The pyramid method: incorporating human content selection variation in summarization evaluation

ANI NENKOVA, REBECCA PASSONNEAU and KATHLEEN MCKEOWN
Columbia University

Human variation in content selection in summarization has given rise to some fundamental research questions: How can one incorporate the observed variation in suitable evaluation measures? How can such measures reflect the fact that summaries conveying different content can be equally good and informative? In this paper we address these very questions by proposing a method for analysis of multiple human abstracts into semantic content units. Such analysis allows us not only to quantify human variation in content selection, but also to assign empirical importance weight to different content units. It serves as the basis for an evaluation method, the Pyramid Method, that incorporates the observed variation and is predictive of different equally informative summaries. We discuss the reliability of content unit annotation, the properties of Pyramid scores, and their correlation with other evaluation methods.

1. INTRODUCTION

The most common way to evaluate the informativeness of automatic summaries is to compare them with human-authored model summaries. For decades, the task of automatic summarization was cast as a sentence selection problem and systems were developed to identify the most important sentences in the input and those were selected to form a summary. It was thus appropriate to generate human models by asking people to produce summary extracts by selecting representative sentences. Systems were evaluated using metrics such as precision and recall [Salton et al. 1997], measuring to what degree automatic summarizers select the same sentences as a human would do. Over time, several related undesirable aspects of this approach have been revealed:

Authors' present addresses: Ani Nenkova, Stanford University Margaret Jacks Hall, Building 460 Stanford, CA 94305-2150; email: anenkova@stanford.edu

Rebecca Passonneau, Center for Computational Learning Systems, Columbia University, Mail Code 7717, 457 Riverside Drive, New York, NY 10115; email: becky@ccls.columbia.edu

Kathleen McKeown, Columbia University, Computer Science Department, 1214 Amsterdam Ave, New York, NY 10027; email: kathy@cs.columbia.edu

Human variation. Content selection is not a determininstic process [Salton et al. 1997; Marcu 1997; Mani 2001]. Different people choose different sentences to include in a summary, and even the same person can select different sentences at different times [Rath et al. 1961]. Such observations lead to concerns about the advisability of using a single human model and suggest that multiple human gold-standards would provide a better ground for comparison.

Analysis granularity. Even ignoring the issue of human variability, comparing the degree of sentence co-selection is not always justified. Even if a system does not choose exactly the same sentence as the human model, the sentence it does choose may overlap in content with one or more of the model sentences, conveying a subset of their content. This partial match should be accounted for, but requires analysis below the sentence level.

Semantic equivalance. An issue related to that of appropriate granularity for analysis is semantic equivalence. Particularly in news, or in multi-document summarization, different input sentences can express the same meaning even when they use different wording. Naturally, annotators would choose only one of the equivalent sentences for their summaries and a system will be penalized if it selects one of the other equally appropriate options.

Extracts or abstracts? When asked to write a summary of a text, people do not normally produce an extract of sentences from the original. Rather, they use their own wording and synthesis of the important information. Thus, exact match of system sentences with human model sentences, as required for recall and precision metrics, is not at all possible. As the field turns to the development of more advanced non-extractive summarizers, we will clearly need to move to a more sophisticated evaluation method which can handle semantic equivalence at varying levels of granularity.

The Pyramid Method, the description and analysis of which are the focus of this paper, provides a unified framework for addressing the issues outlined above. A key assumption of the Pyramid Method is the need for multiple human models which, taken together, yield a gold-standard for system output. The method features a procedure for manual identification of semantic equivalance in abstracts, allowing for variability in the granularity of the analysis. When applied to the human abstracts, it results in a representation that explicitly identifies commonly agreed upon content units. Such semantically motivated analysis allows for the definition of an intuitive evaluation metric that incorporates a differential importance weighting of information based on agreement across abstracts. Thus, the method can also be used to compare system output against the pyramid of human abstracts, yielding a score reflecting how much of the important content the system summary captured. Our analysis of the scoring method shows that despite the inherent difficulty of the task, it can be performed reliably.

In the remainder of this paper, we first define the analysis method, showing how it is used to create pyramids and score system output (Section 2). We then present our analysis confirming the need for multiple models in Section 3, turn to a discussion of the reliability of manual content analysis of automated summaries (Section 4), and before closing, discuss other evaluation approaches and compare them with the Pyramid Method approach (Section 5).

2. DEFINING PYRAMIDS

Quantitative evaluation of content selection of summarization systems requires a gold-standard against which automatically generated summaries can be compared; a pyramid is a representation of a gold-standard summary for an input set of documents. Because a pyramid is used for evaluating summary content (as opposed, for example, to wording), units of comparison within a pyramid correspond to units of meaning. We call these *Summary Content Units* (SCUs). Unlike many gold-standards, a pyramid represents the opinions of multiple human summary writers each of whom has written a model summary for the input set of documents. A key feature of a pyramid is that it quantitatively represents agreement among the human summaries: SCUs that appear in more of the human summaries are weighted more highly, allowing differentiation between important content (that appears in many human summaries) from less important content. Such weighting is necessary in summarization evaluation, given that different people choose somewhat different information when asked to write a summary for the same set of documents. In this section, we define SCUs, outline a procedure for identifying them, and present a method for scoring a new summary against a pyramid.

2.1 Summary content units

SCUs are semantically motivated, sub-sentential units; they are variable in length but not bigger than a sentential clause. This variability is intentional since the same information may be conveyed in a single word or a longer phrase. SCUs emerge from annotation of a collection of human summaries for the same input. They are identified by noting information that is repeated across summaries, whether the repetition is as small as a modifier of a noun phrase or as large as a clause. During the process, annotators label the SCUs in their own words, and can modify the label as they go along. Sentences corresponding to information that appears only in one summary are broken down into clauses, each of which is one SCU in the pyramid. Weights are associated with each SCU indicating the number of summaries in which it appeared. Rather than attempting to provide a formal semantic or functional characterization of what an SCU is, our annotation procedure defines how to compare summaries to locate the same or different SCUs. They are similar in spirit to the automatically identified elementary discourse units [Marcu 2000; Soricut and Marcu 2003], the manually marked information nuggets [Voorhees 2004] and factoids [van Halteren and Teufel 2003], all of which are discussed in greater length in Section 5.

Below is an example of the emergence of two SCUs from six human abstracts. The sentences are indexed by a letter and number combination, the letter showing which summary the sentence came from and the number indicating the position of the sentence within its respective summary.

A1. The industrial espionage case involving GM and VW began with <u>the hiring of Jose Ignacio Lopez, an employee of GM</u> subsidiary Adam Opel, <u>by VW</u> as a production director.

B3. However, <u>he left GM for VW</u> under circumstances, which along with ensuing events, were described by a German judge as "potentially the biggest-ever case of industrial espionage".

C6. He left GM for VW *in March, 1993.*

D6. The issue stems from the alleged recruitment of GM's eccentric and visionary Basque-born procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez's buisness colleagues.

E1. On March 16, 1993, with Japanese car import quotas to Europe expiring in two years, renowned cost-cutter, Agnacio Lopez De Arriortua, left his job as head of purchasing at General Motor's Opel, Germany, to become Volkswagen's Purchasing and Production director.

F3. In March 1993, Lopez and seven other GM executives moved to VW overnight.

