Manual and Automatic Evaluation of Summaries

Chin-Yew Lin and Eduard Hovy

USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
+1-310-448-8711/8731
{cyl,hovy}@isi.edu

Abstract

In this paper we discuss manual and automatic evaluations of summaries using data from the Document Understanding Conference 2001 (DUC-2001). We first show the instability of the manual evaluation. Specifically, the low inter-human agreement indicates that more reference summaries are needed. To investigate the feasibility of automated summary evaluation based on the recent BLEU method from machine translation, we use accumulative n-gram overlap scores between system and human summaries. The initial results provide encouraging correlations with human judgments, based on the Spearman rank-order correlation coefficient. However, relative ranking of systems needs to take into account the instability. Finally, we provide a method of estimating the reliability of scoring by learning to emulate human judgments. Various forms of mismatching between system and human summaries due to synonyms, paraphrases, and regeneration vs. extraction still must be solved to achieve satisfactory correlation.

1 Introduction

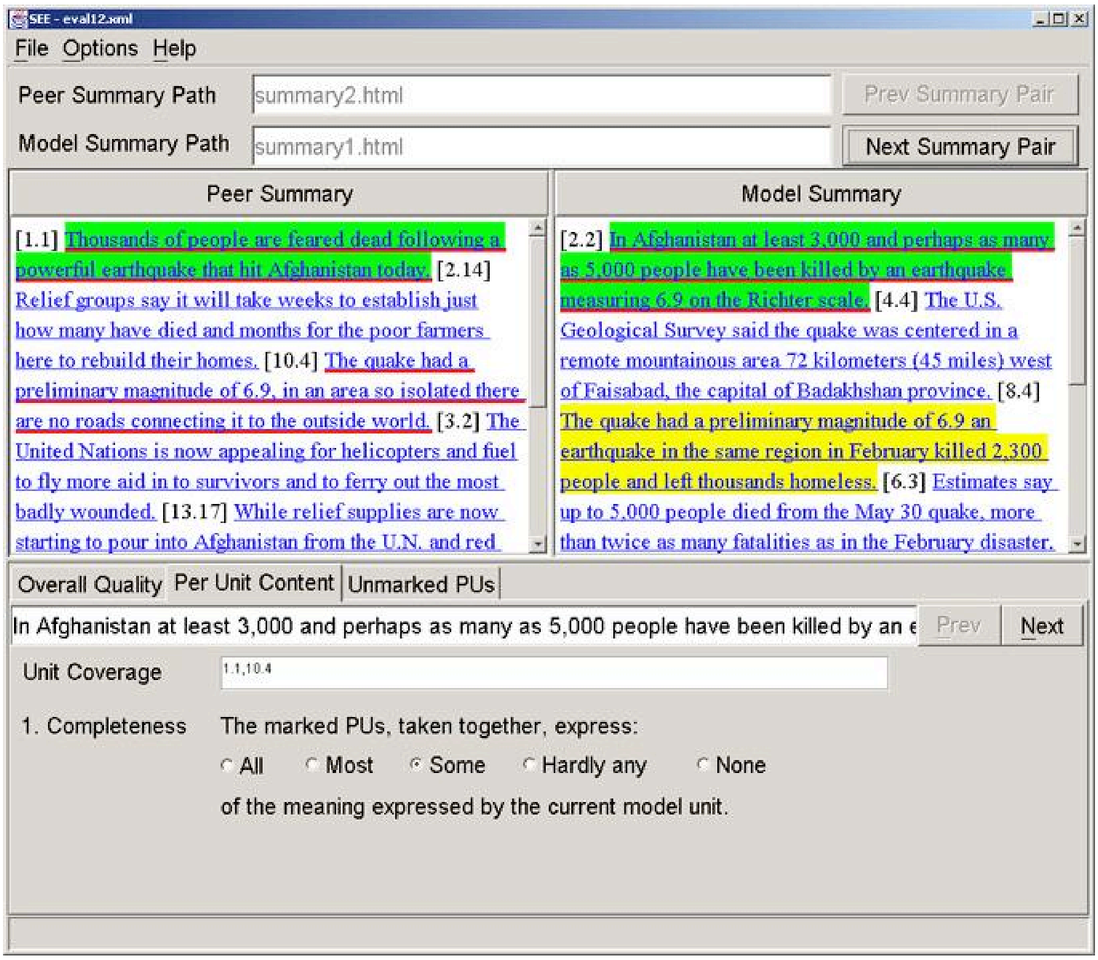
Previous efforts in large-scale evaluation of text summarization include TIPSTER SUMMAC (Mani et al. 1998) and the Document Understanding Conference (DUC) sponsored by the National Institute of Standards and Technology (NIST). DUC aims to compile standard training and test collections that can be shared among researchers and to provide common and large scale evaluations in single and multiple document summarization for their participants.

