participants. This is similar to the "consistency test" applied in SUMMAC, except that it is applied not just to the full-text versions of the documents, but also to all types of summaries. In addition, to validate the hypothesis, a degree of agreement that was higher than chance agreement was required—e.g., a 0.6 Kappa score as opposed to the 0.5 score for agreement by chance (representing at least a 20% increase). In comparison, the SUMMAC experiments achieved only a 0.38 Kappa score, much lower than that of chance agreement. (The reader is referred to (Carletta, 1996) and (Eugenio & Glass, 2004) for further details on Kappa agreement.)

A third hypothesis is that it is possible to demonstrate a correlation between automatic intrinsic measures and extrinsic task-based measures—most notably, a correlation between ROUGE (the automatic intrinsic measure) and recall (the extrinsic measure)—in order to establish an automatic and inexpensive predictor of human performance. In a previous experiment (Zajic et al., 2004), a high correlation was seen with ROUGE and accuracy, so the aim here is to determine if this correlation is consistent.

Crucially, the validation of this third hypothesis—i.e., finding a positive correlation between the intrinsic and extrinsic measures—will result in the ability to estimate the usefulness of different summarization methods for an extrinsic task in a repeatable fashion without the need to conduct user studies. This is important because, as pointed out by Mani et al. (2002), conducting a user study is extremely labor intensive and requires a large number of human participants in order to establish statistical significance. However, as a part of testing this hypothesis, we must also verify (using statistical significance tests) that this is a reliable method for validating the intrinsic measure.

4.2. Experiment resources and design

We used seven types of automatically generated document surrogates and two types of manually generated surrogates. The automatically generated surrogates were:

- *KWIC* – Keywords in Context (Monz, 2004);
- *GOSP* – Global word selection with localized phrase clusters (Zhou & Hovy, 2003);

[3] As a case in point, we conducted initial studies where we evaluated summaries in a more general relevance-assessment task (Zajic et al., 2004) but found that subjects who had been shown the entire document were only able to agree with each other 75% of the time and they agreed with the allegedly correct answers only 70% of the time. These studies did not allow us to draw any conclusions about summary usefulness or to find correlations between human task performance and intrinsic summarization measures.

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

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

- *ISIKWD* – Topic independent keyword summary (Hovy & Lin, 1997);
- *UTD* – Unsupervised Topic Discovery (Schwartz, Sista, & Leek, 2001);
- *Trimmer* – Fluent headline based on a linguistically-motivated parse-and-trim approach (Dorr, Zajic, & Schwartz, 2003);
- *Topiary* – Hybrid topic list and fluent headline based on integration of UTD and Trimmer (Zajic, Dorr, & Schwartz, R., 2004);
- *First75* – The first 75 characters of the document; used as the lower baseline summary.

The two manual surrogates were:

- *Human* – A human-generated 75 character summary (commissioned for this experiment);
- *Headline* – A human-generated headline associated with the original document.

As a control, we used the entire (original) document, referred to as *Full Text*. Note that *Full Text* is considered an upper baseline and *First75* a lower baseline.

The average number of words ranged from 8 to 12 for the surrogates and 594 for the full text. Except for the full text (which averaged 3696 characters), each system output was constrained to 75 characters, as imposed by the DUC-2004 evaluation.

We selected 20 topics from the portion of the Topic Detection and Tracking version 3 (TDT-3) corpus (Allan et al., 1999) containing Associated Press and New York Times news stories. It is possible that the participants had some prior knowledge about the events, yet it is believed that this would not affect their ability to complete the task. Participants' background knowledge of an event can also make this task more similar to real-world browsing tasks, in which participants are often familiar with the event or topic for which they are searching.

Each topic included an event description and a set of 20 documents taken from the top 100 ranked documents retrieved by the FlexIR information retrieval system (Monz & de Rijke, 2001). Crucially, 50% of each subset contained documents relevant to the topic. Because all 20 documents were somewhat similar to the event, this approach ensured that the task would be more difficult than it would be if documents were chosen from completely unrelated events (where the choice of relevance would be obvious even from a poorly written summary). The documents were long enough to be worth summarizing, but short enough to be read within a reasonably short amount of time.

We recruited 20 students at the University of Maryland at College Park as experiment participants.[5] The 20 participants were divided into 10 user groups, each consisting of two users who saw the same two topics for each system (not necessarily in the same order). By establishing these user groups, it was possible to collect data for an analysis of within-group judgment agreement. Because each system/topic pair was judged by two users, there were a total of $20 \times 2 = 40$ judgments made for each system/topic pair, or 800 total judgments per system (across 20 topics). Thus, the total number of judgments, across 10 systems, was 8000 and each user saw each system twice.

A Latin square design was used to ensure that each user group viewed output from each summarization method and made judgments for all 20 event sets (two event sets per summarization system), while also ensuring that each user group saw a distinct combination of system and event. The system/event pairs were presented in a random order (both across user groups and within user groups), to reduce the impact of topic-ordering and fatigue effects.

Users were given a specific set of rules as part of the instructions on how to determine whether a document should be judged relevant or not relevant. The participants performed the experiment on a Windows or Unix workstation, using a web-based interface that was developed to display the event, document descriptions and to record the judgments. They were timed to determine how long it took each user to make all judgments on an event, although participants were not limited in the amount of time they were allowed to complete the experiment.