The annotation starts with identifying similar sentences, like the six above, and then proceeds with finer grained inspection to identify more tightly related subparts. We obtain two SCUs from the underlined and italisized spans of words (called contributors) of the sentences above. It is evident that the contributors for the same content unit in different summaries can vary noticeably since the same meaning can be expressed using very different wording and various syntactic constructions. Contextual information from the entire summary is used to decide semantic equivalance of the contributors, such as resolving pronominal anaphora and filling in arguments inferrable from preceeding sentences (such as *VM being the recruiter* in sentence **D6**). Each SCU has a weight corresponding to the number of summaries it appears in; SCU1 has weight=6 and SCU2 has weight=3. In this manner, information that is included in more human summaries is awarded higher weight and importance. This decison assumes that the summary writers are equally capable, and good at the summarization task.

SCU1 (w=6): *Lopez left GM for VW*
A1. the hiring of Jose Ignacio Lopez, an employee of GM ... by VW

B3. he left GM for VW

C6. He left GM for VW

D6. recruitment of GM's ... Jose Ignacio Lopez

E1. Agnacio Lopez De Arriortua, left his job ... at General Motor's Opel ... to become Volkswagen's ... director

F3. Lopez ... GM ... moved to VW

SCU2 (w =3) *Lopez changes employers in March 1993*
C6. in March, 1993

E1. On March 16, 1993

F3. In March 1993

The remaining parts of the six sentences above end up as contributors to many SCUs of different weight and granularity.

As illustrated above, an SCU consists of a set of contributors that, in their sentential contexts, express the same semantic content. In addition, an SCU has a unique index, a weight, and a natural language label. The label, which is subject to revision throughout the annotation process, has three functions. First, it frees the annotation process from dependence on a semantic representation language. Second, it requires the annotator to be conscious of a specific *meaning* shared by all contributors. Third, because the contributors to an SCU are taken out of context, the label serves as a *reminder* of the full in-context meaning, as in the case of SCU2 above where the temporal PPs are about a specific event, the time of Lopez's recruitment by VW.
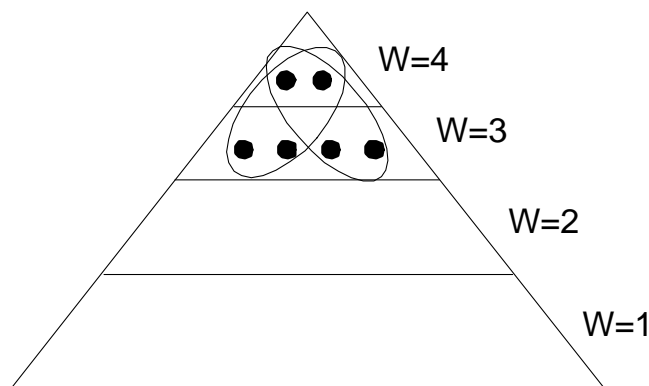
Fig. 1.   Two of six optimal summaries with 4 SCUs

2.2   Scoring a Summary

After the annotation procedure is completed, the final SCUs can be partitioned in a pyramid based on the weight of the SCU; each tier contains all and only the SCUs with the same weight. The number of annotated model summaries n determines the maximum possible number of tiers in the pyramid which we say is *a pyramid of size* n. The number of tiers in the pyramid can be different from its size in cases where there is no overlap between all of the models used for the pyramid creation. The name "pyramid" comes from the observed Zipffian distribution of SCU weights. There are few content units (at the top of the pyramid) that all people expressed in their summaries, and a very large number of content units expressed by only one of the summary writters (forming the base of the pyramid). In descending tiers, SCUs become less important informationally since they emerged from fewer summaries.

We use the term "pyramid of order n" to refer to a pyramid with n tiers. Given a pyramid of order n, we can predict the optimal summary content for a specified number of contributors—it should include all SCUs from the top tier, if length permits, SCUs from the next tier and so on. In short, in terms of maximizing information content value, an SCU from tier $(n-1)$ should not be expressed if all the SCUs in tier n have not been expressed. This characterization of optimal content ignores many complicating factors such as constraints for ordering SCUs in the summary. However, we explicitly aim at developing a metric for evaluating *content selection*, under the assumption that a separate *linguistic quality* evaluation of the summaries will be done as well. The proposed characterization of optimal content is predictive: among summaries produced by humans, many seem equally good without having identical content. Figure 1, with two SCUs in the uppermost tier and four in the next, illustrates two of six optimal summaries of size four (in SCUs) that this pyramid predicts.

Based on a pyramid, the informativeness of a new summary can be computed as the ratio of the sum of the weights of its SCUs to the weight of an optimal summary with the same number of SCUs. Such scores range from 0 to 1, with higher scores indicating that relatively more of the content is as highly weighted as possible.

We now present a precise formula to compute a score for a summary capturing the above intuitions about informativeness, which we term the *original pyramid score*. Suppose the pyramid has n tiers, T_i, with tier T_n on top and T_1 on the bottom. The weight of SCUs in tier T_i will be i. There are alternative ways to assign the weights and the method does not depend on the specific weights assigned: the weight assignment we adopted is simply the most natural and intuitive one. Let $|T_i|$ denote the number of SCUs in tier T_i. Let D_i be the number of SCUs in the summary that appear in T_i. SCUs in a summary that do not appear in the pyramid are assigned weight zero. The total SCU weight $\mathcal{D}$ is $\mathcal{D} = \sum_{i=1}^{n} i \times D_i$

The optimal content score for a summary with X SCUs is:

$$\text{Max} = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|), \text{ where } j = \max_i(\sum_{t=i}^{n} |T_t| \geq X)$$

In the equation above, j is equal to the index of the lowest tier an optimally informative summary will draw from. This tier is the first one top down such that the sum of its cardinality and the cardinalities of tiers above it is greater than or equal to X (summary size in SCUs). For example, if X is less than the cardinality of the most highly weighted tier, then $j = n$ and Max is simply $X \times n$ (the product of X and the highest weighting factor).

Then the pyramid score $\mathcal{P}$ is the ratio of D to Max. Because $\mathcal{P}$ compares the actual distribution of SCUs to an empirically determined weighting, it provides a direct comparison to the way people select information from source texts.

2.3 Other scores based on pyramid annotation

The original pyramid score defined in the previous section represents only one of the possible ways for incorporating the content unit analysis into a score reflecting the appropriateness of a content in a summary. The original pyramid score is similar to a precision metric—it reflects how many of the content units that were included in a summary under evaluation are as highly weighted as possible and it penalizes the use of a content unit when a more highly weighted one is available and not used.

Alternatively, we define a pyramid score corresponding to recall, which we term the *modified pyramid score*. This recall oriented score is defined as the weight of the content units in the summary normalized by the weight of an ideally informative summary of SCU size equal to the average SCU size the human summaries in the pyramid. So again, the score is the ratio between $\mathcal{D}$ (the sum of weights of SCUs expressed in the summary) and Max (the optimal score of a summary of size X,), but this time X is not the SCU length of the evaluated peer, but rather the average number of SCUs in the model summaries used for the creation of the pyramid, X_a.