In this paper we discuss manual and automatic evaluations of summaries using data from the Document Understanding Conference 2001 (DUC-2001). Section 2 gives a brief overview of the evaluation procedure used in DUC-2001 and the Summary Evaluation Environment (SEE) interface used to support the DUC-2001 human evaluation protocol. Section 3 discusses evaluation metrics. Section 4 shows the instability of manual evaluations. Section 5 outlines a method of automatic summary evaluation using accumulative n-gram matching score (NAMS) and proposes a view that casts summary evaluation as a decision making process. It shows that the NAMS method is bounded and in most cases not usable, given only a single reference summary to compare with. Section 6 discusses why this is so, illustrating various forms of mismatching between human and system summaries. We conclude with lessons learned and future directions.

2 Document Understanding Conference (DUC)

DUC2001 included three tasks:

- Fully automatic single-document summarization: given a document, participants were required to create a generic 100-word summary. The training set comprised 30 sets of approximately 10 documents each, together with their 100-word human written summaries. The test set comprised 30 unseen documents.
- Fully automatic multi-document summarization: given a set of documents about a single subject, participants were required to create 4 generic summaries of the entire set, containing 50, 100, 200, and 400 words respectively. The document sets were of four types: a single natural disaster



Peer Summary Path: summary2.html

Model Summary Path: summary1.html

Peer Summary:

[1.1] Thousands of people are feared dead following a powerful earthquake that hit Afghanistan today. [2.14] Relief groups say it will take weeks to establish just how many have died and months for the poor farmers here to rebuild their homes. [10.4] The quake had a preliminary magnitude of 6.9, in an area so isolated there are no roads connecting it to the outside world. [3.2] The United Nations is now appealing for helicopters and fuel to fly more aid in to survivors and to ferry out the most badly wounded. [13.17] While relief supplies are now starting to pour into Afghanistan from the U.N. and red

Model Summary:

[2.2] In Afghanistan at least 3,000 and perhaps as many as 5,000 people have been killed by an earthquake measuring 6.9 on the Richter scale. [4.4] The U.S. Geological Survey said the quake was centered in a remote mountainous area 72 kilometers (45 miles) west of Faisabad, the capital of Badakhshan province. [8.4] The quake had a preliminary magnitude of 6.9 an earthquake in the same region in February killed 2,300 people and left thousands homeless. [6.3] Estimates say up to 5,000 people died from the May 30 quake, more than twice as many fatalities as in the February disaster.

In Afghanistan at least 3,000 and perhaps as many as 5,000 people have been killed by an e

Unit Coverage: 1.1,10.4

1. Completeness The marked PUs, taken together, express:

○ All ○ Most ● Some ○ Hardly any ○ None

of the meaning expressed by the current model unit.

Figure 1. SEE in an evaluation session.

event; a single event; multiple instances of a type of event; and information about an individual. The training set comprised 30 sets of approximately 10 documents, each provided with their 50, 100, 200, and 400-word human written summaries. The test set comprised 30 unseen sets.

- Exploratory summarization: participants were encouraged to investigate alternative approaches to evaluating summarization and report their results.

A total of 11 systems participated in the single-document summarization task and 12 systems participated in the multi-document task.

The training data were distributed in early March of 2001 and the test data were distributed in mid-June of 2001. Results were submitted to NIST for evaluation by July 1st 2001.

2.1 Evaluation Materials

For each document or document set, one human summary was created as the 'ideal' model summary at each specified length. Two other human summaries were also created at each length. In addition, baseline summaries were created automatically for each length as reference points. For the multi-document summarization task, one baseline, *lead baseline*, took the first 50, 100, 200, and 400 words in the last document in the collection. A second baseline, *coverage baseline*, took the first sentence in the first document, the first sentence in the second document and so on until it had a summary of 50, 100, 200, or 400 words. Only one baseline (baseline1) was created for the single document summarization task.

