A Comparison of Rankings Produced by Summarization Evaluation Measures

Robert L. Donaway
Department of Defense
9800 Savage Rd. STE 6409
Ft. Meade, MD 20755-6409
rldonaw@super.org

Kevin W. Drummey
Department of Defense
9800 Savage Rd. STE 6341
Ft. Meade, MD 20755-6341
kwdrumm@super.org

Laura A. Mather
La Jolla Research Lab
Britannica.com, Inc.
3253 Holiday Ct. Suite 208
La Jolla, CA 92037
mather@us.britannica.com

Abstract

Summary evaluation measures produce a ranking of all possible extract summaries of a document. Recall-based evaluation measures, which depend on costly human-generated ground truth summaries, produce uncorrelated rankings when ground truth is varied. This paper proposes using sentence-rank-based and content-based measures for evaluating extract summaries, and compares these with recall-based evaluation measures. Content-based measures increase the correlation of rankings induced by synonymous ground truths, and exhibit other desirable properties.

1 Introduction

The bulk of active research in the automatic text summarization community centers on developing algorithms to produce extract summaries, e. g. (Schwarz, 1990), (Paice and Jones, 1993), (Kupiec et al., 1995), (Marcu, 1997), (Strzalkowski et al., 1998), and (Goldstein et al., 1999). Yet understanding how to evaluate their output has received less attention. In 1997, TIPSTER sponsored a conference (SUMMAC) where various text summarization algorithms were evaluated for their performance in various tasks (Mani et al., 1999; Firmin and Chrzanowski, 1999). While extrinsic evaluation measures such as these are often very concrete, the act of designing the task and scoring the results of the task introduces bias and subject-based variability. These factors may confound the comparison of summarization algorithms. Machine-generated summaries also may be evaluated intrinsically by comparing them with "ideal" human-generated summaries. However, there is often little agreement as to what constitutes the ideal summary of a document.

Both intrinsic and extrinsic methods require time consuming, expert human input in order to evaluate summaries. While the traditional methods have many advantages, they are costly, and human assessors cannot always agree on summary quality. If a numerical measure were available which did not depend on human judgement, researchers and developers would be able to immediately assess the effect of modifications to summary generation algorithms. Also, such a measure might be free of the bias that is introduced by human assessment.

This paper investigates the properties of various numerical measures for evaluating the quality of *generic*, *indicative* document summaries. As explained by Mani et al. (1999), a generic summary is not topic-related, but "is aimed at a broad readership community" and an indicative summary tells "what topics are addressed in the source text, and thus can be used to alert the user as to source content." Section 2 discusses the properties of numerical evaluation measures, points out several drawbacks associated with intrinsic measures and introduces new measures developed by the authors. An experiment was devised to compare the new evaluation measures with the traditional ones. The design of this experiment is discussed in Section 3 and its results are presented in Section 4. The final section includes conclusions and a statement of the future work related to these evaluation measures.

2 Evaluation Measures

An evaluation measure produces a numerical score which can be used to compare different summaries of the same document. The scores are used to assess summary quality across a collection of test documents in order to produce an average for an algorithm or system. However, it must be emphasized that the scores are

most significant when considered per document. For example, two different summaries of a document may have been produced by two different summarization algorithms. Presumably, the summary with the higher score indicates that the system which produced it performed better than the other system. Obviously, if one system consistently produces higher scores than another system, its average score will be higher, and one has reason to believe that it is a better system. Thus, the important feature of any summary evaluation measure is not the *value* of its score, but rather the *ranking* its score imposes on a set of extracts of a document.

To compare two evaluation measures, whose scores may have very different ranges and distributions, one must compare the order in which the measures rank various summaries of a document. For instance, suppose a summary scoring function Y is completely dependent upon the output of another scoring function X, such as $Y = 2^X$. Since Y is an increasing function of X, both X and Y will produce the same ranking of any set of summaries. However, the scores produced by Y will have a very different distribution than those of X and the two sets of scores will not be correlated since the dependence of Y on X is non-linear. Therefore, in order to compare the scores two different measures assign to a set of summaries, one must compare the ranks they assign, not the actual scores.

The ranks assigned by an evaluation measure produce equivalence classes of extract summaries; each rank equivalence class contains summaries which received the same score. When a measure produces the same score for two different summaries of a document, there is a *tie*, and the equivalence class will contain more than one summary. All summaries in an equivalence class must share the same rank; let this rank be the midrank of the range of ranks that would have be assigned if each score were distinct. An evaluation measure should posses the following properties: (i) higher-ranking summaries are more effective or are of higher quality than lower-ranking summaries, and (ii) all of the summaries in a rank equivalence class are more-or-less equally effective.