This score measures if the summary under evaluation is as informative as one would expect given the SCU size of the human models. For example, in cases when a new summary expresses more content units than the average pyramid model, the modified pyramid score is not sensitive to the weight of this additionally packed information. Note that in theory the modified score can be greater than 1. In practice this never happens even when evaluating new human-written summaries because even they contain SCUs with lower weight before expressing all content weigther more highly. In the next section we will discuss the findings from the applica-

tion of the pyramid evaluation method in DUC 2005 where the modified pyramid score showed better qualities than the original pyramid scores—it proved to be less sensitive to peer annotation errors, it distinguished better between systems and had higher correlation with other manual evaluation metrics such as responsiveness judgments. The modified score requires less annotation effort, since the parts of a new summary that don't correspond to any SCU in the pyramid need not be broken down in individual SCUs. This step is necessary for the computation of the original pyramid score because the exact number of content units in the summary needs to be known for the computation of the weight of the ideal summary.

Another possibility for a score can ignore the weighting of content units altogether. The pyramid annotation can be used simply to obtain a pool of content units that are likely to appear in the summary, similarly to the way *nuggets* are used for evaluation of question-answering systems [Voorhees 2004]. In this scenario, the standard precision and recall used in information retrieval can be computed. Earlier we defined D_i as the number of SCUs in a summary under evaluation that appear in tier T_i of the pyramid. In particular, D_0 is the number of SCUs in the peer that do not appear in the pyramid. Then we can straightforwardly define

$$Recall = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} T_i} \tag{1}$$

and

$$Precision = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=0}^{n} D_i} \tag{2}$$

Recall is equal to the fraction of content units in the peer summary that are also in the pyramid and precision is the ratio of SCUs from the pyramid that are expressed in the peer to all SCUs expressed in the peer. Such scores would not incorporate all the knowledge derivable from SCU analysis of multiple summaries—the benefit from the use of multiple models will be only that a bigger pool of potential content units can be collected. But the importance weighting will not be used. A notable difference of the precision/recall approach proposed here and that used in evaluation of question answering systems is that the pyramid method is based on an analysis of *human models*, while the information nuggets in question-answering evaluation are obtained by analyzing (mostly) the output of automatic systems, thus making it impossible to claim that an occurrence in the answer provides empirical evidence for the importance of the nugget.

Another interesting possibility is to use each SCU as the unit of evaluation. Such an approach is similar to the one used in machine translation, where human judgments for quality are collected on a sentence by sentence basis, rather than for a complete text. This kind of score can be used when the goal is to compare systems and will work in the following way. Say there are N content units in a pyramid, $N = \sum_{i=1}^{n} T_i$. Each peer summary will be associated with a binary vector S and $S[k] = 1$ if the kth content unit from the pyramid is expressed in the peer (and is 0 otherwise). Thus, when comparing two summaries from different systems, for each test set we will get a vector of observations, rather than a single number as the

original or modified pyramid scores do. This means that one can apply a paired statistical test and test if *two different summaries for the same set* are *statistically* significantly different. It is not possible to make conclusions of this sort from the original pyramid scores because a vector of observations is necessary to compute the variance of the scores. Similarly, when comparing the performance of two systems across several test sets, the content unit based evaluation would give more data points which can be used to compare the systems. Say there are Y test sets. If the per summary scores are used, the basis for comparison between the two systems will consist of a vector of Y observations. But there will be about $Y * N_a$ data points for content unit based evaluation, where N_a is the expected number of SCUs in a pyramid. This large gain in data can make it possible to reach statistically significant conclusions even when few test sets are available.

3. THE NEED FOR MULTIPLE MODELS IN SUMMARIZATION EVALUATION

It is well known that different people choose different content for inclusion in their summaries and thus a summary under evaluation could receive a rather different score depending on which summary is chosen to be the model. In fact, in previous work, McKeown *et al.* [McKeown et al. 2001] showed that in evaluations based on a single model, the choice of the model had a significant impact on the scores assigned to summaries. If an evaluation uses too few models, the resulting ranking of systems is necessarily suspect: would the ranking have been different if different model summaries were used? In this section, we present a study of the effect of the size of the pyramid on summary scores. The two specific questions we examine are:

(1) *How does variability of scores change as pyramid size increases?*
(2) *At what size pyramid do scores become reliable?*

The data we use to address these questions is 10 100-word summaries for three test sets consisting of about 10 articles each. Empirically, we observed that as more human summaries are added in the pyramid, and the range between higher weight and lower weight SCUs grows larger, scores for held-out summaries for pyramids of growing size change less. This makes sense in light of the fact that a score is dominated by the higher weight SCUs that appear in a summary. However, we wanted to study more precisely at what point scores become independent of the choice of models that populate the pyramid. We conducted three experiments to locate the point at which scores stabilize across our three datasets. Each experiment supports the conclusion that about five summaries are needed.

Specifically, we used three DUC 2003 summary sets for which four human summaries were written. In order to provide as broad a comparison as possible for the least annotation effort, we selected the set that received the highest DUC scores (set D30042), and the two that received the lowest (sets D31041, D31050). For each set, we collected six new summaries from advanced undergraduate and graduate students with evidence of superior verbal skills; we gave them the same instructions used by NIST to produce model summaries.

Our first step in investigating score variability was to examine all pairs of summaries where the difference in scores for a size 9 pyramid was greater than 0.1; there were 68 such pairs out of 135 total. All such pairs exhibit the same pattern

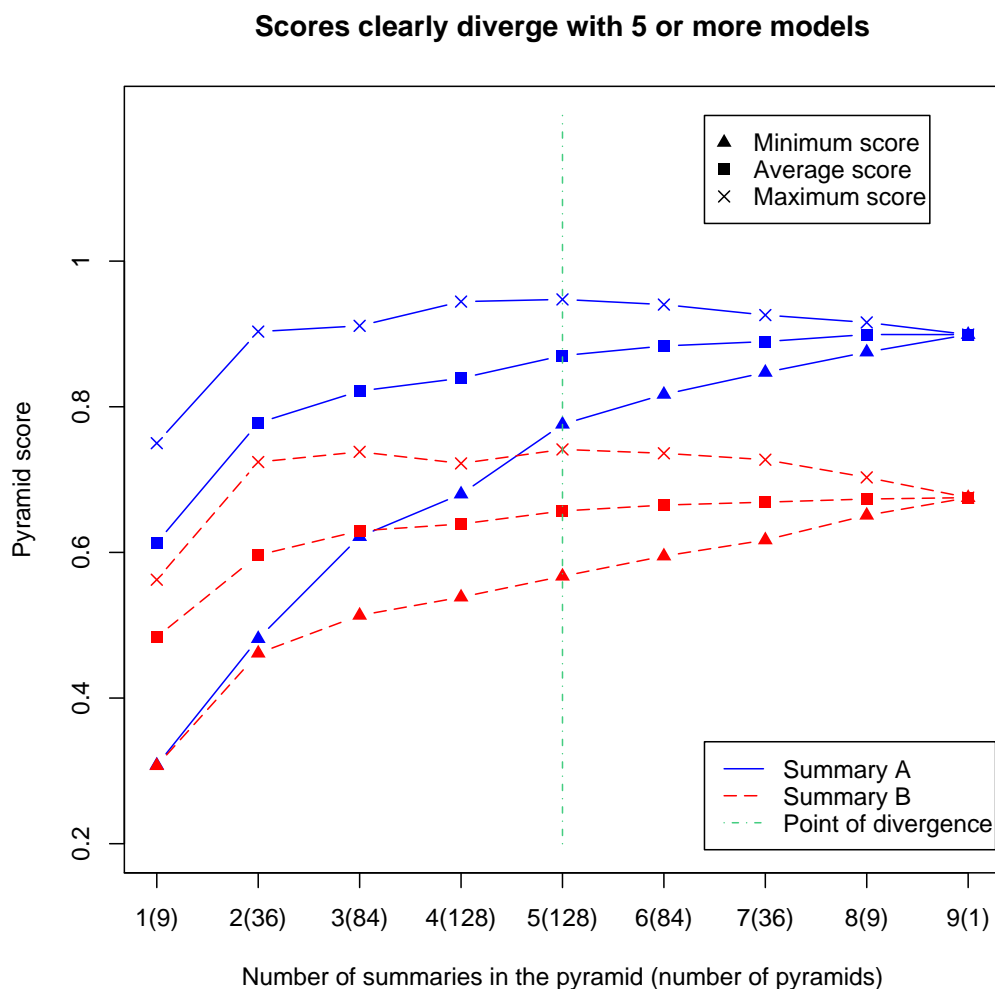## Scores clearly diverge with 5 or more models