2.2 Summary Evaluation Environment

NIST assessors who created the 'ideal' written summaries did pairwise comparisons of their summaries to the system-generated summaries, other assessors' summaries, and baseline summaries. They used the Summary Evaluation Environment (SEE) 2.0 developed by one of the authors (Lin 2001) to support the process. Using SEE, the assessors compared the system's text (the *peer* text) to the ideal (the *model* text). As shown in Figure 1, each text was decomposed into a list of units and displayed in separate windows. In DUC-2001 the sentence was used as the smallest unit of evaluation.

SEE 2.0 provides interfaces for assessors to judge both the content and the quality of summaries. To measure content, assessors step through each model unit, mark all system units sharing content with the current model unit (shown in green highlight in the model summary window), and specify that the marked system units express *all, most, some* or *hardly any* of the content of the current model unit. To measure quality, assessors rate grammaticality[1], cohesion[2], and coherence [3] at five different levels: *all, most, some, hardly any*, or *none*.

For example, as shown in Figure 1, an assessor marked system units 1.1 and 10.4 (shown in red underlines) as sharing *some* content with the current model unit 2.2 (highlighted green).

3 Evaluation Metrics

One goal of DUC-2001 was to debug the evaluation procedures and identify stable metrics that could serve as common reference points. NIST did not define any official performance metric in DUC-2001. It released the raw evaluation results to DUC-2001 participants and encouraged them to propose metrics that would help progress the field.

3.1 Recall, Coverage, Retention and Weighted Retention

Recall at different compression ratios has been used in summarization research to measure how well an automatic system retains important content of original documents (Mani and Maybury 1999). Assume we have a system summary S_s and a model summary S_m. The number of sentences occurring in S_s is N_s, the number of sentences in S_m is N_m, and the number in both S_s and S_m is N_a. Recall is defined as N_a/N_m. The Compression Ratio is defined as the length of a summary (by words or sentences) divided by the length of its original document.

Applying this direct all-or-nothing recall in DUC-2001 without modification is not appropriate because:

1. Multiple system units contribute to multiple model units.
2. Exact overlap between S_s and S_m rarely occurs.
3. Overlap judgment is not binary.

For example in Figure 1, an assessor judged system units 1.1 and 10.4 sharing *some* content with model unit 2.2. Unit 1.1 says "*Thousands of people are feared dead*" and unit 2.2 says "*3,000 and perhaps … 5,000 people have been killed*". Are "thousands" equivalent to "3,000 to 5,000" or not? Unit 10.4 indicates it was an "*earthquake of magnitude 6.9*" and unit 2.2 says it was "*an earthquake measuring 6.9 on the Richter scale*". Both of them report a "6.9" earthquake. But the second part of system unit 10.4, "*in an area so isolated…*", seems to share some content with model unit 4.4 "*the quake was centered in a remote mountainous area*". Are these two equivalent? This example highlights the difficulty of judging the content coverage of system summaries against model summaries and the inadequacy of using simple recall as defined.

For this reason, NIST assessors not only marked the segments shared between system units (SU) and model units (MU), they also indicated the degree of match, i.e., *all*, *most*, *some*, *hardly any*, or *none*. This enables us to compute *weighted recall.*

Different versions of weighted recall were proposed by DUC-2001 participants. (McKeown et al. 2001) treated the completeness of coverage as a threshold: 4 for *all*, 3 for *most* and above, 2 for *some* and above, and 1 for *hardly any* and above. They then proceeded to compare system performances at different threshold levels. They defined recall at threshold *t*, $Recall_t$, as follows:

[1] Does the summary observe English grammatical rules independent of its content?

[2] Do sentences in the summary fit in with their surrounding sentences?

[3] Is the content of the summary expressed and organized in an effective way?