[5] The human participants for all of our studies were required to be native-English speakers to ensure that the accuracy of judgments was not degraded by language barriers.

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 1
Results of extrinsic task measures on 10 systems, sorted by accuracy (using LDC Agreement)

System	TP	FP	FN	TN	*A*	*P*	*R*	*F*	*T* (s)
Full text	328	55	68	349	0.851	0.856	0.828	0.842	23.00
Human	302	54	94	350	0.815	0.848	0.763	0.803	7.38
Headline	278	52	118	652	0.787	0.842	0.702	0.766	6.34
ISIKWD	254	60	142	344	0.748	0.809	0.641	0.715	7.59
GOSP	244	57	152	347	0.739	0.811	0.616	0.700	6.77
Topiary	272	88	124	316	0.735	0.756	0.687	0.720	7.60
First75	253	59	143	345	0.748	0.811	0.639	0.715	6.58
Trimmer	235	76	161	328	0.704	0.756	0.593	0.665	6.67
KWIC	297	155	99	249	0.683	0.657	0.750	0.700	6.41
UTD	271	135	125	269	0.675	0.667	0.684	0.676	6.52
HSD, $p < 0.05$					0.099	0.121	0.180	0.147	4.78

4.3. Results and analysis: LDC Agreement

We computed LDC Agreement on each participant's responses. In addition, the time of each individual's decision was measured from a set of log files and is reported in seconds per document. Finally, we computed the ROUGE scores for all summary types and investigated the correlations between these intrinsic measures and the extrinsic LDC-Agreement rates.

4.3.1. Extrinsic evaluation using LDC Agreement

Table 1 shows LDC Agreement in terms of IR metrics, focusing primarily on accuracy, and also the breakdown of *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN) (as defined in Section 3.1), for all 10 systems. The table also shows the average *T* (time) it took users on each document (in s). The rows are sorted by Accuracy, which is the focus for the remainder of this discussion.

We used one-factor repeated-measures ANOVA and found that at least one pair of systems is significantly different. Tukey's Studentized Range criterion, called the honestly significant difference (HSD) (for a description, see (Hinton, 1995)) was used to determine which pairs of systems were significantly different as shown in the bottom row of Table 1. If the difference in measures between two systems is greater then the HSD, then a significant difference between the systems can be claimed. Unfortunately, significant differences with $p < 0.05$ cannot be claimed among any of automatic systems for any of the measures.

Table 1 shows that the Headline system demonstrates some loss in accuracy as compared to the Human system, but not a statistically significant difference. This suggest that users can make judgments with the headline or a human-generated system with almost the same level of accuracy.[6] Similarly, significant differences are not seen between the first 75 characters of the document and either the Headline or Human system.[7] Achieving statistical significance is important for making claims about summary usefulness. We address this point in the later studies with the Relevance Prediction measure described in Section 5.

Although the accuracy differences cannot be shown to be significant across all pairs of systems, the decision-making was sped up significantly—3 times as much (e.g., 7.38 s/summary for the Human system compared to 23 s/document for the Full Text)—by using summaries instead of the full text document. In fact, it is possible that the summaries provide even more of a timing benefit than is revealed by these results. Because the full texts are significantly longer than 3 times the length of the summaries, it is likely that the

[6] One might conclude from Table 1 that existing headlines are good enough—and even human short summaries cannot help further. However, there are many types of documents that do not have headlines, e.g., technical papers (which often have topic-oriented titles that are not as informative as a short summary) and non-text sources (broadcast news, conversational speech, and informal genres like blog pages, etc.). Thus, there are important applications for automatic generation of short summaries—and it is appropriate to design a measure that adequately predicts summary usefulness for these applications.

[7] These results further motivate the need for a different evaluation methodology, in that LDC Agreement does not distinguish among the human-produced systems (Headline and Human) or between the human-produced systems and the first 75 characters of the document on the level of accuracy.

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 2
Kappa score for LDC Agreement and Between-Participant Agreement, sorted by Kappa score

System	Kappa score for LDC Agreement	Between-Participant Agreement
Full Text	0.670	0.840
Human	0.630	0.815
Trimmer	0.610	0.805
Headline	0.600	0.800
GOSP	0.570	0.785
First75	0.556	0.778
ISIKWD	0.492	0.746
Topiary	0.470	0.735
KWIC	0.442	0.721
UTD	0.350	0.680

human users were able to use the bold-faced descriptor words to skim the texts-whereas skimming is less likely for a one-line summary. However, even with skimming, the timing differences are very clear.

Note that the human-generated systems—Text, Human and Headline—performed best with respect to Accuracy, with the Text system as the upper baseline, consistent with the initial expectations. However, the tests of significance indicate that many of the differences in the values assigned by extrinsic measures are small enough to support the use of machine-generated summaries for relevance assessment. For example, four of the seven automatic summarization systems show about a 5% or less decrease in accuracy in comparison with the performance of the Headline system. This validates our first hypothesis: that reading document summaries saves time over reading the entire document text without an adverse impact on accuracy. This finding is consistent with the results obtained further in the previous SUMMAC experiments.

Recall that our second hypothesis is that this task supports a very high degree of interannotator agreement—beyond the low rate of agreement (16–69%) achieved in the SUMMAC experiments. Additionally, a Kappa score of 0.6 (higher than chance agreement) is expected from this task as opposed to the 0.38 Kappa score (lower than chance agreement) of the SUMMAC experiments. Kappa is computed as $(P_A - P_E/1 - P_E)$, where we P_A is taken to be LDC Agreement and P_E to be expected agreement.[8]

We also measured Between-Participant Agreement, defined as follows:

$$\frac{\text{total number of times two participants made same judgment on same doc, sys}}{\text{total number of times two participants judged same doc, sys}}$$

Table 2 shows the Kappa and Between-Participant Agreement scores, sorted by Kappa score. The kappa scores for all systems except UTD are well above the kappa scores for chance agreement (0.5) thus supporting the hypothesis that this task is unambiguous enough that users can perform it with a high level of agreement.

4.3.2. Automatic intrinsic evaluation: ROUGE

Whereas SUMMAC focused only on extrinsic task evaluation, we investigate the problem of validating automatic intrinsic evaluation measures by testing for correlations with extrinsic measures of task performance. The intrinsic measure used in our experiments is ROUGE (Lin & Hovy, 2003), which requires reference summaries for the input documents. Three 75-character summaries were commissioned (in addition to the summaries in the Human system) to use as references. Although we computed 1-grams through 4-grams for both measures, for brevity, we show only the results with 1 and 2-grams—abbreviated as R1 and R2 in Table 3. We also computed ANOVA and found that the differences were statistically significant with $p < 0.05$. The last row of the table shows the honestly significant differences for each measure.

Note that ROUGE yields higher values for *Full Text* than for the automatic methods, e.g., *ISIKWD* and *Topiary*. This was an expected result because the full text contains almost all *n*-grams that appear in the reference summaries.

[8] It is assumed that the expected agreement will be 0.5 because 50% of the documents presented to the users are actually relevant.

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 3
ROUGE scores on 10 systems, sorted by ROUGE-1

System	R1	R2
Full text	0.81808	0.35100
First75	0.25998	0.09824
ISIKWD	0.24188	0.00866
Topiary	0.22476	0.06992
KWIC	0.20265	0.06093
Headline	0.20084	0.04744
GOSP	0.20035	0.06285
Trimmer	0.18901	0.07095
Human	0.16838	0.03872
UTD	0.12802	0.01444
HSD, $p < 0.05$	0.05	0.0289

4.3.3. Correlating ROUGE with LDC Agreement

To test our third hypothesis—demonstrating that intrinsic measures correlate positively with extrinsic measures—the results of the automatic metrics were compared to those of the human system performance. Two methods were used for computing this correlation—Pearson r and Spearman ρ (Siegel & Castellan, 1988)—both of which are commonly used in summarization and machine translation evaluation (see e.g., Lin, 2004, Lin & Och, 2004).

Pearson r is computed as follows:

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

where s_i is the score of system i with respect to a particular measure (e.g., precision) and $\bar{s}$ is the average score over all systems, including the full text. Spearman ρ is used to produce correlation results more suitable for this specific task. It is computed exactly like the Pearson r correlation, but instead of comparing actual scores, one compares the system ranking based on an intrinsic measure with the system ranking based on an extrinsic measure.

In computing the correlations, we treated *Full Text* as an outlier; otherwise, the significantly longer texts would lead to spuriously high correlations due to high ROUGE scores that were purely length-induced. Table 4 shows the Pearson and Spearman correlations between the average system scores assigned by the task-based metrics from Table 1 and the automatic metrics from Table 3. While these results indicate there is a positive correlation in some cases, all positive correlations are rather low. Tests of statistical significance indicate that none of the correlations are statistically significant with $p < 0.05$ (using one-tailed testing).

Computing correlation on the basis of the average performance of a system for all topics has the disadvantage that there are only 10 data points which leads to rather unstable statistical conclusions. In order to increase the number of data points, we redefined a data point to be a system-topic pair, e.g., First75/topic3001 and Topiary/topic3004 are two different data points. In general, a data point is defined as system-i/topic-n, where $i = 1\ldots 10$ (ten summarization systems are compared) and $n = 1\ldots 20$ (20 topics are being used). This new definition of a data point resulted in 180 data points for the current experiment (200 data points minus the 20 data points corresponding to *Full Text*. The resulting correlations are shown in Table 5. As before, the correlations are not very strong, but in some cases, a statistically significant positive correlation can be detected

Table 4
Pearson r and Spearman ρ correlation between extrinsic and intrinsic scores grouped by system (excluding full text)

	Pearson r				Spearman ρ			
	A	*P*	*R*	*F*	*A*	*P*	*R*	*F*
R1	0.229	0.389	−0.271	0.171	0.233	0.083	−0.116	0.300
R2	0.000	0.055	−0.222	−0.051	−0.100	−0.150	−0.350	−0.150

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 5
Pearson r and Spearman ρ correlation between extrinsic and intrinsic scores grouped by system-topic pair (excluding full text)

	Pearson r				Spearman ρ			
	A	*P*	*R*	*F*	*A*	*P*	*R*	*F*
R1	*0.181	*0.178	0.108	*0.170	0.176	0.214	0.095	0.172
R2	0.078	0.057	0.034	0.058	0.104	0.093	0.055	0.097

between certain intrinsic and extrinsic evaluation measures (those marked with a single asterisk (*)). Note that Pearson r indicates significant differences in three cases and Spearman ρ indicates no significant differences. This might be because Spearman ρ is a stricter test that is less likely to cause a Type-I error, i.e., to incorrectly reject the null hypothesis that there is no difference.

Although grouping the individual scores in the form of system-topic pairs resulted in more data points than using only the systems as data points it introduced another source of noise. In particular, given two data points system-*i*/topic-*n* and system-*j*/topic-*m*, where the former has a higher ROUGE-1 score than the latter but a lower accuracy score, the two data points are inversely correlated. The problem is that the reordering of this pair with respect to the two evaluation measures may be caused not only by the quality of the summarization method, but also by the difficulty of the topic. For some topics it is easier to distinguish between relevant and non-relevant documents than for others.

Since our interest lies in the effect of system performance, our aim is to eliminate the effect of topic difficulty while maintaining a reasonable sample size of data points. This was achieved by normalizing each of the original data points in the following way: For each data point we computed the score of the intrinsic measure m_i and the score of the extrinsic measure m_e. Then, for a given data point d, we computed the average score of the intrinsic measure m_i for all data points that used the same topic as d and subtracted the average score from each original data point on the same topic. The same procedure was applied to the extrinsic measure m_e. The goal was to produce a distribution where the data points belonging to the same topic were normalized with respect to their difference from the average score for that topic. Since absolute values were not being used anymore, the distinction between hard and easy topics disappeared.

Table 6 shows the adjusted correlations—using both Pearson and Spearman—for all pairs of intrinsic and extrinsic measures on all systems (again, excluding *Full Text*). Both the Pearson r and Spearman ρ correlations indicate that only one of the pairs shows a statistically significant correlation, viz. ROUGE-1 and Precision at a level of $p < 0.05$.

In all tests above, we were unable to confirm our third hypothesis—that we could demonstrate a correlation between the intrinsic and extrinsic measures, specifically ROUGE-1 and Recall.

4.4. Experimental findings

These experiments show that there is a small yet statistically significant correlation between some of the intrinsic measures and a user's performance in an extrinsic task. Unfortunately, the strength of this correlation depends heavily on the correlation measure: Although Pearson r shows statistically significant differences in a some cases, a stricter non-parametric correlation measure such as Spearman ρ only showed a significant correlation in one case.

The overall conclusion that can be drawn from this experiment is that ROUGE-1 does correlate with precision and to a somewhat lesser degree with accuracy, but that the stability of these correlations remains to be

Table 6
Adjusted Pearson r and Spearman ρ correlation between extrinsic and intrinsic scores grouped by system-topic pair (excluding full text)

	Pearson r				Spearman ρ			
	A	*P*	*R*	*F*	*A*	*P*	*R*	*F*
R1	0.114	*0.195	−0.038	0.082	0.123	*0.248	−0.070	0.064
R2	−0.034	0.015	−0.097	−0.050	0.022	0.072	−0.073	−0.011

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

established. In addition, it is important to determine what differences in ROUGE-1 are needed to yield significant differences in human performance in an extrinsic task.

5. Validation of an automatic measure using Relevance Prediction

The previous experiment demonstrated that a measure that uses low-agreement human-produced annotations does not yield stable results. We argued (in Section 3) that this is a significant hurdle in determining the effectiveness of a summarizer for an extrinsic task such as relevance assessment. Therefore, our second experiment—referred to as *RP with Human Summaries*—explores the human performance scoring and correlations using both the LDC-Agreement method and the new Relevance Prediction method.

For the purpose of this comparison, we simplified the experiment by using only the human-generated summaries—the original news story Headline (*Headline*), and human summaries that were commissioned for this experiment (*Human*).[9] Although neither summary is produced automatically, this experiment focuses more narrowly on summary usefulness and differences in presentation style, rather than on rankings between different automatic summarization systems.

5.1. Hypotheses

One hypothesis for our RP with Human Summaries experiment is that the summaries would allow participants to achieve a Relevance Prediction rate of 70–90%. This rate was predicted because we expected human-produced summaries to yield a rate higher than 50% (higher than that of random judgments) but not as high as 100% (lower than that of judgments made on the full text document).

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

A third hypothesis was also tested: that the 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 and, ultimately, for validating automatic intrinsic measures).

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

5.2. Experiment resources and design

Three distinct events and their related document sets were selected from TDT-3.[10] As in the previous experiment, the 20 documents were selected from a larger set of documents that were automatically retrieved by FlexIR such that exactly half (10) had been judged relevant by the LDC annotators.

We recruited 10 experiment participants to evaluate three different presentation types: the full text documents and two summary types (described below). Each document was pre-annotated with the Headline associated with the original newswire source. These Headline surrogates were used as the first summary type and had an average length of 53 characters. In addition, human-generated summaries were commissioned for each document as the second summary type. The average length of these Human surrogates was 75 characters.

For each event, each of 10 participants was given a description of the event (pre-written by LDC) and then asked to judge relevance of 20 documents associated with that event (using the three different presentation types described above). After reading each document or summary, the participant clicked on a radio button corresponding to their judgment and clicked a *submit* button to move to the next document description. Participants were not allowed to move to the next summary/document until a valid selection was made and no

[9] 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.

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

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 7
Results of extrinsic task measures on three presentation types, sorted by accuracy (using LDC Agreement)

System	TP	FP	FN	TN	*A*	*P*	*R*	*F*	*T* (s)
Full text	226	102	74	198	0.707	0.689	0.753	0.720	13.38
Human	196	90	104	210	0.677	0.685	0.653	0.669	4.57
Headline	171	67	129	233	0.673	0.718	0.570	0.636	4.60
HSD, $p < 0.05$					0.037	0.037	0.057	0.045	7.23

backing up was allowed. Judgment time was computed as the number of seconds it took the participant to read the full text document or surrogate, comprehend it, compare it to the event description, and make a judgment (timed up until the participant clicked the *submit* button).

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

Experiments were conducted using a web browser (Internet Explorer) on a PC in the presence of the experimenter. Participants were given written and verbal instructions for completing their task. For example, in an election event, participants were instructed that documents describing new people in office, new public officials, change in governments or parliaments were potentially relevant.

Two main factors were measured: (1) differences in judgments for the three presentation types (Headline, Human, and the Full Text document) and (2) judgment time. Each participant 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 participant's judgments were constrained to two possibilities, *relevant* or *not relevant*.[11]

5.3. Results and analysis: LDC Agreement and Relevance Prediction

As before, we computed the time and accuracy of each participant's performance. However, in this experiment we computed *both* LDC Agreement (using Accuracy, as before) and Relevance Prediction rates, rather than just one or the other. We then investigated the correlations between ROUGE-1 and both extrinsic measures: LDC Agreement and Relevance Prediction.

5.3.1. Extrinsic evaluation using LDC Agreement and Relevance Prediction

Tables 7 and 8 show the humans' judgments using LDC Agreement and Relevance Prediction, respectively. Using the Relevance Prediction measure, the Human surrogates yield an average of 0.813 for accuracy, significantly higher than the rate of 0.707 for LDC Agreement with $p < 0.01$ (using a paired t-test), thus confirming the first hypothesis. The Relevance Prediction Precision and F-score results were also significantly higher than the LDC Agreement results with $p < 0.01$.

However, the second hypothesis was not confirmed. The Headline Relevance Prediction yielded a rate of 0.760, which was lower than the rate for Human (0.813), but the difference was not statistically significant at the $p < 0.05$ level. We do note that the actual significance level for Relevance Prediction was $p < 0.07$, which is very close to the generally accepted level for significance testing and much better than the level achieved by LDC Agreement ($p < 0.38$). We believe that with additional systems and datapoints, we would expect Relevance Prediction to achieve significance at the 95% level.

[11] If participants were allowed to make additional judgments such as *somewhat relevant*, this could possibly encourage participants 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).

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 8
Results of extrinsic task measures on three presentation types, sorted by accuracy (using Relevance Prediction)

System	TP	FP	FN	TN	*A*	*P*	*R*	*F*	*T* (s)
Human	251	35	77	237	0.813	0.878	0.765	0.818	4.57
Headline	211	27	117	245	0.760	0.887	0.643	0.746	4.60

Table 9
Average ROUGE scores for headline and human surrogates

Surrogate	R1	R2
Headline	0.211	0.068
Human	0.269	0.079

As for the third hypothesis, that the Relevance Prediction measure would be more reliable than that of LDC Agreement, Tables 7 and 8 illustrate a substantial difference between the two agreement measures. The Relevance Prediction rate (Accuracy) is 20% higher for the Human summaries and 13% higher for the Headline summaries. These differences are statistically significant for Human summaries (with $p < 0.01$) and Headline summaries (with $p < 0.05$) using single-factor ANOVA. The higher Relevance Prediction rate supports our hypothesis and confirms that this approach provides a more stable framework for evaluating different summarization techniques.

Finally, the average timing results confirm the fourth hypothesis. The users took 4–5 s (on average) to make judgments on both the Headline and Human summaries, as compared to about 13.4 s to make judgments on full text documents. This shows that it takes users almost 3 times longer to make judgments on full text documents as it took to make judgments on the summaries (Headline and Human). This finding is not surprising since text summaries are an order of magnitude shorter than full text documents.

5.3.2. Automatic intrinsic evaluation: ROUGE

In Section 4, ROUGE was shown to have a positive—but low—correlation with LDC Agreement. This weak correlation was attributed to low interannotator agreement in the gold standard. The goal here is to test whether ROUGE is better correlated with the new Relevance Prediction technique.

Table 9 shows the average ROUGE scores, based on 3 reference summaries per document.[12] (As before, we show only the results with 1 and 2-g—abbreviated as R1 and R2.) The ROUGE scores for Headline surrogates were slightly lower than those for Human surrogates. This is consistent with the earlier statements 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 Human surrogates) makes it more likely that the summary would match the reference.[13]

A summary in which the word choice is less constrained—as in the non-extractive Headline surrogates—is less likely to share *n*-grams with the reference. Thus, non-extractive summaries can be found that have almost identical meanings, but very different words. This raises the concern that ROUGE may be highly sensitive to the style of summarization that is used. Section 5.4 discusses this point further.

5.3.3. Correlating ROUGE with LDC Agreement and Relevance Prediction

To test whether ROUGE correlates more highly with Relevance Prediction than with LDC Agreement, Pearson's *r* (for a full definition, refer back to Section 4.3.3) is used to determine the correlations for the results of both techniques. We restricted our attention to ROUGE-1, which has been shown to have the highest correlations with human judgments on headlines in DUC (Harman & Over, 2004).

[12] A total of 180 human-generated reference summaries (3 for each of 60 documents) were commissioned (in addition to the human generated summaries used in the experiment).
[13] Recently, Zhou, Lin, Munteanu, & Hovy (2006) used paraphrases to overcome the issue of exact word-matching.

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

Table 10
Pearson correlations with ROUGE-1 for Relevance Prediction (RP) and LDC-Agreement (LDC), where partition size (P) = 1, 2, and 4

Surrogate	$P = 1$	$P = 2$	$P = 4$
Headline (RP)	0.127	0.194	0.314
Human (RP)	0.063	0.109	0.139
Headline (LDC)	−0.096	−0.066	−0.009
Human (LDC)	−0.039	−0.023	−0.018

Since there are only 3 systems: Human, Headline, and Full Text, it would not be very meaningful to compute the correlation between the different measures with only three points. There are 3 distinct topics, but this is still a relatively small number of points. Each judgment could be considered an independent data point, but in this case, the ROUGE-1 scores would be computed on single summaries and the agreement would be either zero or one, which would make it difficult to compute correlations. Therefore, we created groups of documents that would be scored together, so that the agreement score would be continuous and there would be enough data in each group to be able to have a meaningful number.

We tested 3 ways of partitioning the data; with 1, 2, or 4 documents (or their summaries) in each group. Partitions of size 4 provide a reasonable tradeoff between having a good estimate and having several data points. (Larger partition sizes would result in too few data points and compromise the statistical significance of the correlation results).

In order to increase the number of data points, we chose 10,000 random sets of 4 documents from the test set. Obviously, these data points are not independent, so the statistical significance is still based on the original number of samples. But grouping the documents in many ways provides a smoother estimate of the correlation between the different measures. This idea of partitioning the data are similar to the idea of re-sampling for the bootstrap significance test (Davison & Hinkley, 1997).

To correlate the partitioned agreement scores with the intrinsic measure, ROUGE-1 was also run on all 120 individual surrogates in the experiment (i.e., the Human and Headline surrogates for each of the 60 event/document pairs) and the resulting scores were averaged for all surrogates belonging to the same partitions (for each of the three partition sizes). These partitioned ROUGE-1 values were then used for detecting correlations with the corresponding partitioned agreement scores described above.

Across partitions, the max/min Relevance Prediction rates for Headline and Human surrogates (0.93/0.60 and 0.98/0.68, respectively) were all higher than the corresponding LDC Agreement rates (0.85/0.50 and 0.88/0.55, respectively). This provides further support for our hypothesis that Relevance Prediction produces better results than LDC Agreement for evaluation of summary usefulness.

Table 10 shows the Pearson Correlations between ROUGE-1 and both Relevance Prediction and LDC Agreement. As one might expect, there is some variability in the correlation between ROUGE and human judgments for the different partitions. However, the standard deviation for both Headline (0.179) and Human (0.162) indicates that the variabilities in both cases are rather small.

For Relevance Prediction, a positive correlation for both surrogate types was observed, with a slightly higher correlation for Headline than for Human. For LDC Agreement, no correlation (or a minimally negative one) was observed with ROUGE-1 scores, for both the Headline and Human surrogates. The highest correlation was observed for Relevance Prediction on Headline.

The conclusion is that ROUGE correlates more highly with the Relevance Prediction measurement than the LDC-Agreement measurement, although it must be noted that none of the correlations in Table 10 were statistically significant at $p < 0.05$. The low LDC-Agreement scores are consistent with previous studies where poor correlations were attributed to low interannotator agreement rates.

5.4. Experimental findings

As observed above, many of the Headline surrogates were not actually summaries of the full text, but were eye-catchers. Often, these surrogates did not allow the user 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. (It was noticed that most words in the Human surrogates

appeared in the corresponding stories.) There were three consequences of this difference between Headline and Human: (1) The rate of agreement was lower for Headline than for Human; (2) The average ROUGE score was lower for Headline than for Human; and (3) The correlation of ROUGE scores with agreement was higher for Headline than for Human.

A further analysis explains the (somewhat counterintuitive) third point above. We computed the ROUGE scores for the true positives/negatives and false positives/negatives, for both Headline and Human surrogates. We found that the average ROUGE-1 scores for true positives and true negatives for Headline surrogates (0.2127 and 0.2162) were significantly lower than the corresponding scores for Human surrogates (0.2696 and 0.2715). On the other hand, the number of false negatives was substantially higher for Headline surrogates than for Human surrogates (see Table 8) and these corresponded to much lower ROUGE scores for Headline surrogates (0.1996) than for Human (0.2586) surrogates.

Although there were very few false positives (less than 27 and 35—i.e., under 6%—for Headline and Human, respectively), the number of false negatives was particularly high for Headline (50% higher than for Human). This difference was statistically significant at $p < 0.01$ using the t-test. The large number of false negatives with Headline may be attributed to the eye-catching nature of these surrogates. A user 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 user 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), it is speculated that, if a correlation exists between Relevance Prediction and ROUGE, the false negatives may be a major contributing factor.

Based on this experiment, it is conjectured 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.

6. Memory and priming study

One concern with the evaluation methodology associated with Relevance Prediction is the issue of possible memory effects or priming: if the same users saw a summary and a full document about the same event, their judgments for the second system may be biased by the information provided by the first system. Thus, we conducted an experiment—referred to as *Document Presentation Methods*—to determine whether the order in which summaries and corresponding full text documents are displayed can affect user's judgments.

Ten different summary and document orderings were tested, with the presentation methods ranging from an extreme form of influence—the summary and full text being presented in immediate succession—to a method where the information source (e.g. summary) is presented on one week and the alternative source (e.g. full text) is presented one week later.

Two of the methods that were used were labeled D1S2D2 and S1D1D2. In D1S2D2, the user saw only the document set on week one, and on week two, the user saw the corresponding summary set followed by the same document set. For S1D1D2, the user saw the reverse: a summary set and the corresponding document set on week one, and then the same document set on week two.

Table 11 shows these labels as column headers, with underlining to indicate the judgments that are being compared. D1S2D2 (in column 1) refers to the percentage of summary judgments in the second week (S2) that match the corresponding full document judgments in that same week (D2). D1S2D2 (in column 2) refers to the percentage of full document judgments in the first week (D1) that match those same full document judgments in the second week (D2). S1D1D2 (in column 3) refers to the percentage of summary judgments in the first

Table 11
Two-user study to compare summary/document judgments both across and within weeks

	D1S2D2	D1S2D2	S1D1D2	S1D1D2
User 1	70%	100%	70%	100%
User 2	60%	100%	50%	90%

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002

week (S1) that match the corresponding full document judgments in that same week (D1). Finally, S1D1D2 (in column 4) refers to the percentage of full document judgments in the first week (D1) that match those same full document judgments in the second week (D2).

Two study participants were recruited through emailed experiment advertisements. Our results indicate that User 1's judgment remained the same for both cases, and User 2 changed only a single judgment.[14] From this, we can conclude that the order in which the summaries and corresponding full texts are shown do not bias the user's selections for subsequent judgments. The judgments users made on a document after seeing its corresponding summary were almost the same when they were presented with the document only. (For a full description of the experiment and its results, see (President & Dorr, 2006).) This study demonstrates that it is possible to use our Relevance Prediction approach reliably.

7. Conclusion and future work

This work has led to a number of important contributions, most notably the introduction of a new method for measuring agreement on extrinsic tasks called Relevance Prediction. This method was compared with the previous gold-standard based LDC-Agreement method and was shown to be more reliable for evaluation of summary usefulness. The validity of the approach was confirmed with a memory and priming study.

Although Relevance Prediction has thus far been used only for human-generated summaries, our future work will incorporate both human and automatic summaries. This will allow us to further investigate the reliability of the Relevance Prediction method, the human performance and correlation differences with automatic summaries, and to make a comparison of both types of summaries with the upper baseline (the full text document) and the lower baseline (first 75 characters of the document). In addition, the experiments described above used only short, 75 character summaries generated from a single source document (single document summarization). With the shift in focus of the summarization community to multi-document summarization (Dang, 2005), our subsequent experiments will investigate 150 character summaries and will also apply Relevance Prediction to tasks involving multi-document summaries.

It is expected that this work yields a usable framework for testing hypotheses and drawing definitive conclusions about summary usefulness, thus justifying continued research and development of new summarization methods.

Acknowledgements

We are indebted to David Zajic for insights and comments regarding the Relevance Prediction measure. The second author thank Steve, Carissa, and Ryan for their energy enablement. This work has been supported, in part, under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001 and the Nuffield Foundation, Grant No. NAL/32720. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

References

Ahmad, K., Vrusias, B., & de Oliveira, P. C. F. (2003). Summary evaluation and text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*. Toronto, Canada (July).

Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., & Lavrenko, V. (1999). Topic-based novelty detection. *Tech. Rep. 1999 summer workshop at CLSP final report*. Maryland: Johns Hopkins.

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing and Management, 31*(5), 675–685.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 22*(2), 249–254, June.

Dang, H. T. (2005). Overview of DUC 2005. In *Proceedings of the document understanding conferences (DUC)*. Vancouver: Canada (October).

[14] Note, that in comparing columns 1 and 3 for User 2, the percentage changes by 10%—a single judgment. The same is seen in comparing columns 2 and 4 of Table 11.

Davison, A., & Hinkley, D. (Eds.). (1997). *Bootstrap methods and their application*. Cambridge, United Kingdom: Cambridge University Press.