Fig. 2. Minimum, maximum and average scores for two summaries for pyramids of different size. Summary A is better than summary B as can be seen from the scores for pyramids of size 9, but with few models in the pyramid, it can be assigned lower scores than those for summary B.

illustrated in Figure 2 for two summaries we call 'Summary A' (shown with solid lines) and 'Summary B' (shown with dotted lines). The x-axis on the plot shows how many summaries were used in the pyramid (and in brackets, the number of pyramids of that size that could be constructed with the nine available model summaries) and the y-axis shows the minimum (marked on the plot by a triangle), maximum (marked by a cross) and average (marked by a square) scores for each

of the summaries for a given size of pyramid.[1] Of the two, A has the higher score for the size 9 pyramid, and is perceivably more informative. Averaging over all size-1 pyramids, the score of summary A is higher than that for B (with all sizes of pyramids, including that for size-1, the square representing the average score for summary A across all possible pyramids is above the square that represents the average score for summary B). But some individual size-1 pyramids might yield a higher score for summary B: the minimum score assigned by some pyramid to summary A (triangle) is lower than the average score for the worse summary B.

The score variability at size-1 is huge: it can be as high as 0.5, with scores for summary A varying between around 0.3 to close to 0.8. With pyramids of bigger size, scores stabilize: the difference between the maximum and minimum score each summary could be assigned diminishes, and even the lowest score assigned to the better summary (A) is higher than any score for the worse summary (B). Specifically, in our data, if summaries *diverge* at some point as in Figure 2, meaning that the minimum score for the better summary is higher than the maximum score for the worse summary, the size of the divergence never decreases as pyramid size increases. This is visually expressed in the figure by the growing distance between the triangles and the crosses. The vertical dotted line at pyramids of size 5 marks the first occurrence of divergence in the graph. For pyramids of size > 4, summaries A and B never receive scores that would reverse their ranking, regardless of which model summaries are used in the pyramids.

For all pairs of divergent summaries, the relationship of scores follows the same pattern we see in Figure 2. The point of divergence where the scores for one summary become consistently higher than those of the other was found to be stable. In all instances, if summary A gets higher scores than summary B for all pyramids of size n, then A gets higher scores for pyramids of size $\geq n$. We analyzed the score distributions for all 67 pairs of summaries the size-9 scores for which differed by more than 0.1, in order to determine what size of pyramid is required to reliably determine which is the better one. The expected value for the point of divergence of scores, in terms of number of summaries in the pyramid, is 5.5.

We take the scores assigned at size 9 pyramids as being a reliable metric on the assumption that the pattern we have observed in our data is a general one, namely that variance always decreases with increasing sizes of pyramid. When testing all combinations of pyramids with four or five models, if the observed range of scores for one summary is lower than the score range for another summary, it will remain lower for all pyramids with a larger number of models. We postulate that summaries whose scores differ by less than 0.06 have roughly the same informativeness. The assumption is supported by two facts. First, this corresponds to the difference in scores for the same summary when the pyramid annotation has been performed by two independent annotators (see [Nenkova and Passonneau 2004] for details). In later studies in the context of DUC 2005, it was also shown that scores based on peer annotations produced by novice annotators given the same pyramid also differ on average by 0.06 [Passonneau et al. 2005]. Second, the pairs of summaries whose scores never clearly diverged, had scores differing by less than 0.06 at pyramid size 9. So we assume that differences in scores by less than 0.06 do not translate

---

[1]Note that we connected data points with lines to make the graph more readable.

to meaningful differences in information quality and proceed to examine how the relative difference between two summaries at size-9 pyramids could change if we used pyramids of smaller size instead.

Now, for each pair of summaries $(sum1, sum2)$, we can say whether they are roughly the same when evaluated against a pyramid of size n and we will denote this as $|sum1| ==_n |sum2|$, (scores differ by less than 0.06 for some pyramid of size n) or different (scores differ by more than 0.06 for all pyramids of size n) and we will use the notation $|sum1| <_n |sum2|$ if the score for $sum2$ is higher.

When pyramids of smaller size are used, the following errors can occur, with the associated probabilities:

E_1∴ $|sum1| ==_9 |sum2|$ but $|sum1| <_n |sum2|$ or $|sum1| >_n |sum2|$ at some smaller size n pyramid. The conditional probability of this type of error is $p_1 = P(|sum1| >_n |sum2|||sum1| ==_9 |sum2|)$. In this type of error, summaries that are essentially the same in terms of informativeness will be falsely deemed different if a pyramid of smaller size is used.

E_2∴ $|sum1| <_9 |sum2|$ but with a pyramid of smaller size $|sum1| ==_n |sum2|$. This error corresponds to "losing ability to discern", and a pyramid with fewer models will not manifest a difference that can be detected if nine models were used. Here, $p_2 = P(|sum1| ==_n |sum2|||sum1| <_9 |sum2|)$.

E_3∴ $|sum1| <_9 |sum2|$ but for a smaller size pyramid $|sum1| >_n |sum2|$ Here, $p_3 = P(|sum1| >_n |sum2|||sum1| <_9 |sum2|) + P(|sum1| <_n |sum2|||sum1| >_n |sum2|)$. This is the most severe kind of mistake and ideally it should never happen, with the better summary getting a much lower score than the worse one. Note that such error can happen only for gold-standards with size smaller than their point of divergence.

Empirical estimates for the probabilities p_1, p_2 and p_3 can be computed directly by counting how many times the particular error occurs for all possible pyramids of size n. By taking each pyramid that does not contain either of $sum1$ or $sum2$ and comparing the scores they are assigned, the probabilities in Table 3 are obtained. We computed probabilities for pairs of summaries for the same set, then summed the counts for error occurrence across sets. The size of the pyramid is shown in the first column of the table, labeled n. The last column of the table, "Data points", shows how many pyramids of a given size were examined when computing the probabilities. The total probability of error $p = p1 * P(|sum1| ==_9 |sum2|) + (p2 + p3) * (1 - P(|sum1| ==_9 |sum2|))$ is also shown in Table 3.

Table 3 shows that for size-4 pyramids, the errors of type E_3 are ruled out. At size-5 pyramids, the total probability of error drops to 0.1 and is mainly due to error E_2, which is the mildest one.

Choosing a desirable size of pyramid involves balancing the two desiderata of having less data to annotate and score stability. Our data suggest that for this corpus, 4 or 5 summaries provide an optimal balance of annotation effort with score stability. This is reconfirmed by our following analysis of ranking stability.