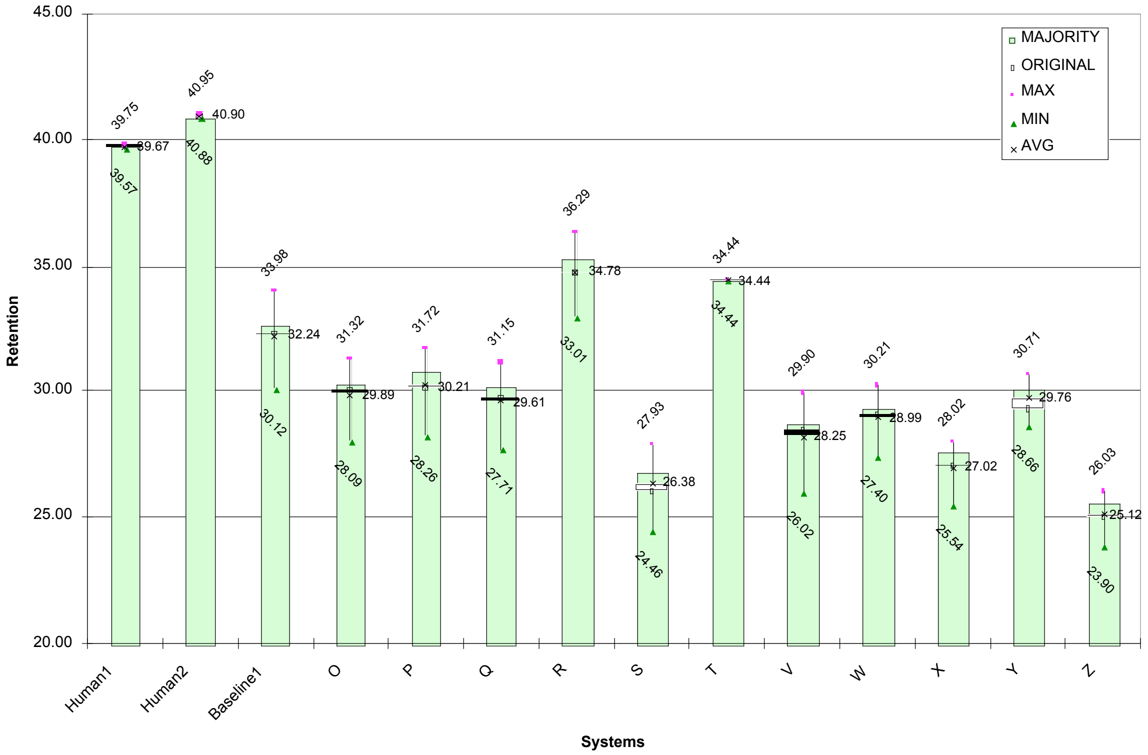


Figure 2. DUC 2001 single-document retention score distribution.

$$\frac{\text{Number of MUs marked at or above } t}{\text{Total number of MUs in the model summary}}$$

Instead of thresholds, we use here as coverage score the ratio of completeness of coverage C: 1 for *all*, 3/4 for *most*, 1/2 for *some*, 1/4 for *hardly any*, and 0 for *none*. To avoid confusion with the recall used in information retrieval, we call our metric weighted retention, $Retention_w$, and define it as follows:

$$\frac{(\text{Number of MUs marked}) \bullet C}{\text{Total number of MUs in the model summary}}$$

If we ignore C (set it to 1), we obtain an unweighted retention, $Retention_1$. We used $Retention_1$ in our evaluation to illustrate that relative system performance (i.e., system ranking) changes when different evaluation metrics are chosen. Therefore, it is important to have common and agreed upon metrics to facilitate large scale evaluation efforts.

4 Instability of Manual Judgments

In the human evaluation protocol described in Section 2, nothing prevents an assessor from assigning different coverage scores to the same system units produced by different systems against the same model unit. (Since most systems produce extracts, the same sentence may appear in many summaries, especially for single-document summaries.) Analyzing the DUC-2001 results, we found the following:

- Single document task
  - A total of 5,921 judgments
  - Among them, 1,076 (18%) contain multiple judgments for the same units
  - 143 (2.4%) of them have three different coverage scores
- Multi-document task
  - A total of 6,963 judgments
  - Among them 528 (7.6%) contain multiple judgments
  - 27 (0.4%) of them have three different coverage scores

Intuitively this is disturbing; the same phrase compared to the same model unit should always have the same score regardless of which system produced it. The large percentage of multiple judgments found in the single document evaluation are test-retest errors that need to be addressed in computing performance metrics.