The following sections contrast the ranking properties of three types of evaluation measures: recall-based measures, a sentence-rank-based measure and content-based measures. These types of measures are defined, their properties are described and their use is explained.

2.1 Recall-Based Evaluation Measures

Recall-based evaluation measures are intrinsic. They compare machine-generated summaries with sentences previously extracted by human assessors or judges. From each document, the judges extract sentences that they believe make up the best extract summary of the document. A summary of a document generated by a summarization algorithm is typically compared to one of these "ground truth" summaries by counting the number of sentences the ground truth summary and the algorithm's summary have in common. Thus, the more sentences a summary has *recalled* from the ground truth, the higher its score will be. See work by Goldstein et al. (1999) and Jing et al. (1998) for examples of the use of this measure.

The recall-based measures introduce a bias since they are based on the opinions of a small number of assessors. It is widely acknowledged (Jing et al., 1998; Kupiec et al., 1995; Voorhees, 1998) that assessor agreement is typically quite low. There are at least two sources of this disagreement. First, it is possible that one human assessor will pick a particular sentence for inclusion in their summary when the content of another sentence or set of sentences is approximately equivalent. Jing et al. (1998) agree: "...precision and recall are not the best measures for computing document quality. This is due to the fact that a small change in the summary output (e.g., replacing one sentence with an equally good equivalent which happens not to match majority opinion [of the assessors]) can dramatically affect a system's score." We call this source of summary disagreement 'disagreement due to *synonymy*.' Here is an example of two human-generated extracts from the same 1991 *Wall Street Journal* article which contain different sentences, but still seem to be describing an article about violin playing in a film:

> EXTRACT 1: Both Ms. Streisand's film husband, played by Jeroen Krabbe, and her film son, played by her real son Jason Gould, are, for the purposes of the screenplay, violinists. The actual sound – what might be called a fiddle over – was produced off camera by Pinchas Zucker-

man. The violin program in "Prince of Tides" eliminates the critic's usual edge and makes everyone fall back on his basic pair of ears.

EXTRACT 2: Journalistic ethics forbid me from saying if I think "Prince of Tides" is as good as "Citizen Kane," but I don't think it's wrong to reveal that the film has some very fine violin playing. But moviegoers will hear Mr. Zuckerman cast off the languor that too often makes him seem like the most bored of great violinists. With each of these pieces, Mr. Zuckerman takes over the movie and shows what it means to play his instrument with supreme dash.

Another source of disagreement can arise from judges' differing opinions about the true focus of the original document. In other words, judges disagree on what the document is about. We call this second source 'disagreement due to *focus*.' Here is a human-generated extract of the same article which seems to differ in focus:

EXTRACT 3: Columbia Pictures has delayed the New York City and Los Angeles openings of "Prince of Tides" for a week. So Gothamites and Angelenos, along with the rest of the country, will have to wait until Christmas Day to see this film version of the Pat Conroy novel about a Southern football coach (Nick Nolte) dallying with a Jewish female psychotherapist (Barbra Streisand) in the Big Apple. Perhaps the postponement is a sign that the studio is looking askance at this expensive product directed and co-produced by its female lead.

Whatever the source, disagreements at the sentence level are prevalent. This has serious implications for measures based on a single opinion, when a slightly different opinion would result in a significantly different score (and rank) for many summaries.

For example, consider the following three-sentence ground truth extract of a 37-sentence 1994 *Los Angeles Times* article from the TREC collection. It contains sentences 1, 2 and 13.

(1) Clinton Trade Initiative Sinks Under G-7 Criticism. (2) President Clinton came to the high-profile Group of Seven summit to demonstrate new strength in foreign policy but instead watched his premier initiative sink Saturday under a wave of sharp criticism. (13) The negative reaction to the president's trade proposal came as a jolt after administration officials had built it up under the forward-looking name of "Markets 2000" and had portrayed it as evidence of his interest in leading the other nations to more open trade practices.

An extract that replaces sentence 13 with sentence 5:

(5) In its most elementary form, it would have set up a one-year examination of impediments to world trade, but it would have also set an agenda for liberalizing trade rules in entirely new areas, such as financial services, telecommunications and investment.

will receive the same recall score as one which replaces sentence 13 with sentence 32:

(32) Most nations have yet to go through this process, which they hope to complete by January.