In order to study the issue of how the pyramid scores behave when more than two summarizers are compared, for each set we randomly selected five peer summaries and constructed pyramids consisting of all possible subsets of the remaining five.

n	p1	p2	p3	p	data points
1	0.41	0.23	0.08	0.35	1080
2	0.27	0.23	0.03	0.26	3780
3	0.16	0.19	0.01	0.18	7560
4	0.09	0.17	0.00	0.14	9550
5	0.05	0.14	0.00	0.10	7560
6	0.02	0.10	0.00	0.06	3780
7	0.01	0.06	0.00	0.04	1080
8	0.00	0.01	0.00	0.01	135

Table I. Probabilities of errors E1, E2, E3 (p_1, p_2 and p_3 respectively), and total probability of error (p). The first column shows the pyramid size and the last column gives the number of observations used to compute the probabilities.

We computed the Spearman rank-correlation coefficient for the ranking of the five peer summaries compared to the ranking of the same summaries given by the size-9 pyramid. Spearman coefficient r_s [Dixon and Massey 1969] ranges from -1 to 1, and the sign of the coefficient shows whether the two rankings are correlated negatively or positively and its absolute value shows the strength of the correlation. The statistic r_s can be used to test the hypothesis that the two ways to assign scores leading to the respective rankings are independent. The null hypothesis can be rejected with a one-sided test with level of significance $\alpha = 0.05$, given our sample size $N = 5$, if $r_s \geq 0.85$.

Since there are multiple pyramids of size $n \leq 5$, we computed the average ranking coefficient, as shown in Table II. Again we can see that in order to have a ranking of the summaries that is reasonably close to the rankings produced by a pyramid of size $n = 9$, 4 or more summaries should be used.

n	average r_s	# pyramids
1	0.41	15
2	0.65	30
3	0.77	30
4	0.87	15
5	1.00	3

Table II.   Spearman correlation coefficient average for pyramids of order $n \leq 5$

4. APPLICATION OF THE METHOD IN DUC 2005

In the 2005 Document Understanding Conference, special attention was devoted to the study of evaluation metrics. The pyramid semantic-centered evaluation was one of the metrics used. 20 test sets were evaluated with the pyramid method, in addition to the linguistic quality evaluation, responsiveness and the automatic ROUGE metrics. The task in the conference was to produce a 250-word summary in response to a topic such as this shown below:

```
Explain the industrial espionage case involving VW and GM. Identify
the issues, charges, people, and government involvement. Report the
progress and resolution of the case. Include any other relevant
```

factors or effects of the case on the industry.

Pyramid evaluation was applied to 27 peers for each of the 20 test sets. The 20 pyramids were constructed by a team at Columbia University, and the peer annotation was performed by DUC participants and additional volunteers who were interested in the pyramid annotation. There were a total of 26 peer annotators and all 27 summaries in some sets were annotated by two different annotators allowing for a study of annotation reliability.

4.1 Peer Annotation Reliability

Pyramid scores rely on two kinds of human annotation: creation of the pyramids, and annotation of peer summaries against pyramids. Use of human annotations requires an assessment of their reliability. The reliability of pyramid construction is discussed in [Passonneau 2006]. Here, we analyze the reliability of the DUC 2005 peer annotations, which were performed by untrained annotators under time pressure. For six of the twenty document sets, we have annotations made by two different annotators (12 total annotators).

The purpose of a reliability analysis is to determine whether a change in annotator would have a significant impact on results. This is typically assessed by measuring inter-annotator agreement. In general, we believe it is important to combine measures of agreement with an independent assessment of the impact of substituting different annotations of the same data on the end results [Passonneau 2006], such as scoring peer systems. We use interannotator agreement to see if two annotators find largely the same SCUs in a peer, meaning the same content. We also measure the correlation of scores from different peer annotations to determine whether two annotations yield equivalent scores. Note that if two annotators find different SCUs in the same peer, the overall sum can be the same; two SCUs of weight two yield the same sum as one SCU of weight four. Further, a pyramid might contain two SCUs that are similar in content but have distinct weights, thus in principle, two annotators could find roughly the same content in a peer, but the sum of the SCU weights could be quite distinct. We find high interannotator agreement on peer annotations, which indicates that different annotators find largely the same SCUs in the peer summaries. We also find high correlations between the scores resulting from the different annnotations. Finally, the absolute differences in score values, on average, are generally smaller than the .06 difference we identified earlier as the threshold for a meaningful difference (§3).

4.1.1 *Data Representation.* Typically, an annotation task begins with a set of items to be annotated, and the annotation categories to use. Comparing annotations involves a one-to-one comparison between each decision made by annotators for each item. The peer annotation task does not fall directly into this model because the content units in a summary are not given in advance. Annotators are instructed to make their own selections of word sequences in the peer that express the same information as expressed in some SCU of the pyramid, and different annotators find different cardinalities of SCUs per peer. However, the pyramid is given in advance, and in order to correctly perform the task, an annotator must review every SCU in the pyramid at least once. Thus, for each document set, we take the annotator to make j decisions, where j is the number of SCUs in the pyramid.

We follow Krippendorff [1980] in representing agreement data in an $i \times j$ matrix for each decision made by each of the i coders (rows) on the j items (columns); so each annotation matrix has two rows, and a number of columns given by the number of SCUs per pyramid, which ranges from 88 to 171 for the 6 document sets considered here. To calculate interannotator agreement, we use Krippendorff's α.

The values in each cell (i, j) of the agreement matrix indicate how often annotator i finds SCU j in the peer. Annotators often find a peer summary to contain multiple instances of a given SCU. This will occur when a peer summary contains repetitive information, which occurs rather often in automatic multi-document summarization. There were a total of 359 cases of SCUs that were multiply-selected. Cell values for these cases indicate the number of times an SCU was matched to a peer, which ranged from zero to seven. We will illustrate below how we extend α to scale the comparison of choices made by annotators when they agree on an SCU, but not on how often it appears.

Our data representation contrasts with the method adopted in [Teufel and van Halteren 2004] for a similar annotation task, involving factoid annotation of 100-word summaries where the number of factoids is approximately the same order of magnitude (e.g., 153 factoids) as in pyramids. There, the authors assume that annotators consider each factoid for each sentence of the summary; for one twenty-sentence summary, they should arrive at 3,630 coding units.[2] In contrast, if a pyramid has 153 SCUs, our data representation of peer annotation represents 153 decisions per annotator. The difference in choice of data representation accounts for some of the difference in the agreement we find (see Table III) versus the figures of 0.86 and 0.87 reported in [Teufel and van Halteren 2004].

4.1.2 *Multiply Annotated Units.* In [Passonneau et al. 2005], we discussed in detail the motivation for using Krippendorff's α so as to give partial credit when annotators agree that a peer expresses a given SCU, but differ as to how often. Here, we briefly summarize the argument.

In the formula for α shown below, where j is the number of coders and i is the number of units, the numerator is a summation over the product of counts of all pairs of values b and c, times a weight or distance metric δ, within rows. The denominator is a summation of comparisons of paired values within columns.

$$1 - \frac{ij - 1 \sum_k \sum_b \sum_{b>c} n_{b_k} n_{c_k} \delta_{bc}}{j \sum_b \sum_c n_b n_c \delta_{bc}} \tag{3}$$

The choice of distance metric δ depends on the scale of values that is used in the annotation; because α measures disagreements, δ is set to zero when annotators agree. For nominal (categorical) data, when any pair of values are compared, say s and t, δ_{st} is 0 if they are the same (no disagreement) and 1 otherwise (disagreement). Applied here, the result is that if annotator A finds an SCU three times in a given peer, and annotator B finds the same SCU twice, they are said to disagree completely (δ=1). In order to quantify the cases of partial agreement on SCUs, we will report agreement using a δ based on the Dice coefficient [L.R.Dice 1945], a ratio for comparing the size of two sets.