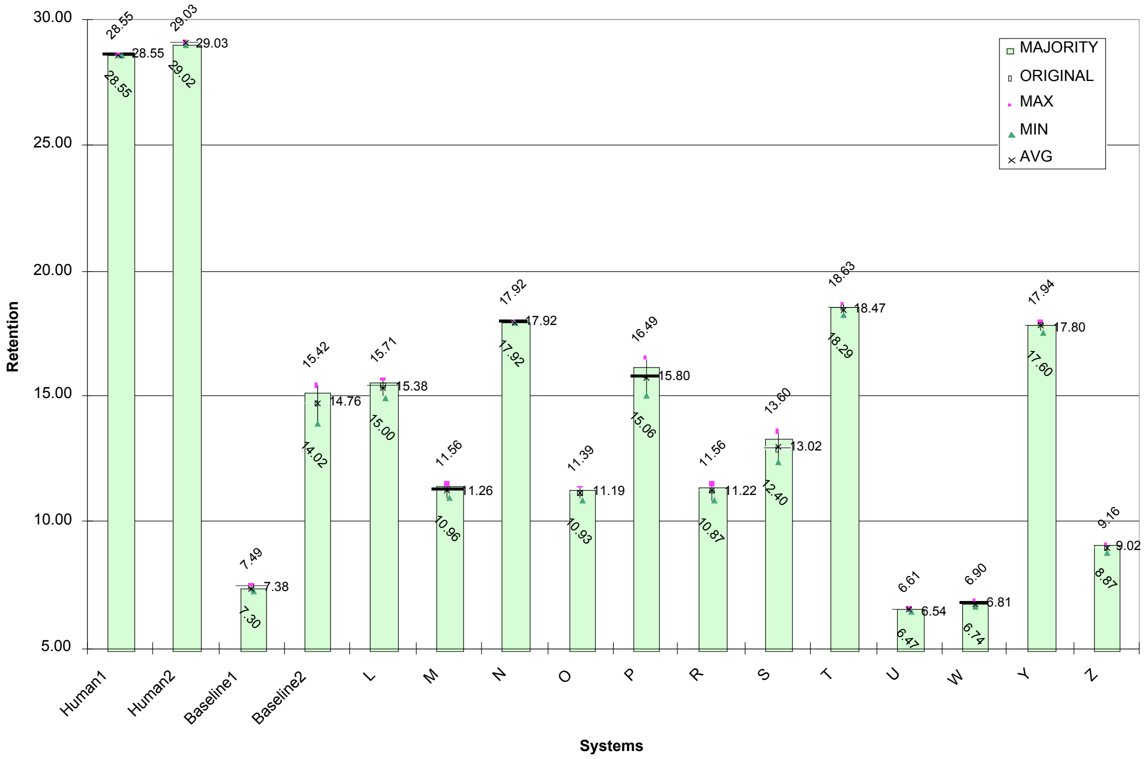


Figure 3. DUC 2001 multi-document retention score distribution.

Figure 2 and Figure 3 show the retention scores for systems participating in the single- and multi-document tasks respectively. The error bars are bounded at the top by choosing the maximum coverage score (MAX) assigned by an assessor in the case of multiple judgment scores and at the bottom by taking the minimum assignment (MIN). We also compute system retentions using the majority (MAJORITY) and average (AVG) of assigned coverage scores. The original (ORIGINAL) does not consider the instability in the data.

Analyzing all systems' results, we made the following observations.

(1) Inter-human agreement is low in the single-document task (~40%) and even lower in multi-documents task (~29%). This indicates that using a single model as reference summary is not adequate.

(2) Despite the low inter-human agreement, human summaries are still much better than the best performing systems.

(3) The relative performance (rankings) of systems changes when the instability of human judgment is considered. However, the rerankings remain local; systems remain within performance groups. For example, we have the following groups in the multi-document summarization task (Figure 3):

a. {Human1, Human2}
b. {N, T, Y}
c. {Baseline2, L, P}
d. {S}
e. {M, O, R}
f. {Z}
g. {Baseline1, U, W}

The existence of stable performance regions is encouraging. Still, given the large error bars, one can produce 162 different rankings of these 16 systems. We conclude the following:

(1) We need more than one model summary.

(2) We need more than one evaluation for each summary against each model summary.

(3) We need to ensure a single rating for each system unit.

In the next section we discuss automatic summary evaluation.