These two alternative summaries both have the same recall rank, but are obviously of very different quality.

Considered quantitatively, the only important component of either precision or recall is the 'sentence agreement' J, the number of sentences a summary has in common with the ground truth summary. Following Goldstein et al. (1999), let M be the number of sentences in a ground truth extract summary and let K be the number of sentences in a summary to be evaluated. With precision $P = J/K$ and recall $R = J/M$ as usual, and $F_1 = 2PR/(P+R)$, then elementary algebra shows that $F_1 = 2J/(M+K)$. Often, a fixed summary length K is used. (In terms of word count, this represents varying compression rates.) When a particular ground truth of a given document is chosen, then precision, recall and F_1 are all constant multiples of J. As such, these measures produce different *scores*, but the same *ranking* of all the K-sentence extracts from the document. Since only this ranking is of interest, it is not necessary to examine more than one of P, R and F_1.

The sentence agreement J can only take on integer values between 0 and M, so J, P, R,

and F_1 are all discrete variables. Therefore, although there may be thousands of possible extract summaries of a document, only $M+1$ different scores are possible. This will obviously create a large number of ties in rankings produced by the P, R, and F_1 scores, and will greatly increase the probability that radically different summaries will be given the same score and rank. On the other hand, two summaries which express the same ideas using different sentences will be given very different scores. Both of these problems are illustrated by the example above. Furthermore, if a particular ground truth includes a large proportion of the document's sentences (perhaps it is a very concise document), shorter summaries will likely include only sentences which appear in the ground truth. Consequently, even a randomly selected collection of sentences might obtain the largest possible score. Thus, recall-based measures are likely to violate both properties (i) and (ii), discussed at the beginning of Section 2. These inherent weaknesses in recall-based measures will be further explored in Section 4.

2.2 A Sentence-Rank-Based Measure

One way to produce ground truth summaries is to ask judges to rank the *sentences* of a document in order of their importance in a generic, indicative summary. This is often a difficult task for which it is nearly impossible to obtain consistent results. However, sentences which appear early in a document are often more indicative of the content of the document than are other sentences. This is particularly true in newspaper articles, whose authors frequently try to give the main points in the first paragraph (Brandow et al., 1995). Similarly, adjacent sentences are more likely to be related to each other than to those which appear further away in the text. Thus, sentence position alone may be an effective way to rank the importance of sentences.

To account for sentence importance within a ground truth, a summary comparison measure was developed which treats an extract as a ranking of the sentences of the document. For example, a document with five sentences can be expressed as $(1,2,3,4,5)$. A particular extract may include sentences 2 and 3. Then if sentence 2 is more important than sentence 3, the sentence ranks are given by $(4,1,2,4,4)$. Sentences 1, 4, and 5 all rank fourth, since 4 is the midrank of ranks 3, 4 and 5. Such rank vectors can be compared using Kendall's tau statistic (Sheskin, 1997), thus quantifying a summary's agreement with a particular ground truth. As will be shown in Section 4, sentence rank measures result in a smaller number of ties than do recall-based evaluation measures.

Although it is also essentially recall-based, the sentence rank measure has another slight advantage over recall. Suppose a ground truth summary of a 20-sentence document consists of sentences $\{2,3,5\}$. The machine-generated summaries consisting of sentences $\{2,3,4\}$ and $\{2,3,9\}$ would receive the same recall score, but $\{2,3,4\}$ would receive a higher tau score (5 is closer to 4 than to 9). Of course, this higher score may not be warranted if the content of sentence 9 is more similar to that of sentence 5.

The use of the tau statistic may be more appropriate for ground truths produced by classifying all of the sentences of the original document in terms of their importance to an indicative summary. Perhaps four different categories could be used, ranging from 'very important' to 'not important.' This would allow comparison of a ranking with four equivalence classes (representing the document) to one with just two equivalence classes (representing inclusion and exclusion from the summary to be evaluated).

2.3 Content-Based Measures

Since indicative summaries alert users to document content, any measure that evaluates the quality of an indicative summary ought to consider the similarity of the content of the summary to the content of the full document. This consideration should be independent of exactly which sentences are chosen for the summary. The content of the summary need only capture the general ideas of the original document. If human-generated extracts are available, machine-generated extracts may be evaluated alternatively by comparing their contents to these ground truths. This section defines content-based measures by comparing the term frequency (tf) vectors of extracts to tf vectors of the full text or to tf vectors of a ground truth extract. When the texts and summaries are tokenized and token aliases are determined by a thesaurus, summaries that disagree due to *synonymy* are likely to have similarly-distributed

term frequencies. Also, summaries which happen to use synonyms appearing infrequently in the text will not be penalized in a summary-to-full-document comparison. Note that term frequencies can always be used to compare an extract with its full text, since the two will always have terms in common, but without a thesaurus or some form of term aliasing, term frequencies cannot be used to compare abstracts with extracts.

The vector space model of information retrieval as described by Salton (1989) uses the inner product of document vectors to measure the content similarity $sim(d_1, d_2)$ of two documents d_1 and d_2. Geometrically, this similarity measure gives the cosine of the angle between the two document vectors. Since $\cos 0 = 1$, documents with high cosine similarity are deemed similar. We apply this concept to summary evaluation by computing document-summary content similarity $sim(d, s)$ or ground truth-summary content similarity $sim(g, s)$.

Note that when comparing a summary with its document, a prior human assessment is not necessary. This may serve to eliminate the ambiguity of a human assessor's bias towards certain types of summaries or sentences. However, the scores produced by such evaluation measures cannot be used reliably to compare summaries of drastically different lengths, since a much longer summary is more likely than a short summary to produce a term frequency vector which is similar to the full document's tf vector, despite the normalization of the two vectors. (This contrasts with the bias of recall towards short summaries.)

This similarity measure can be enhanced in a number of ways. For example, using term frequency counts for a large corpus of documents, term weighting (such as log-entropy (Dumais, 1991) or tf-idf (Salton, 1989)) can be used to weight the terms in the document and summary vectors. This may improve the performance of the similarity measure by increasing the weights of content-indicative terms and decreasing the weights of those terms that are not indicative of content. It is demonstrated in Section 4 that term weighting caused a significant increase in the correlation of the rankings produced by different ground truths; however, it is not clear that this weighting increases the scores of high quality summaries.

There are two potential problems with using the cosine measure to evaluate the performance of a summarization algorithm. First of all, it is likely that the summary vector will be very sparse compared to the document vector since the summary will probably contain many fewer terms than the document. Second, it is possible that the summary will use key terms that are not used often in the document. For example, a document about the merger of two banks, may use the term "bank" frequently, and use the related (yet not exactly synonymous) term "financial institution" only a few times. It is possible that a high quality extract would have a low cosine similarity with the full document if it contained only those few sentences that use the term "financial institution" instead of "bank." Both of these problems can be addressed with another common tool in information retrieval: latent semantic indexing or LSI (Deerwester et al., 1990).

LSI is a method of reducing the dimension of the vector space model using the singular value decomposition. Given a corpus of documents, create a term-by-document matrix A where each row corresponds to a term in the document set and each column corresponds to a document. Thus, the columns of A represent all the documents from the corpus, expressed in a particular term-weighting scheme. (In our testing, the document vectors' entries are the relative frequencies of the terms.) Compute the singular value decomposition (SVD) of this matrix (for details see Golub and van Loan (1989)). Retain some number of the largest singular values of A and discard the rest. In general, removing singular values serves as a dimension reduction technique. While the SVD computation may be time-consuming when the corpus is large, it needs to be performed only once to produce a new term-document matrix and a projection matrix. To calculate the similarity of a summary and a document, the summary vector s must also be mapped to this low-dimensional space. This involves computing a vector-matrix product, which can be done quickly.

The effect of using the scaled, reduced-dimension document and summary vectors is two-fold. First, each coordinate of both the document and summary vector will contribute to

the overall similarity of the summary to the document (unlike the original vector space model, where only terms that occur in the summary contribute to the cosine similarity score). Second, the purpose of using LSI is to reduce the effect of near-synonymy on the similarity score. If a term occurs infrequently in the document but is highly indicative of the content of the document, as in the case where the infrequent term is synonymous with a frequent term, the summary will be penalized less in the reduced-dimension model for using the infrequent term than it would be in the original vector space model. This reduction in penalty occurs because LSI essentially averages the weights of terms that co-occur frequently with other terms (both "bank" and "financial institution" often occur with the term "account"). This should improve the accuracy of the cosine similarity measure for determining the quality of a summary of a document.

3 Experimental Design

This section describes the experiment that tests how well these summary evaluation metrics perform. Fifteen documents from the Text Retrieval Conference (TREC) collection were used in the experiment. These documents are part of a corpus of 103 newspaper articles. Each of the documents was tokenized by a language processing algorithm, which performed token aliasing. In our experiments, the term set was comprised of all the aliases appearing in the full corpus of 103 documents. This corpus was used for the purposes of term weighting. Four expert judges created extract summaries (ground truths) for each of the documents. A list of the first 15 documents, along with some of their numerical features is found in Table 1. The judges were instructed to select as many sentences as were necessary to make an "ideal" indicative extract summary of the document. In terms of the count of sentences in the ground truth, the lengths of the summaries varied from document to document. Ground truth compression rates were generally between 10 and 20 percent. The inter-assessor agreement also varied, but was often quite high. We measured this by calculating the average pairwise recall in the collection of four ground truths.

A suite of summary evaluation measures $\{E_k\}$ which produce a score for a summary was developed. These measures may depend on none, one, or all of the collection of ground truth summaries $\{g_j\}$. Measures which do not depend on ground truth compute the summary-document similarity $sim(s, d)$. Content-based measures which depend on a single ground truth g_i compute the summary-ground truth similarity $sim(s, g_i)$. A measure which depends on all of the ground truths $g_1, \ldots, g_4$, computes a summary's similarity with each ground truth separately and averages these values. Table 2 enumerates the 28 different evaluation measures that were compared in this experiment. Note that the Recall and Kendall measures require a ground truth.

In this study, the measures will be used to evaluate extract summaries of a fixed sentence length K. In all of our tests, $K = 3$ for reasons of scale which will become clear. A summary length of three sentences represents varying proportions of the number of sentences in the full text document, but this length was usually comparable to the lengths of the human-generated ground truths. For each document, the collection $\{S_j\}$ was generated. This is the set of all possible K-sentence extracts from the document. If the document has N sentences total, there will be N choose K

$$\binom{N}{K} = \frac{N!}{K!\,(N-K)!}$$

extracts in the exhaustive collection $\{S_j\}$. *The focus now is only on the set of all possible summaries and the evaluation measures, and not on any particular summarization algorithm.* For each document, each of the measures in $\{E_k\}$ was used to rank the sets $\{S_j\}$. (Note that the measures which do not depend on ground truths could, in fact, be used to *generate* summaries if it were possible to produce and rank the exhaustive set of fixed-length summaries in real time. Despite the authors' access to impressive computing power, the process took several hours for each document!) The next section compares these different rankings of the exhaustive set of extracts for each document.

4 Experimental Results

One way to compare the different rankings produced by two different evaluation measures is to

Table 1: Test Document & Summary Statistics

Doc. No.	TREC File Name	Sent. Count	Token Count	Gnd. Truth Sent. Cnt.	Gnd. Truth Avg. Recall
1	WSJ911211-0057	34	667	3, 4, 12, 3	44%
2	WSJ900608-0126	34	603	4, 4, 9, 3	54%
3	WSJ900712-0047	18	364	2, 3, 5, 2	78%
4	latwp940604.0027	23	502	4, 5, 5, 4	69%
5	latwp940621.0116	27	579	12, 11, 10	84%
6	latwp940624.0094	17	460	5, 5, 5, 4	79%
7	latwp940707.0400	33	503	6, 9, 8, 8	52%
8	latwp940709.0051	37	877	3, 5, 5, 4	53%
9	latwp940713.0013	34	702	9, 4, 5, 8	35%
10	latwp940713.0014	30	528	6, 5, 7, 5	88%
11	latwp940721.0080	28	793	3, 3, 5, 2	88%
12	latwp940725.0030	36	690	9, 2, 7, 5	45%
13	latwp940725.0128	18	438	6, 3, 5, 5	63%
14	latwp940729.0109	25	682	4, 3, 4, 3	96%
15	latwp940801.0010	28	474	4, 5, 4, 5	43%

Table 2: Evaluation Measures

Similarity Measure	Details	Ground Truth Dependency					
		None	g_1	g_2	g_3	g_4	All
Recall	J_i/M, $J_i = \#(s \cap g_i)$	N.A.	E_1	E_2	E_3	E_4	E_5
Kendall Tau	see Section 2.2	N.A.	E_6	E_7	E_8	E_9	E_{10}
tf Cosine	$sim(s, d$ or $g_i)$ on tf vectors	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}
tf-idf Cosine	$sim(s, \cdot)$ on tf-idf weighted vectors	E_{17}	E_{18}	E_{19}	E_{20}	E_{21}	E_{22}
SVD tf Cosine	$sim(s, \cdot)$ on low-dim. vectors	E_{23}	E_{24}	E_{25}	E_{26}	E_{27}	E_{28}

calculate their Spearman rank correlation coefficient. When two evaluation measures produce nearly the same ranking of the summary set, the rank correlation will be near 1 and a scatterplot of the two rankings will show points nearly lying on a line with slope 1. When there is little correlation between two rankings, the statistic will be near 0 and the scatterplot will appear to have randomly-distributed points. A negative correlation indicates that one ranking often reverses the rankings of the other and in this case a rank scatterplot will show points nearly lying on a line with negative slope.

Table 3 compares the Spearman correlation of the rankings produced by a specific pair of ground truths. The first row contains the correlations of two highly similar ground truth extracts of document 14. Both of these extracts consisted of three sentences; two of the sentences were common to both extracts. Not surprisingly, the correlation is high regardless of what measure produced the rankings. The second row demonstrates an increase (across the row) in correlation between rankings produced by two different ground truth summaries of document 8. These two ground truths did not disagree in focus, but did disagree due to *synonymy* — they contain just one sentence in common. In general, the correlation among the rankings produced by synonymous ground truths was increased most by using the SVD content-based comparison. Figure 1 illustrates the correlation increase graphically for this pair of ground truths. By contrast, the third row of Table 3 displays a decrease (across the row) in correlation between rankings produced by two different ground truths. In this case, the two ground truths disagreed in *focus*: they are Extracts 2 and 3 contrasted in Section 2.1. Again, the correlation among the rankings produced by the four ground truths was decreased most by using a weighted content-based comparison such

Table 3: Correlation of Ground Truths Depends on Level of Disagreement

	recall	tau	tf cosine	tf-idf	SVD
Agree Sentences	0.87	0.96	0.95	0.87	0.99
Disagree synonymy	0.34	0.37	0.53	0.72	0.96
Disagree focus	0.22	0.31	0.32	0.20	-0.29

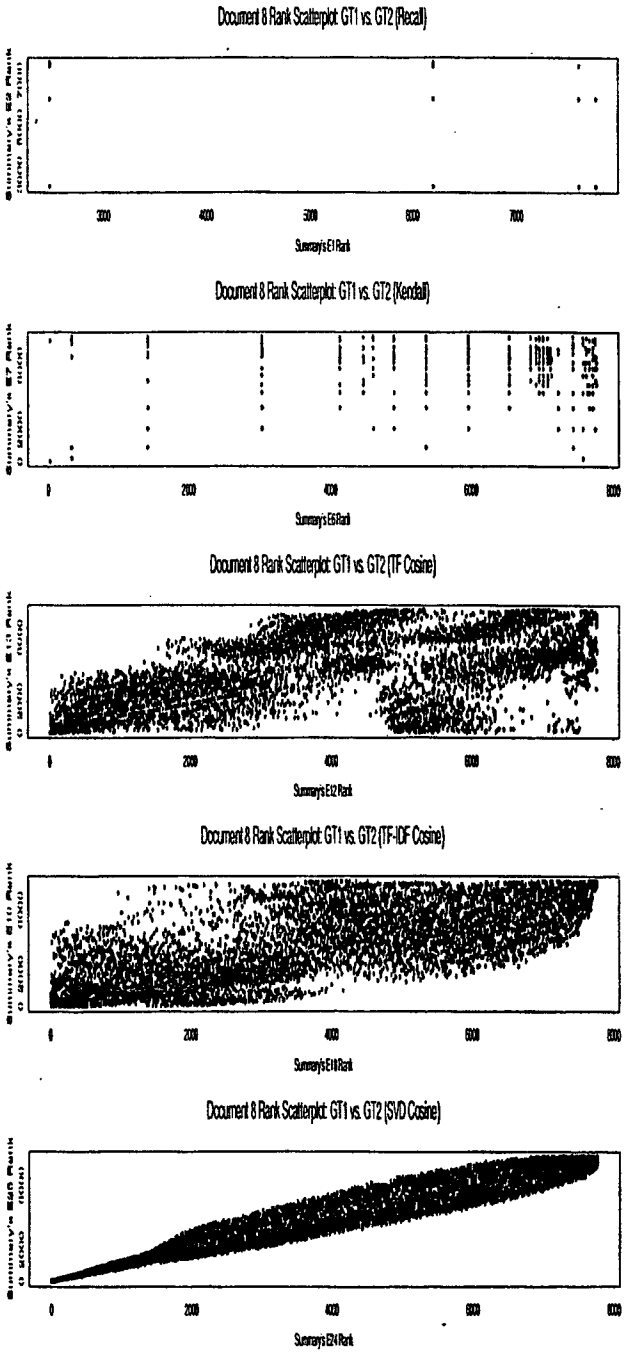
as tf-idf or SVD. These patterns were typical for rankings produced by ground truths which differed in focus, allaying the fear that applying the SVD weighting would produce correlated rankings based on *any* two ground truths.

Of course, the lack of correlation among recall-based rankings whenever ground truths did not contain exactly the same sentences implies that a different collection of extracts would rank highly if one ground truth were replaced with the other. This effect would surely carry through to system averages across a set of documents. To exemplify the size of this effect, for each document, the summaries which scored highest using one ground truth were scored (using recall) against a second ground truth. With the first ground truths, these high-scoring summaries averaged over 75% recall; using the second ground truths, the same summaries averaged just over 25% recall. Thus, by simply changing judges, an automatic system which produced these summaries would appear to have a very different success rate. This disparity is lessened when content-based measures are used, but the outcomes are still disparate.

Evidence suggests that the content-based measures which do not rely on a ground truth may be an acceptable substitute to those which do. Over the set of 15 documents, the average within-document inter-assessor correlation is 0.61 using term frequency, 0.72 using tf-idf, and 0.67 using SVD. The average correlation of the ground truth dependent measures with those that perform summary-document comparisons is 0.48 using term frequency, 0.70 using tf-idf, and 0.56 using SVD. This means that on average, the rankings based on single ground truths are only slightly more correlated to each other than they are to the rankings that do not depend on any ground truth.

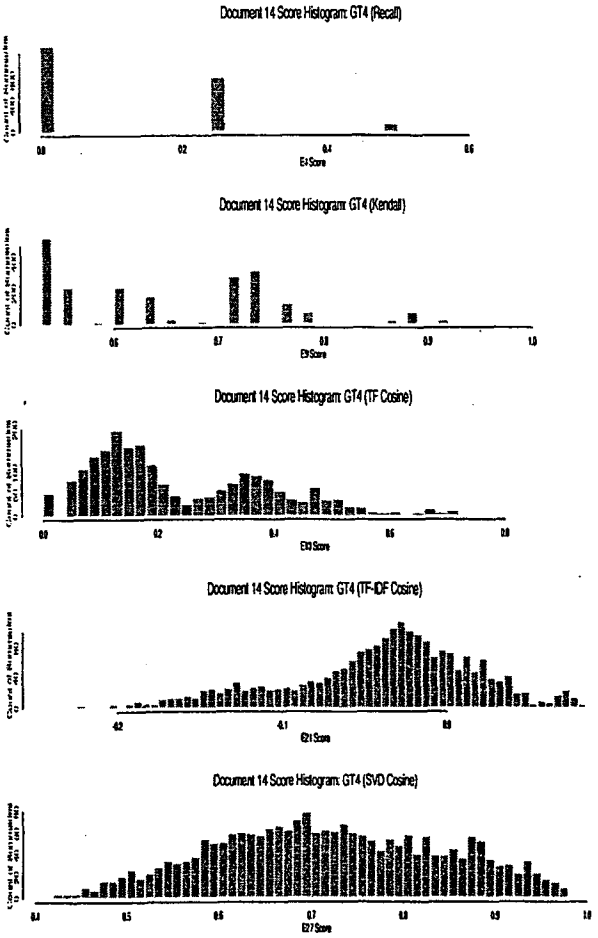
As noted in Section 2.1, the recall-based measures exhibit unfavorable scoring properties. Figure 2 shows the histogram of scores assigned to the exhaustive summary set for doc-

Figure 1: Synonymy: Content-based Measures Increase Rank Correlation



ument 14 by five different measures. Each of these measures was based on the same ground truth summary of this document, which contained four sentences. Clearly, the measures based on a more sophisticated parsing method have a much greater ability to discriminate between summaries. By contrast, the recall metric can assign one of only four scores to a length 3 summary, based on the value of J. Elementary combinatorics shows that 4 extracts will receive the highest possible score (and thus will rank first), 126 summaries will rank second, 840 summaries will rank third, and 1330 summaries will rank last (with a score of 0). This accounts for all of the 2300 three-sentence extracts that are possible. It seems very unlikely that all of the second-ranking summaries are equally effective. The histogram depicting this distribution is shown at the top of Figure 2. This is followed by the histograms for the Kendall metric, and the content-based metrics using term frequency, tf-idf, and SVD weighted vectors, respectively. The tf-idf and SVD weighted measures produced a very fine distribution of scores, particularly near the top of the range. That is, these metrics are able to distinguish between different high-scoring summaries. These patterns in the score histograms were typical across the 15 documents.

Figure 2: Score Histograms for Document 14



5 Conclusions and Future Work

There is wide variation in the rankings produced by recall scores from non-identical ground truths. This difference in scores is reflected in averages computed across documents. The low inter-assessor correlation of ranks based on recall measures is distressing, and indicates that these measures cannot be effectively used to compare performances of summarization systems. Measures which gauge content similarity produce more highly correlated rankings whenever ground truths do not disagree in focus. Content-based measures assign different rankings when ground truths do disagree in focus. In addition, these measures provide a finer grained score with which to compare summaries.

Moreover, the content-based measures which rely on a ground truth are only slightly more correlated to each other than they are to the measures which perform summary-document comparisons. This suggests that the effectiveness of summarization algorithms could be measured without the use of human judges. Since the cosine measure is easy to calculate, feedback of summary quality can be almost instantaneous.

The properties of these content-based measures need to be further investigated. For example, it is not clear that content-based measures satisfy properties (i) and (ii), discussed in Section 2. Also, while they do produce far fewer ties than either recall or tau, such a fine distinction in summary quality is probably not justified. When human-generated ground truths are available, perhaps some combination of recall and the content-based measures could be used. For instance, whenever recall is not perfect, the content of the non-overlapping sentences could be compared with the missed ground truth sentences. Also, the effects of compression rate,

summary length, and document style are not known.

The authors are currently performing further experiments to see if users prefer summaries that rank highly with content-based measures over other summaries. Also, the outcomes of extrinsic evaluation techniques will be compared with each of these scoring methods. In other words, do the high-ranking summaries help users to perform various tasks better than lower-ranking summaries do?

6 Acknowledgements

The authors would like to thank Mary Ellen Okurowski and Duncan Buell for their support, encouragement, and advice throughout this project. Thanks go also to Tomek Strzalkowski, Inderjeet Mani, Donna Harman, and Hal Wilson for their suggestions of how to improve the design of the experiment. We greatly appreciate the fine editing advice Oksana Lassowsky provided. Finally, we are especially grateful to the four expert judges, Benay, Ed, MEO, and Toby, who produced our ground truth summaries.

References

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Susan T. Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments & Computers*, 23(2):229–236.

Therese Firmin and Michael J. Chrzanowski. 1999. An evaluation of automatic text summarization systems. In *Advances in Automatic Text Summarization*, chapter 21, pages 325–336. MIT Press, Cambridge, Massachusetts.

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the ACM SIGIR*, pages 121–128.

Gene H. Golub and Charles F. van Loan. 1989. *Matrix Computations*. The Johns Hopkins University Press, Baltimore.

Hongyan Jing, Kathleen McKeown, Regina Barzilay, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *American Association for Artificial Intelligence Spring Symposium Series*, pages 60–68.

Julian Kupiec, Jan Pederson, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the ACM SIGIR*, pages 68–73.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of the Ninth Conference of the European Chapter of the ACL*, pages 77–85.

Daniel Marcu. 1997. From discourse structure to text summaries. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88.

C. D. Paice and P. A. Jones. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the ACM SIGIR*, pages 69–78.

Gerard Salton. 1989. *Automatic Text Processing*. Addison-Wesley Publishers, Massachusetts.

C. Schwarz. 1990. Content based text handling. *Information Processing and Management*, 26(2):219–226.

David J. Sheskin. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press LLC, United States.

T. Strzalkowski, J. Wang, and B. Wise. 1998. A robust practical text summarizaton system. In *AAAI Intelligent Text Summarization Workshop*, pages 26–30.

Ellen M. Voorhees. 1998. Variations in relevance judgements and the measurement of retrieval effectiveness. In *Proceedings of the ACM SIGIR*, pages 315–323.