---

[2]Their paper reads, *N=153 factoids times 20 sentences = 2920*, clearly an unintended mistake.

Annotators	Setid	α_{Dice}
102,218	324	0.67
108,120	400	0.53
109,122	407	0.49
112,126	426	0.63
116,124	633	0.68
121,125	695	0.61

Table III.   Interannotator agreement on peer annotation.

Let a represent the cases where two annotators A and B agree that an SCU appears in the current peer, b the cases where A finds SCUs that B does not, and c the cases where B finds SCUs that A does not. Dice is then: $\frac{(2a)}{(2a+b+c)}$. Where A finds two instances of an SCU and B finds three, Dice equals .8. Because α measures disagreements, δ is (1-Dice), which in this case is = .2. (1-Dice) increases as the disparity in SCU counts grows larger.

Table III summarizes the interannotator agreement for the six pairs of doubly annotated peers. The rows represent pairs of annotators on a particular document set and pyramid. The rightmost column represents average agreement across the 27 peers in each set. Analysis of variance of the full data set, with each agreement measured in turn as the dependent variable, and annotator pair, set, and peer as factors, shows no significant difference in variance on agreement, thus it is reasonable here to report the average agreement as representative of the individually computed agreement measures.

Values for α_{Dice} range from 0.49, about halfway between chance and perfect agreement, to 0.68 for sets 324 and 633, indicating that annotators agree rather well, especially considering the large number of degrees of freedom in their decisions. They had between 88 and 171 SCUs to select from, and each summary had 250 words. Annotators were free to select any sequence of words within a sentence as expressing an SCU, and could reselect words to match a different SCU, as long as the total selection was not a duplicate. Regarding the difference in performance across document sets, it is possible that at least one of the annotators who did sets 400 and 407 (the pair with lowest agreemnt) was less careful, or that these sets were more difficult to annotate.

4.1.3 *Interpretation of Results and Score Correlation.* We now look at the correlation of summary scores resulting from the two different annotations, followed by an examination of the size of the difference in score values. Investigators rarely report reliability results paired with independent results indicating at what level of reliability the annotations become useful. Instead, they often rely on an *a priori* threshold, such as the 0.67 value offered by Krippendorff [1980]. In [Passonneau 2006], we introduced the term *paradigmatic reliability study* to refer to the type of reliability study exemplified here, where interannotator agreement is reported along with an independent assessment of how much the results based on the annotated data would vary, given a different annotator.

Table IV shows Pearson's correlations on scores from the two different annotations for each set, and the confidence intervals, for both types of pyramid score. The

Annot.	Set id	original scores		modified scores	
		Cor.	Conf. Int.	Cor.	Conf. Int.
102,218	324	.76	(.54,.89)	.83	(.66, .92)
108,120	400	.84	(.67,.92)	.89	(.77, .95)
109,122	407	.92	(.83,.96)	.91	(.80, .96)
112,126	426	.90	(.78,.95)	.95	(.90, .98)
116,124	633	.81	(.62,.91)	.78	(.57, .90)
121,125	695	.91	(.81,.96)	.92	(.83, .96)

Table IV. Pearson's correlations for original and modified scores of the paired annotations. P-value = 0 for all correlations.

Set	Original score	Modified score
324	0.0713	0.1048
400	0.0062	0.0401
407	0.0413	0.0401
426	0.0142	0.0238
633	0.0289	0.0200
695	0.0506	0.0357

Table V. Average difference between the original and modified pyramid scores from two independent annotations.

correlations are high, and the differences between the correlations for the original and modified score are relatively small. For the four sets 400, 407, 426 and 695 the correlations are relatively higher (≥ 0.90), especially on the modified scores.

In addition to the correlations, we examined absolute differences in average scores. The average difference in scores across the 162 pairs of doubly annotated peers was 0.0617 for the original score and 0.0555 for the modified score. These numbers are very close to the empirically estimated difference of scores that we postulated in Section 2. Table V shows the average paired difference for each set for the original (in column 2) and modified scores (in column 3) respectively. The average differences in the six sets are overall in the expected range, smaller than 0.06. The only exception is *set 324*, where the scores differed on average by 0.1 for the modified score and 0.07 for the original pyramid scores. One of the annotators for this set reported that he was pressed for time and did not use the script provided to annotators to ensure consistency, so it comes as no surprise that the set this annotator was involved in exhibits the largest differences across annotator scores.

The smaller differences for the modified score compared to the original score is consistent with the fact that many annotators reported that they were unsure how to annotate content in the peer that is not in the pyramid. For the modified score, which does not make use of the annotation of content that does not appear in the pyramid, the differences are more systematic, indicated by the lower p-values for each set. Annotator training, or a protocol for doublechecking the annotations, possibly by another annotator, are likely to further reduce the observed differences.

4.1.4   *Discussion.* For the peer annotation data, the use of α_{Dice} is more intuitive, given the characteristics of the data, and more accurately quantifies the amount of agreement, than an unweighted metric. This makes it possible to place

pairs of values from different annotators on a scale, as opposed to a binary contrast between agreement and disagreement.

Here we have provided multiple assessments of the peer annotations, noting that scores from different annotations correlate very highly, with values generally above 0.80, and often above 0.90. The highly significant results on correlations of scores provide an independent assessment of the reliability of the peer annotations, as does the size of the score differences, which tend to be well below the sensitivity threshold of the pyramid metric. By combining three different types of evidence, we provide a comprehensive assessment of the reliability of peer annotation.

4.2   Correlations with other evaluation metrics

In this section, we will overview the correlations between the manual and automatic metrics used in DUC. The study of correlations is important in order to identify which metrics are mutually redundant or substitutable. For example, if two metrics A and B have correlation exceeding 0.95, and if we know the scores for one metric, say A, then we can predict the scores for the other (B) with very high accuracy. If the scores for metric B are more difficult to obtain than those for metric A (e.g., they require more annotation, more human subjects, etc.) then we can say that the metrics are mutually substitutable and simply use metric A in place of metric B. This situation usually arises when one of the metrics is automatic (easier to produce) and the other is manual (more difficult and expensive to produce). In the case when scores for both metrics with high correlation above 0.95 are equally easy/difficult to produce, it is advisable to chose and report only one of them, since the other does not bring in any new information into the analyses. Likewise, if two metrics are not perfectly correlated, they give information on some orthogonal qualities of the summaries and can be used jointly for overall assessment.

Table VI shows the correlations between pyramid scores and the other official metrics from DUC 2005—responsiveness and bigram overlap (ROUGE-2) and skip bigram (ROUGE-SU4). The responsiveness judgments were solicited by two NIST annotators (responsiveness-1 and responsiveness-2), who ranked all summaries for the same input on a scale from 1 to 5. The two ROUGE metrics are automatically computed by comparing a summary to a pool of human models on the basis of n-gram overlap. The numbers are computed using only the 25 automatic peers as this gives a more fair and realistic analysis of correlations. When the humans are included, all correlations exceed 0.92. This is related to the fact that in all metrics the human KM corrected misspelling performance is much better and consequently become outliers among the scores for each set and inflate the correlation.

All correlations are rather high and significantly different from zero. The two variants of the pyramid scores (original and modified) are very highly correlated, with Pearson's correlation coefficient of 0.96; so are the two automatic metrics as well ($\rho = 0.98$), indicating that the two pairs of metrics are mutually redundant. At the same time, the two sets of responsiveness judgments, given by two different judges under the same guidelines, have a correlations of only 0.83, confirming that the metric is subjective and different scores are likely to be assigned by different humans. The correlation between responsiveness-2 and the modified pyramid score is as high as 0.9 but still the metrics are not mutually redundant and each reveals information about the summary quality that is not captured by the other. The

	Pyr (mod)	Respons-1	Respons-2	ROUGE-2	ROUGE-SU4
Pyr (orig)	0.96	0.77	0.86	0.84	0.80
Pyr (mod)		0.81	0.90	0.90	0.86
Respons-1			0.83	0.92	0.92
Respons-2				0.88	0.87
ROUGE-2					0.98

Table VI. Pearson's correlation between the different evaluation metrics used in DUC 2005. Computed for 25 automatic peers over 20 test sets.

automatic metrics correlate quite well with the manual metrics, with Pearson's correlation in the range 0.80 to 0.92 but still do not seem to be high enough to suggest that the automatic metrics can be used to replace manual metrics. The findings are comparable with results from previous years on multi-document test sets [Lin 2004]. In previous years, a manual evaluation protocol based on a comparison between a *single* model and a peer was used. In his studies, Lin compared the correlations between these manual scores and several versions of automatic scores. Very good results were achieved for single document summarization and for very short summaries of 10 words where the correlation between the automatic and manual metrics was 0.99 and 0.98 respectively. But for 100-word multi-document summaries, the best correlation between an automatic metric and the manual metric was 0.81: the correlations for multi-document summarization are not as high as the ones achieved in automatic evaluation metrics for machine translations and for other summarization tasks, where Pearson's correlations between manual and automatic scores was close to perfect 0.99 [Papineni et al. 2002].

5. RELATED WORK

Summarization evaluation has been seen as a research topic in its own right for quite some time, with the difficulties stemming from the fact that there are multiple good summaries for the same input document(s). Many researchers have identified problems that arise as a consequence [Rath et al. 1961; Minel et al. 1997; Jing et al. 1998; Goldstein et al. 1999; Donaway et al. 2000]. Perhaps because of these early acknowledgements of the difficulties, there have been many recent efforts to develop evaluation methodology that is accurate, easy to use and can be applied on a wide scale. In this section, we discuss the annual summarization evaluation run by NIST, as well as other manual evaluation methods.

5.1 NIST-run Summarization Evaluation

The largest of recent efforts on evaluation has been developed within the Document Understanding Conference (DUC) series, which began in KM "or" to "on" 2001 and in which each year a large number of participants test their systems on a common test set. The DUC approach (until 2004) for evaluating summary content involves the comparison of *peer* summaries (i.e., summaries generated by automatic systems) against a single human-authored model. Each year, NIST collected multipled models, one of which was used for comparison while the other human models were scored against the first. To do the scoring, the human model was automatically broken down into *elementary discourse units* (EDUs), capturing the need for

analysis on a level smaller than a sentence. Software developed at ISI [Soricut and Marcu 2003] was used for this task and since it was done automatically, the granularity of each EDU varied from as short as a noun phrase to as long a complex sentence with multiple clauses. For each EDU in the model, the human evaluator had to decide on a 1 to 5 scale the degree to which the peer expresses its information. In addition, for sentences in the peer that did not express any model EDU, the evaluators assigned a score reflecting whether the sentence contained important information. Different proposals were made on how to incorporate the model EDU judgments into a final score, and average model EDU per summary was eventually adopted as a metric that was used throughout DUC 2004.

Two main drawbacks of the DUC approach were the use of a single model and the granularity of EDUs. Post evaluation analysis by McKeown *et al.* [McKeown et al. 2001] indicated that *the model* had larger impact on peer scores than *which summarizer* performed the task. In addition, Marcu [Marcu 2001] reported that some systems were overly penalized since they contained content ranked as highly relevant for the topic, but not included in the model summary, again pointing out a shortcoming of the use of a single model. The second drawback of the evaluation was the granularity of automatically identified EDUs. NIST evaluators reported having trouble deciding when an EDU can be said to match content in the peer and were also unsure how to use context in order to interpret the meaning of EDUs. Our work on pyramid evaluation aims at addressing these problems, and we are grateful to the DUC organizers and participants for giving us the opportunity to analyze some of the problems and look for solutions.

5.2   Other Manual Methods

There have been several other approaches to manual evaluation that address the problem of matching semantically similar units of information in a system summary against a model summary. These include a method for scoring sentences ranked for relevance, the use of nuggets as part of the TREC evaluation of definition questions, and the development of factoids and analysis of their impact on evaluation methodology.

Relative utility [Radev et al. 2000; Radev and Tam 2003] was one of the possible evaluation approaches listed in the "Evaluation Road Map for Summarization Research"[3], prepared in the beginning of the Document Understanding Conferences. In this method, all sentences in the input are ranked on a scale from 0 to 10 as to their suitability for inclusion in a summary. In addition, sentences that contain similar information are explicitly marked, so that in the evaluation metric one could penalize for redundancy and reward equally informationally equivalent sentences. The ranking of sentences from the entire input allows for a lot of flexibility, because summaries of any size or compression rate can be evaluated. At the same time, the method is applicable only to extractive systems that select sentences directly from the input and do not attempt any reformulation or regeneration of the original journalist-written sentence. The relative utility approach is very similar in spirit to the evaluation used by Marcu [Marcu 2000], Chapter 9), who asked multiple independent subjects to rank the importance of information units following older

---

[3]`www-nlpir.nist.gov/projects/duc/papers/summarization.roadmap.doc`

research strategies [Johnson 1970; Garner 1982]. The main difference is that earlier research directly concentrated on subsentential units rather than sentences.

Information nuggets have served for evaluation of question answering systems on non-factoid questions, which require a longer answer, very similar to summarization. Information nuggets are identified by human annotators through the analysis of *all systems' responses* to the question, as well as the searches made by the person who designed the question. They are atomic pieces of interesting information about the target, each of which is marked as vital (i.e., required) or non-vital (i.e., acceptable but not required) [Voorhees 2004]. In theory, the requirement that information nuggets be atomic distinguishes nuggets from our SCUs. SCUs vary in granularity—usually highly-weighted SCUs are characterized by shorter contributors and more "atomicity" than lower-weight SCUs. The information nuggets are also tailored to the contents of peer answers and are, at least in theory, meant to be atomic with respect to peers. But when we look at actual question answering evaluations, the identification of nuggets in the systems' answer allows for a lot of freedom and subjective interpretation by the annotator. The classification of nuggets into vital and non-vital is subjective, and can differ between different humans. In the question-answering settings, it is not possible to assign an empirical weight to a nugget, depending on the number of answers that contain it, since the nuggets are derived mainly from systems' answers rather than from answers that a human would produce. It will be interesting to further explore the parallels of the pyramid method and the nugget-based evaluation approach, possibly combining desirable characteristics from both in order to reach a unified evaluation framework for non-factoid questions answering and summarization, as has already been suggested, for example, in Lin and Demner-Fushman [Lin and Demner-Fushman 2005].

The most thorough analysis on the consensus of human summaries of the same text was presented by van Haltren and Teufel [van Halteren and Teufel 2003]. They collected 50 abstractive summaries of the same text and developed an annotation scheme for content units called *factoids*, analyzing the 50 abstracts in terms of factoids. Their large pool of summaries allowed for insightful observations and an empirical judgment that the appearance of new content with the addition of new summaries does not tail off. Their initial work was semantically oriented, also including an analysis between the relations among different factoids, much in the spirit of the van Dijk tradition—"factoids correspond to expressions in a FOPL-style semantics, which are compositionally interpreted" and they envisioned even further formalization of their mark-up. In their later work [Teufel and van Halteren 2004], where they included the analysis of another set of 20 summaries, they seem to settle to a representation closer to SCUs than on a first order logic language.

In their work, van Halteren and Teufel also address the question of *How many summaries are enough"* for stable evaluation results. Their investigation leads to the conclusion that 20 to 30 model summaries are necessary [Teufel and van Halteren 2004][4]. This conclusion is dramatically different from our study of pyramid evaluation where we established that about five human models are necessary for stable results. A careful analysis shows that there is no contradiction as it might

---

[4]In earlier work [van Halteren and Teufel 2003], they conclude that at least 30–40 are needed, but presumably the later results supercede these.

seem at a first glance and that actually two different questions were addressed in their work and ours.

The approach that Teufel and Halteren take is the following: they resample their pool of summaries (with possible repetitions) in order to get sets of N summaries for different values of N. Then for each pool of summaries derived in this manner, they score summaries against the factoid inventory using the weight of factoids in the peer summaries (without the normalization factor we propose). Then, for each pair of *system rankings*, regardless of the difference in scores, they compute the Spearman correlation coefficient and then take the average of the correlation coefficients for a given N. They deem a scoring reliable when the average correlation for a given N exceeds 0.95.

Our approach is more practical in nature—we assumed that a small difference in pyramid score does not necessarily entail a difference in summary quality. In fact, summaries with pyramid scores that differed by less than 0.06 were considered *equal* with respect to their information content. Then we proceeded to investigate what errors can arise in identifying summaries as being informationally equal or different (i.e., result in a change in system ranking). Consider, for example, the following scores for six systems under two different pyramid inventories.

system	sys1	sys2	sys3	sys4	sys5	sys6
Inventory 1	0.69	0.68	0.67	0.66	0.65	0.64
Inventory 2	0.64	0.65	0.66	0.67	0.68	0.69

In the pyramid analysis, all systems' summaries will be considered informationally equal under both inventories and thus the scores will be considered stable. But the rank correlation is perfectly negative, -1. So the apparent difference between their conclusion and ours in fact is due to the required strength of the expected results.

6. CONCLUSIONS

In this paper, we presented the Pyramid evaluation method, which is based on the semantic analysis of multiple human models. We demonstrated that the semnatic analysis into content units can be performed reliably and that Pyramid scores lead to stable evaluation results. Pyramid scores are highly correlated with direct overall judgments of the summary quality (summary responsiveness), but in addition they are also diagnostic, providing an indication for what important information is missing from a summary.

Part of the motivation for developing the Pyramid method was to provide a much needed in the summarization community evaluation metric that transparently incorporates the complexities of the the summarization task. The wide use of the method, both in large-scale evaluations and in individual studies, indicates that this goal has mostly been fulfilled. We hope that in the future data from Pyramid annotations will be also used to further research in abstractive summarization through the study of different verbalizations of the same content and the packaging of information in sentences.

REFERENCES

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37–46.

DI EUGENIO, B. AND GLASS, M. 2004. The kappa statistic: A second look. *Computational Linguistics*, 95–101.

DIXON, W. AND MASSEY, F. 1969. *Introduction to statistical analysis*. McGraw-Hill Book Company.

DONAWAY, R., DRUMMEY, K., AND MATHER, L. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of NAACL-ANLP Workshop on Automatic Summarization*.

GARNER, R. 1982. Efficient text summarization. *Journal of Educational Research 75*, 275–279.

GOLDSTEIN, J., KANTROWITZ, M., MITTAL, V., AND CARBONELL, J. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of ACM SIGIR'99*. 121–128.

JING, H., BARZILAY, R., MCKEOWN, K., AND ELHADAD, M. 1998. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*.

JOHNSON, R. 1970. Recall of prose as a function of structural importance of linguistic units. *Journal of Verbal Learning and Verbal Behaviour 9*, 12–20.

KRIPPENDORFF, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

LIN, C.-Y. 2004. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop in Text Summarization, ACL'04*.

LIN, J. AND DEMNER-FUSHMAN, D. 2005. Evaluating summaries and answers: Two sides of the same coin? In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Measures for MT and Summarization*.

L.R.DICE. 1945. Measures of the amount of ecologic association between species. *Ecology*, 297–302.

MANI, I. 2001. Summarization evaluation: An overview. In *NAACL 2001 Workshop on Automatic Summarization*.

MARCU, D. 1997. From discourse structure to text summaries. In *Proceedings of ACL/EACL-97 summarization workshop*. 82–88.

MARCU, D. 2000. *The Theory and Practice of Discourse and Summarization*. The MIT Press.

MARCU, D. 2001. Discourse-based summarization in duc 2001. In *Document Understanding Conference 2001*.

MCKEOWN, K. R., BARZILAY, R., EVANS, D., HATZIVASSILOGLOU, V., SCHIFFMAN, B., AND TEUFEL, S. 2001. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of DUC 2001*.

MINEL, J.-L., NUGIER, S., AND PIAT, G. 1997. How to appreciate the quality of automatic text summarization? In *Proceedings of the ACL/ECL'97 Workshop on Intelligent Scalable Text Summarization*. 25–30.

NENKOVA, A. AND PASSONNEAU, R. J. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Joint Annual Meeting of Human Language Technology (HLT) and the North American chapter of the Association for Computational Linguistics (NACL)*. Boston, MA.

PAPINENI, K., ROUKOS, S., WARD, T., AND ZHU, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

PASSONNEAU, R. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

PASSONNEAU, R., NENKOVA, A., MCKEOWN, K., AND SIGELMAN, S. 2005. Applying the pyramid method in duc 2005. In *Proceedings of the 2005 DUC Workshop*. Vancouver, B.C.

RADEV, D., JING, H., AND BUDZIKOWSKA, M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*.

RADEV, D. AND TAM, D. 2003. Single-document and multi-document summary evaluation via relative utility. In *Poster session, proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'03)*.

RATH, G. J., RESNICK, A., AND SAVAGE, R. 1961. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation 2,* 12, 139–208.

SALTON, G., SINGHAL, A., MITRA, M., AND BUCKLEY, C. 1997. Automatic text structuring and summarization. *Information Processing and Management 33,* 2, 193–208.

SORICUT, R. AND MARCU, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *HLT-NAACL*.

TEUFEL, S. AND VAN HALTEREN, H. 2004. Evaluating information content by factoid analysis: human annotation and stability. In *EMNLP-04*.

VAN HALTEREN, H. AND TEUFEL, S. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*.

VOORHEES, E. M. 2004. Overview of the trec 2003 question answering track. In *Proceedings of the 12th Text REtrieval Conference (TREC 2003)*. 54–68.

.....