5 Automatic Summary Evaluation

	Original	No stemmer			Lovin stemmer			Porter stemmer		
SYSCODE	Retention	X-1G	X-2G	X-3G	L-1G	L-2G	L-3G	P-1G	P-2G	P-3G
	ranking	(unigram)			(unigram)			(unigram)		
Human1	1	2	1	1	1	1	1	1	1	1
Human2	2	1	2	2	2	2	2	2	2	2
Baseline1	14	15	15	15	16	15	14	16	15	14
Baseline2	8	8	7	6	8	8	6	8	8	6
L	7	7	6	7	7	7	7	7	7	7
M	10	10	10	10	10	11	11	9	10	11
N	4	4	4	4	4	4	4	4	4	4
O	11	12	12	12	12	12	12	11	12	12
P	6	5	5	5	5	5	5	5	5	5
R	11	11	11	11	11	10	10	12	11	10
S	9	9	9	9	9	9	9	9	9	9
T	3	3	3	3	3	3	3	3	3	3
U	16	14	14	14	14	14	15	14	14	15
W	15	16	16	16	15	16	16	15	16	16
Y	5	6	8	8	6	6	8	6	6	8

Figure 4. Manual and automatic ranking comparisons.

Inspired by recent progress in automatic evaluation of machine translation (BLEU; Papineni et al. 2001), we would like to apply the same idea in the evaluation of summaries. Following BLEU, we used the automatically computed accumulative n-gram matching scores (NAMS) between a model unit (MU) and a system summary (S)[4] as performance indicator, considering multi-document summaries. Only content words were used in forming n-grams. NAMS is defined as follows:

$$a_1\_NAM_1 + a_2\_NAM_2 + a_3\_NAM_3 + a_4\_NAM_4$$

NAM_n is n-gram hit ratio defined as:

$$\frac{\#\text{ of matched n-grams between MU and S}}{\text{total }\#\text{ of n-grams in MU}}$$

We tested three different configurations of a_i:

C1: $a_1 = 1$ and $a_2 = a_3 = a_4 = 0$;

C2: $a_1 = 1/3$, $a_2 = 2/3$, and $a_3 = a_4 = 0$;

C3: $a_1 = 1/6$, $a_2 = 2/6$, $a_3 = 3/6$, and $a_4 = 0$;

C1 is simply unigram matching. C2 and C3 give more credit to longer n-gram matches. To examine the effect of stemmers in helping the n-gram matching, we also tested all configurations with two different stemmers (Lovin's and Porter's). Figure 4 shows the results with and without using stemmers and their Spearman rank-order correlation coefficients (rho) compared against the original retention ranking from Figure 3. X-*n*G is configuration *n* without using any stemmer, L-*n*G with the Lovin stemmer, and P-*n*G with the Porter stemmer.

The results in Figure 4 indicate that unigram matching provides a good approximation, but the best correlation is achieved using C2 with the Porter stemmer. Using stemmers did improve correlation. Notice that rank inversion remains within the performance groups identified in Section 4. For example, the retention ranking of Baseline1, U, and W is 14, 16, and 15 respectively. The P-2G ranking of these three systems is 15, 14, and 16. The only system crossing performance groups is Y. Y should be grouped with N and T but the automatic evaluations place it lower, in the group with Baseline2, L, and P. The primary reason for Y's behavior may be that its summaries consist mainly of headlines, whose abbreviated style differs from the language models derived from normal newspaper text. We address these mismatch issues in Section 6.

In the next section, we propose an alternative view of automatic summary evaluation, one that allows us to estimate measurability.

5.1 Summary Evaluation as a Decision Making Process

Instead of calculating the NAMS scores directly from pairs of model and system units and using the results to assess system performance, we view evaluation of summaries as a decision making process. For a given model unit and a system summary, we want an automated evaluation system to learn how to assign a correct coverage score to a system-model unit pair automatically. We are interested in determining how well such a system can emulate the performance of human evaluators.

[4] The whole system summary was used to compute NAMS against a model unit.

SYSCODE	Single-Doc-Org	Single-Doc-Exp
Human1	0.54280	0.60827
Human2	0.44228	0.48276
Baseline1	0.76247	0.85551
O	0.71816	0.83974
P	0.71515	0.81878
Q	0.68194	0.79931
R	0.73706	0.83022
S	0.68084	0.81413
T	0.73274	0.83746
V	0.75088	0.86544
W	0.71104	0.84115
X	0.62760	0.78258
Y	0.64527	0.76962
Z	0.62242	0.78445

Figure 5. Krippendorff agreement coefficients on the original model unit set and the expanded set (single-document).