Dorr, B. J., Zajic, D., & Schwartz, R. (2003). Hedge trimmer: a parse-and-trim approach to headline generation. In *Proceedings of the human language technology – North American chapter of the association for computational linguistics (HLT-NAACL) text summarization workshop*. Alta., Canada (May).

Edmundson, H. P. (1969). New methods in automatic abstracting. *Journal of the Association for Computing Machinery, 16*(2), 264–285.

Eugenio, B. D., & Glass, M. (2004). Squibs and discussions – the kappa statistic: a second look. *Computational Linguistics*, 95–101.

Hand, T. F. (1997). A proposal for task-based evaluation of text summarization systems. In *Proceedings of the ACL/EACL-97 summarization workshop*. Madrid: Spain (July).

Harman, D., & Over, P. (2004). In *Proceedings of the document understanding conference (DUC) 2004*. Boston, MA.

Hinton, P. R. (1995). *Statistics explained: A guide for social science students*. New York, NY: Routledge.

Hovy, E., & Lin, C.-Y. (1997). Automated text summarization in SUMMARIST. In *Proceedings of the association for computational linguistics (ACL) workshop on intelligent scalable text summarization*. Madrid, Spain (August).

Hovy, E., Lin, C.-Y., & Zhou, L. (2005). Evaluating DUC 2005 using basic elements. In *Proceedings of the document understanding conferences (DUC)*. Vancouver, Canada (October).

Jing, H., Barzilay, R., McKeown, K., & Elhadad, M. (1998). Summarization evaluation methods: experiments and analysis. In *Proceedings of the AAAI symposium on intelligent summarization*. Stanford University, CA (March 23–25).

LDC (2006). Data annotation. Linguistic Data Consortium, University of Pennsylvania: Philadelphia, PA. <http://www.ldc.upenn.edu>.

Lin, C.-Y. (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).

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using *N*-gram co-occurrence statistics. In *Proceedings of the joint annual meeting of human language technology (HLT) and the North American chapter of the association for computational linguistics (HLT-NAACL)* (pp. 71–78). Edmonton, Canada (May–June).

Lin, C.-Y., & Och, F.J. (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).

Lin, J., & Demner-Fushman, D. (2005). Automatically evaluating answers to definition questions. In *Proceedings of the 2005 human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP 2005)* (pp. 931–938). Vancouver, Canada.

Mani, I. (2001). Summarization evaluation: an overview. In *Proceedings of the North American chapter of the association for computational linguistics (NAACL) workshop on automatic summarization*.

Mani, I., Klein, G., House, D., & Hirschman, L. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering, 8*(1), 43–68.

Monz, C. (2004). Minimal span weighting retrieval for question answering. In *Proceedings of the special interest group on information retrieval (SIGIR) workshop on information retrieval for question answering*. Pittsburgh, PA (May).

Monz, C., & de Rijke, M. (2001). The university of Amsterdam at CLEF 2001. In *Proceedings of the cross language evaluation forum workshop (CLEF 2001)* (pp. 165–169). Darmstadt, Germany (September).

Nenkova, A., & Passonneau, R. J. (2004). Evaluating content selection in summarization: the pyramid method. In *Proceedings of the joint annual meeting of human language technology and the North American chapter of the association for computational linguistics (HLT-NAACL)*. Boston, MA (May).

Over, P., & Yen, J. (2003). An introduction to DUC-2003: intrinsic evaluation of generic news text summarization systems. In *Proceedings of the DUC 2003 workshop on text summarization*. <http://duc.nist.gov/pubs/2003slides/duc2003intro.pdf>.

Paice, C. D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management, 26*(1), 171–186.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics (ACL 2004)*. Philadelphia, PA (July).

President, S., & Dorr, B. (2006). Text summarization evaluation: correlating human performance on an extrinsic task with automatic intrinsic metrics. Tech. rep., University of Maryland, College Park, MD, LAMP-TR-133, CS-TR-4808, UMIACS-TR-2006-28.

Schwartz, R., Sista, S., & Leek, T. 2001. Unsupervised topic discovery. In: *Proceedings of the advanced research and development activity in information technology (ARDA) workshop on language modeling and information retrieval*. Pittsburgh, PA (May).

Siegel, S., & Castellan, N. J. Jr., (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Tombros, A., & Sanderson, M. (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* (pp. 2–10).

Zajic, D., Dorr, B. J., & Schwartz, R. (2004). BBN/UMD at DUC 2004: topiary. In *Proceedings of the document understanding conference (DUC)*. Boston, MA (May).

Zajic, D., Dorr, B. J., Schwartz, R., & President, S. (2004). Headline evaluation experiment results. Tech. rep., University of Maryland, College Park, MD, UMIACS-TR-2004-18.

Zhou, L., & Hovy, E. (2003). Web-trained extraction summarization system. In: *Proceedings of the joint annual meeting of human language technology (HLT) and the North American chapter of the association for computational linguistics (HLT-NAACL)*. Alta., Canada (May).

Zhou, L., Lin, C.-Y., Munteanu, D. S., & Hovy, E. (2006). ParaEval: using paraphrases to evaluate summaries automatically. In *Proceedings of HLT/NAACL-2006* (pp. 447–454).

Please cite this article in press as: Hobson, S. P. et al., Task-based evaluation of text summarization using ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.01.002