We used a version of the C4.5 rule learner developed by Witten and Frank (1999), together with eight features: the four n-gram matching scores NAM_n, n = 1, 2, 3, and 4, and four counts of the corresponding number of n-grams in the model unit (length). The assumption is if the C4.5 rule learner can learn to make correct coverage assignments by using NAM_n and length features alone, then the decisions (assigned coverage scores) should have a high degree of agreement with the reference human evaluation score decisions. We used Krippendorff's (1980) agreement coefficient α to assess the degree of agreement, which is defined as follows:

$$\alpha = 1 - \frac{D_o}{D_e}$$

D_o is the observed disagreement and D_e is the expected disagreement. Disagreement is measured on the ordinal scale. We performed the following experiments:

(1) Single document task:

a. Learn C4.5 rules to score system X using the DUC-2001 single-document test collection but leaving out system X's decisions;

b. Test the learned rules on system X's decisions: compute the coefficient between the decisions given by humans and by the learned rules;

c. Repeat the above for all systems.

SYSCODE	Multi-Doc-Org	Multi-Doc-Exp
Human1	0.21734	0.29003
Human2	0.12474	0.22316
Baseline1	0.32158	0.44581
Baseline2	0.38393	0.59318
L	0.37001	0.53253
M	0.33022	0.48626
N	0.33279	0.53872
O	0.37254	0.55113
P	0.35391	0.56048
R	0.31613	0.49922
S	0.35606	0.56148
T	0.44037	0.57502
U	0.22757	0.36744
W	0.20099	0.44844
Y	0.16930	0.40383
Z	0.32778	0.56396

Figure 6. Krippendorff agreement coefficients on the original model unit set and the expanded set (multi-document).

(2) Multi-document task:

a. Same as (1) but use DUC-2001 multi-document test collection.

We observed that for most MUs, several SUs were considered equivalent (scored 4). Therefore we can expand the ideal set by including all the 4-scoring SUs, and performing the n-gram matching against this expanded set. We performed an additional experiment:

(3) Repeat (1) and (2) on expanded model units.

Expanded model units were obtained using the following procedure:

(1) Initialize the expanded model unit, M_{xj}, with the original model unit M_j.

(2) For every model unit M_j,

a. for every system unit S_i of system X such that S_i has a human assigned coverage score of 4, add S_i to M_{xj}.

(3) For a set of system units S_i, compute its NAM_n against the expanded model unit by:

$$NAM_n = \max(NAM_n(S_i, S_j))$$
$$\text{where } S_i \neq S_j.$$

Following Krippendorff, we consider an agreement coefficient greater than 0.8 as reliable. The results shown in Figures 5 and 6 indicate that we cannot learn reliable C4.5 rules to assign coverage scores using just NAMS and length related features. Expanding model units helps, but only for some systems and only in the single document summarization task. Notice the comparably lower reliability of HUMAN1 and HUMAN2, which is probably largely caused by C4.5 learning from the more frequent extracts instead of the human summaries. The much lower reliability in the multi-document task indicates this is a more difficult task to learn.

6 Discussion

Several types of mismatches between human and system summaries make manual and automatic evaluations of summaries difficult.

6.1 Synonyms and Paraphrases

Since there are many ways to say the same thing, mismatches are bound to occur. Even if both model and system summaries are extracts, the system may choose different extracts to say the same thing, especially in the multi-document case. For example (D14):

- Model:
  - `Crashes occurred in widely separated areas, and varied in the degree of disaster.`
- Systems:
  - `F-16S CRASH IN MID-AI.`
  - `ANOTHER F-16 CRASHES IN BLACK FOREST.`
  - `MILITARY CARGO PLANE CRASH SITE YIELDS FEW CLUES. F-111 CRASHES IN SAUDI ARABIA, KILLING TWO.`
  - `B-52 BOMBER CRASHES IN MICHIGA. CREW SURVIVES.`
  - `U.S. PILOT PARACHUTES TO SAFETY AFTER MILITARY JET CRASHES IN JAPAN.`

6.2 Generation vs. Extraction

A summary is intended to generalize details. Few current automatic text summarization systems perform the necessary regeneration. Human abstraction and regeneration invariably highlights different aspects, and employs different phrasing and word usage. For example (D05-LA060490-0083):

- Model 1: `On May 11 it was reported that a Siamese cat had died from a form of BSE and the press went wild suggesting that if cats can get it people can too.`
- Model 2: `The British press has been making a big deal of the Ministry of Agriculture's report that a Siamese cat had died of a disease similar to "mad cow", saying that it proved that other species, including humans, were vulnerable.`

6.3 Different Styles

The style of system and reference summaries may differ; for example, normal newspaper text vs. headlines (D15):

- Model:
  - `The tuberculosis was fueled by AIDS patients who were vulnerable when their lowered immune system allowed the latent bacteria to develop into active tuberculosis.`
- System:
  - `INCREASE OF TUBERCULOSIS DUE TO AIDS VIRUS POSES NEW HEALTH THREAT.`
  - `STUDY RECOMMENDS TB TREATMENT FOR AIDS-INFECTED ADDICTS.`
  - `RESEARCHERS DECLARE SUCCESS IN PUTTING AIDS IN REMISSION.`

6.4 Different Perspectives

DUC-2001 did not require participants or human abstractors to generate summaries from specific perspectives. It is quite possible that summaries generated by systems are good, but made from different perspectives, and therefore fail to match reference summaries. Figure 7 shows such an example, where the reference summary focuses on the causes of the accidents while the system summary focuses on the accidents themselves.

Reference Summary
(Focus on the causes of accidents)
<D59.M.050.K.K>

1. A number of domestic and foreign planes crashed during the years 1987-1991.
2. Some of the suspected causes are mechanical failure, pilot error, extreme weather and terrorist activity.
3. Occasionally causes overlap, such as pilot error during a mechanical crisis, or navigational problems in bad weather.
4. Some crashes remain a mystery.

System Summary
(Focus on the accidents themselves)
<D59.M.050.K.T>

1. Many civil air crashes occurred in late 80s and early 90s.
2. Type of airplanes involved were: single-engine airplane Cessna P210, Boeing 737, DC-10 jetliner, Boeing 737-200, Boeing 747, Cessna 177, MD-82, DC-7.
3. Airlines involved were: Midland Airways, UTA, United Airlines, Pan American World Airways, Northwest Airlines, American Airlines.
4. Many people died and were injured.

Figure 7. Summaries from different perspectives.

7 Conclusions

We described manual and automatic evaluation of single and multi-document summarization in DUC-2001. We showed the instability of human evaluations and the need to consider this factor when comparing system performances. As we factored in the instability, systems tended to form separate performance groups. One should treat with caution any interpretation of performance figures that ignores this instability.

Automatic evaluation of summaries using accumulative n-gram matching scores seems promising. System rankings using NAMS and retention ranking had a Spearman rank-order correlated coefficient above 97%. Using stemmers improved the correlation. However, satisfactory correlation is still elusive.

Our initial attempt to gauge the reliability and learnability of automated summary evaluation by learning C4.5 rules to assign coverage scores to individual model units did not work out well. One factor might be that the features we used in training were too weak. However, the main problem we ascribe to the small size of the learning data. The core problem in automated summary evaluation is the large expressive range of English. More even than in machine translation, human summarizers create fresh text. No n-gram matching evaluation procedure can overcome the paraphrase/synonym problem unless (many) model summaries are available. Even enriching the test corpus by expanding model units, as described in Section 5, does not help if only a single model summary and some evaluations are provided.

We cannot estimate how many model summaries and judgments are required to achieve reliable automated summary evaluation. We did however in section 5 provide a method for quantifying the reproducibility of a set of judgments and estimating its reliability using Krippendorff's coefficient. This method may be useful in the future when more summary judgments are available.

References

DUC. 2001. The Document Understanding Conference 2001. http://www-nlpir.nist.gov/projects/duc/2001.html.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology.* Sage Publications.

Lin, C.-Y. 2001. *Summary Evaluation Environment.* http://www.isi.edu/~cyl/SEE.

Mani, I., D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. *The TIPSTER SUMMAC Text Summarization Evaluation: Final Report.* MITRE Corp. Tech. Report.

Witten, I.H. and E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann.