Applying the Pyramid Method in DUC 2005

Rebecca J. Passonneau and **Ani Nenkova** and **Kathleen McKeown** and **Sergey Sigelman**
Columbia University
Computer Science Department
New York, NY 10027
{becky,ani,kathy,ss1792}@cs.columbia.edu

Abstract

In DUC 2005, the pyramid method for content evaluation was used for the first time in a cross-site evaluation. We discuss the method used in creating pyramid models and performing peer annotation. Analysis of score averages for the peers indicates that the best systems score half as well as humans, and that systems can be grouped into better and worse performers. There were few significant differences among systems. High score correlations between sets from different annotators, and good interannotator agreement, indicate that participants can perform annotation reliably. We found that a modified pyramid score gave good results and would simplify peer annotation in the future.

1 Introduction

Since 2001, the annual Document Understanding Conferences (DUC) have pursued the goal established in a 2000 roadmap to develop and evaluate sophisticated automated techniques for document summarization. However, developing evaluation methods for summarization has been difficult because human summaries vary for many reasons, including the knowledge, biases, goals, and intended audience of the summary writer. The pyramid method for content evaluation (Nenkova and Passonneau, 2004) addresses the variation in content across human summaries of the same source texts.

Designed to handle abstractive summarization, the pyramid method differs from previous evaluation methods primarily in assigning weights to content units, based on a model constructed from multiple human summaries. A new summary is rewarded more for containing information that occurs more often across a sample of human summaries. The research focus is thus to distinguish between more and less relevant information. As in previous work (van Halteren and Teufel, 2003), content is identified on the basis of shared meaning, not shared words or word strings (ngrams), thus this evaluation method leaves systems relatively unconstrained with respect to the way in which content is expressed. Here it is applied to systems which are primarily extractive.

To apply the pyramid method, DUC 2005 relied on manual methods for constructing the pyramid models associated with each document cluster for 20 sets, and for annotating the 25 peer summaries produced by systems, plus two by humans for each set. Columbia University constructed the pyramids, and participants in the evaluation did the peer annotations. Scores for the annotated peers were computed automatically as part of the annotation tool distributed by Columbia.

Our results show that pyramid scores group systems into better and worse performers, based on individual comparisons, although no single system can be identified as best across the different metrics used in DUC05 (original and modified pyramid, responsiveness, and ROUGE scores). Our analyses indicate that peer annotation is reliable on two measures, interannotator agreement, and consistency of scores. We also discuss results of a modified pyramid score that is analogous to recall; it correlates highly with the original score, but is easier to produce annotations for. Finally, analysis of the pyramids themselves show that humans produced summaries in 2005 that had more variation than summaries produced in DUC 2003 and we suspect that this is due to increased summary and document length as well as larger cluster size.

2 Pyramids

Twenty document clusters, or topics, were prepared by NIST assessors from TREC documents, following instructions provided at http://www-nlpir.nist.gov/projects/duc/duc2005/tasks.html. Each cluster was to contain between 25 and 50 documents relevant to a request for information created by the assessor; the average cluster size was 30.4 documents of 720 words each. For each topic, nine summaries of approximately 250 words each were written by humans. Of these, seven were used for each of the twenty pyramids. The remaining two were included in the peer evaluation.

Based on previous work, the use of seven summaries

SCU LABEL: Dogs are used in tracking suspects

W=7 **Contributors** *in context*

C. **They are excellent at tracking suspects and missing persons,** *and add muscle to law enforcement personnel in controlling crowds.*

D. **Working as a team with police, dogs help hunt down** *and capture criminal* **suspects.**

E. *Dogs can also pick-up scent from a crime scene* **and track persons who were at that site**

F. *Besides patrolling and narcotics detection, dogs are used in many types of police work, including detection of explosives and flammable liquids, crowd control* **and manhunts.**

G. **Such dogs cannot only track the suspects** *but are trained to tackle them as well.*

H. **They track** *and capture* **fleeing criminals** *or escaped prisoners.*

J. **Dogs are used for tracking a suspect,** *missing person or lost child.*

Figure 1: A sample SCU of weight 7

was assumed to be sufficient for creating the DUC 2005 pyramids. In (Nenkova and Passonneau, 2004), we found that between four and five summaries yielded scores that, for any pair of systems, would produce the same relative ranking as that given by larger pyramids. There, however, the peer and model summaries were shorter (100 words), the clusters were smaller (10 documents), and the average document length shorter (500 words). Another important consideration for choosing the number of model summaries was the time and labor costs for annotation.

2.1 Method of Pyramid Annotation

To create a pyramid, annotators identify units of information found within a pool of summaries written by different humans who have read the same cluster of documents. We refer to these units as *Summary Content Units*, or SCUs. An SCU includes a semantic label assigned by the annotator, primarily for mnemonic purposes, and consists of one continuous or discontinuous sequence of words (referred to as a *contributor*) from each summary that expresses the same information. Due to the concision typical of human summaries, we enforce a constraint that an SCU can have at most one contributor per model summary. The number of contributors per SCU thus ranges from a minimum of one to a maximum equal to the number of model summaries.

A copy of the annotation guidelines used for DUC 2005 is at http://www.cs.columbia.edu/∼ani/DUC2005 /AnnotationGuide.htm. A sample SCU is given in Figure 1. All 7 summaries express the information that the annotator labelled "Dogs are used in tracking suspects." For purposes of illustration, the contributors (boldface) are presented here in their sentential contexts (italics).

We mention three considerations for designing the annotation method. First, in order to minimize training requirements and make the annotation as widely useful as possible, annotation does not depend on knowledge of semantics, or logical formalisms. The judgments annotators are asked to make resemble those required in ordinary language use: to determine if there is something expressed in summary A that is also expressed in summary B. As mentioned in the guidelines, one summary might refer to the same time more precisely than another, as in "1993" versus "the early 90's". The semantic precision of an SCU depends on the relevance of the information, and is left to the annotator's judgment.

	SCU weight									Mean
Pyr.	1	2	3	4	5	6	7	N	SUM	SCU wt
D311	54	18	5	7	5	4	5	98	217	2.21
D324	39	20	11	11	4	3	1	89	201	2.26
D345	81	19	8	2	4	2	1	117	190	1.62
D366	55	14	5	13	5	3	1	96	200	2.08
D376	47	17	9	7	9	4	4	97	233	2.40
D391	93	28	12	5	1	1	1	141	223	1.58
D393	69	26	13	6	5	3	1	123	234	1.90
D400	130	18	9	11	1	1	2	172	262	1.52
D407	82	24	10	9	4	2	2	133	242	1.82
D413	53	29	12	2	11	5	2	114	254	2.23
D422	92	21	10	6	2	1	1	133	211	1.59
D426	80	22	12	8	9	5	7	143	316	2.21
D431	53	19	8	8	7	7	1	103	231	2.24
D435	78	21	17	2	5	2	3	128	237	1.85
D632	69	30	18	7	3	4	1	132	257	1.95
D633	55	21	10	5	4	1	2	98	187	1.91
D654	62	12	6	5	2	1	1	89	147	1.65
D671	62	21	14	9	6	5	1	118	249	2.11
D683	94	15	11	7	3	1	1	132	213	1.61
D695	114	20	2	1	1	1	1	140	182	1.30
AVG	73	21	10	7	5	3	2	120	224	1.90

Table 1: Distribution of SCUs at each weight

The second consideration is that the annotator is required to select specific words (continuous or discontinuous strings) as contributors, i.e., that express the labelled SCU content. This provides greater insurance that the annotator is comparing what is actually expressed across summaries than if annotators were simply required to create a list of ideas expressed.

The third consideration is that SCUs emerge from the annotators' judgment that the same information is expressed, independent of what words are used, or how many. Thus there is no *a priori* size in words or grammatical structures associated with SCUs, apart from the requirement that an SCU should not consist of more than one atomic tensed clause. There is no requirement that a contributor taken out of context should express the full meaning of the SCU. Although the phrase **and manhunts** in F of Figure 1 does not explicitly mention dogs or police work, they are clearly inferrable in the context.

Column 1 of Table 1 lists pyramids by the document set number (e.g., D435). Columns 2 through 8 indicate the number of SCUs at each weight from 1 to 7 for the twenty pyramids. Columns 9 through 12 give the total number of SCUs (N), the sum of their weights (SUM), and the mean SCU weight. As illustrated, there are typically a large number of SCUs of weight 1, meaning they occur in only one summary, and relatively few SCUs of the maximum weight. For SCUs of intermediate weight, the cardinality at each weight decreases as the weight increases along

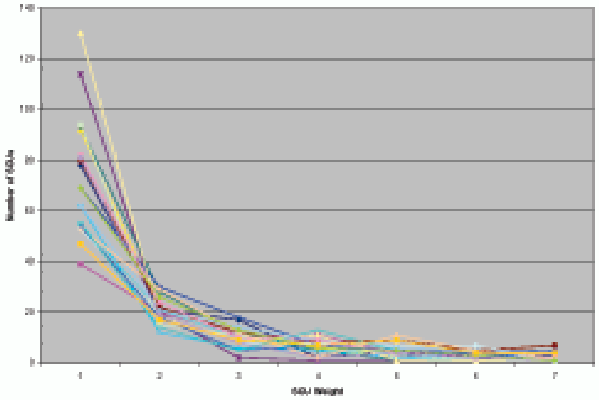

Figure 2: Line chart of cardinality of SCUs at each weight

a curve that is roughly the shape of a negative binomial, as illustrated in the line chart in Figure 2, which plots the information from columns 2-8 of Table 1. The x-axis represents each SCU weight from 1 to 7, and the y-axis represents the cardinality of SCUs of each weight. Though there is a clear pattern of global convergence, it is also clear that there are differences across document sets.

A comparison of distribution of SCU weights this year with SCU weights in DUC03 reveals that there are a larger proportion of SCUs of weight one this year. This year's mean SCU weight is 1.9, using 250-word models. For DUC 2003 document sets with 100-word models, we have three pyramids with ten model summaries each, having a mean SCU weight of 2.9, and five pyramids with seven model summaries each, with a mean SCU weight of 2.7. This year, on average, 60% of SCUs are weight one, whereas in the eight pyramids from 2003, 40% on average were weight one.

2.2 Annotation Task for Pyramid Creation

Five annotators at Columbia University consisting of graduate students, postdocs, research scientists and programmers participated in the pyramid creation task. Each was already familiar with SCU annotation, and used a tool designed specifically for pyramid and peer annotation, called DucView. Each annotator was assigned the task of creating four pyramids from sets of human-written summaries. As noted, each set consisted of seven 250-word summaries. Each annotator was also assigned four different pyramids to review. Annotators and checkers discussed how to handle cases of disagreement, and annotation decisions that remained problematic were discussed as a group until consensus was achieved. The two stages of review constitute a variant of a commonly used adjudication method.

Due to the resource demands for creating pyramids, we did not produce multiple annotations for the same pyramid. In (Nenkova and Passonneau, 2004), we reported an interannotator reliability value from two annotators on pyramid creation of $\alpha = .81$, using a variant of the method proposed in (Passonneau, 2004). The problematic issue in comparing pyramid annotations is that pyramids produced by different annotators generally yield differences in the total number of SCUs in the pyramid. This makes it difficult to apply agreement measures such as Cohen's κ or Siegel and Castellan's K, where the cardinality of coding units must be the same across coders. In addition, these and most other interannotator metrics fail to capture partial matches. A similar problem arises in comparing coreference annotations, as noted in (Passonneau, 2004), where it is argued that using Krippendorff's α (1980) to measure partial agreement is more appropriate for set-valued data. (Artstein and Poesio, 2005) extend this approach by showing how to capture the best features of α, κ and related metrics.

3 Peer Annotation and Score Computation

Peers were annotated by evaluation participants, with each annotator doing a full set of peers produced for one document cluster. Of the twenty document sets, six were given to more than one peer annotator and we report on interannotator agreement in section 5. Given a model pyramid for a document set, each peer summary is annotated against the corresponding pyramid. Peer annotation is simpler than pyramid annotation because the set of SCUs is already fixed.

Full instructions for peer annotation are in the guidelines referred to above. After familiarizing themselves with the SCUs in a pyramid, annotators select words in the peer summary that express the same information expressed in an SCU, and co-select the matching pyramid SCU. As with pyramid annotation, discontinuities are allowed. Also, if the peer has expressed the same information more than once, the annotator can reselect the same SCU. If the peer expresses information that does not appear in the pyramid, the annotator selects a special 'nonmatching SCU' in the pyramid.

Two types of scores for peers were computed from the peer annotations. Both scores are a ratio of the sum of the weights of the SCUs found in the peer (OBServed) to the sum for an ideal summary (MAXimum). If the number of SCUs of a given weight i that occur in a summary is O_i, the sum of the weights of all the SCUS in a summary is:

$OBS = \sum_{i=1}^{n} i \times O_i$

In the original pyramid score, the number of SCUs used in computing MAX is the same as the number used to compute OBS. If we label the pyramid tiers by their weight (T_i), it is given by:

$\text{MAX}_O = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|)$

where $j = \max_i(\sum_{t=i}^{n} |T_t| \geq X)$

Like precision, the ratio $\frac{OBS}{MAX_O}$ indicates the proportion of SCUs in the peer that were as highly weighted as they

peer	score	95% conf. interval
B	0.5428	(0.4719; 0.6138)
A	0.4900	(0.4360; 0.5440)
14	0.2477	(0.1816; 0.3139)
17	0.2398	(0.1848; 0.2949)
10	0.2340	(0.1825; 0.2855)
15	0.2322	(0.1816; 0.2829)
7	0.2307	(0.1723; 0.2891)
4	0.2197	(0.1720; 0.2674)
16	0.2170	(0.1540; 0.2799)
32	0.2134	(0.1640; 0.2628)
6	0.2110	(0.1535; 0.2684)
19	0.2089	(0.1543; 0.2636)
12	0.2086	(0.1597; 0.2575)
11	0.2085	(0.1480; 0.2690)
21	0.2063	(0.1638; 0.2488)
26	0.1970	(0.1288; 0.2652)
28	0.1944	(0.1496; 0.2393)
3	0.1894	(0.1495; 0.2292)
13	0.1855	(0.1217; 0.2492)
25	0.1691	(0.1144; 0.2237)
1	0.1666	(0.0944; 0.2388)
27	0.1631	(0.1084; 0.2177)
31	0.1587	(0.0875; 0.2298)
24	0.1491	(0.1010; 0.1972)
20	0.1446	(0.0876; 0.2014)
30	0.1376	(0.0970; 0.1782)
23	0.1216	(0.0662; 0.1769)

Table 2: Average peer original pyramid scores

peer	mod. score	95% conf. interval
B	0.4818	(0.4050; 0.5585)
A	0.4629	(0.3989; 0.5269)
10	0.2000	(0.1509; 0.2491)
17	0.1972	(0.1471; 0.2473)
14	0.1874	(0.1283; 0.2466)
7	0.1840	(0.1251; 0.2428)
15	0.1793	(0.1378; 0.2209)
4	0.1722	(0.1298; 0.2146)
16	0.1706	(0.1124; 0.2288)
11	0.1691	(0.1150; 0.2233)
19	0.1672	(0.1187; 0.2158)
12	0.1645	(0.1189; 0.2101)
6	0.1639	(0.1216; 0.2062)
32	0.1607	(0.1184; 0.2030)
21	0.1589	(0.1238; 0.1940)
3	0.1459	(0.1082; 0.1837)
26	0.1413	(0.0895; 0.1931)
13	0.1412	(0.0915; 0.1908)
28	0.1400	(0.1045; 0.1756)
25	0.1395	(0.0900; 0.1889)
27	0.1306	(0.0862; 0.1749)
1	0.1258	(0.0603; 0.1912)
31	0.1215	(0.0633; 0.1796)
24	0.1140	(0.0770; 0.1511)
30	0.1131	(0.0794; 0.1469)
20	0.0937	(0.0547; 0.1327)
23	0.0608	(0.0284; 0.0933)

Table 3: Average peer modified pyramid scores

could be. In the modified score, MAX_M is computed using the average number of SCUs found in the seven human model summaries in the corresponding pyramid, which is the total number of SCUs in the pyramid, divided by the number of models:

$$\text{MAX}_M = \frac{\sum_{i=1}^{n} |T_i|}{j} \text{where } j = \max_i.$$

Like recall, $\frac{OBS}{MAX_M}$ indicates the proportion of the *target* highly weighted SCUs that were found in the peer. The average number of SCUs in all models was 17, and in all peers was 7.4, thus the modified scores are lower.

4 Score Results

As noted in section 3, we compute two scores for each peer summary, the original and modified pyramid scores. Tables 2 and 3 show the score averages across the 20 document sets for each of the 27 system and human peers. In the six cases where we have scores from two annotators (324, 400, 407, 426, 633, 695), the pairs of scores were first averaged for each peer on each document set.

The human peer average is roughly double that of the best automatic system, and several systems have an average performance lower than that of the baseline. The 95% confidence intervals for the mean of each system, also listed in the tables, are quite wide, indicating that there was considerable variability in the peer performance by set. Analysis of variance with the pyramid scores as a dependent variable and the set and peer as factors show that both factors are significant, $p = 0$.

The ranking of the peers based on average scores is very similar for the original and the modified pyramid scores. In fact, correlations between the original and modified average peer scores are very high:

pearson	0.9941
spearman	0.9810
kendal's tau	0.9047

The top six systems, for example, are the same for both modified and original scores. Five of these top six systems were also rated as the top systems by responsiveness showing consistency across manual scores. Despite consistency in group, each metric, whether manual or automated, ranks a different system as the best; the original pyramid score ranks system 14 as best, the modified ranks 10 best, responsiveness 4, and ROUGE-SU4 15.

The high correlations between the peer averages suggest that the modified score can be used for comparing systems. This would diminish the annotation load, because the modified pyramid score does not require content in the peer that does not correspond to SCUS in the model pyramid to be broken down into distinct SCUs of zero weight. This is both a tedious task for annotators and one on which they are very inconsistent. For the remainder of the paper, we present our analyses for the modified scores only.

Modified scores	
peer	better than
10	25 27 31 24 30 20 23
17	25 27 24 30 20 23
14	25 27 1 24 30 20 23
7	13 25 27 31 24 30 20 23
15	3 25 27 1 24 30 20 23
4	25 27 31 24 30 20 23
16	24 30 23
11	24 30 23
19	24 30 23
12	30 23
6	31 30 20 23
32	24 30 20 23
21	24 30 23
3	30 23
26	23
28	30 20 23

Table 4: Significant difference based on individual comparisons. Based on paired Wilcoxon rank-sum, $\alpha = 0.05$

Modified scores	
peer	better than
21	23
32	23
6	23
12	23
19	23
11	23
16	23
4	23
15	23
7	23
14	23
17	23 20
10	23 20

Table 5: Significant differences between peers according to multiple comparisons with Tukey's method, $\alpha = 0.05$.

In previous years, confidence intervals were used to show significant differences between systems. Confidence intervals are an approximation of a direct test that we use here (Schenker and Gentleman, 2001), specifically using paired comparisons to determine significant differences between any one peer and all others. To make these individual comparisons on a peer-to-peer basis, the exact paired Wilcoxon rank sum test, $\alpha = 0.05$, was used. The use of individual comparisons is particularly helpful for system development; any one developer can see how their system compares with others. The significant differences for modified scores between individual pairs of systems are listed in Table 4. In each row, the peer system appears in column one and in the remaining columns of the row, the systems for which the peer performed significantly better are shown. This table reveals a set of six systems (10, 17, 14, 7, 15 and 4) which tend to outperform a larger number of systems.[1]

Peer	Better than (primary)	Better than (secondary)
26	23	
20	23	
3	23	23
32	23	
25	23	
7	23	23
12	23	23
27	23	
6	23 31	
16	23 31	23
19	23 31	
24	23 31	
21	23 31	
28	23 31	23
11	23 31	
17	23 31	
15	23 31 1	23
10	23 31 1	23 31 20
14	23 31 1 30 26 13 20	23 31 20
4	23 31 1 30 26 13 20 3	23
B	*All systems*	*All systems*
A	*All systems*	*All systems*

Table 6: Significant difference between peers based on the primary and secondary responsivness judgements

With both confidence intervals and individual comparisons, if they are used to compute comparisons between every combination of pairs, there is the risk of introducing experiment-wise Type I errors given the large number of comparisons. To determine significant differences between all combinations of systems, Tukey's honest significant difference method based on confidence intervals was used to compare peers for significant differences, with experiment-wise type I error ≤ 0.05. This is a more conservative test and thus, there are few significant differences between peers when we control for the overall error for the multiple comparisons; these are shown in Table 5. From these results we can conclude that, in most cases, only system 23 performed significantly worse than the others, although system 20 was significantly worse than two systems, 17 and 10. The lack of significant differences between systems is similar to system performance on Task 4 for DUC 2004, a summarization task also focused by question (Nenkova, 2005).

The manual responsiveness score, computed by NIST, shows a similar trend (see table 6). There were two sets of responsiveness judgements—primary, done by the person who wrote the topic, and a secondary one, done by a different NIST annotator.

Table 7 shows correlations between the different types of scoring used in DUC-05. Clearly the original and modified scores are mutually substitutable (correlation above

[1]The same pattern occurs with original scores.

	Pyr modified	Resp (primary)	Resp (secondary)	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
Resp (prim)			0.83	0.92	0.92
Resp (sec)				0.88	0.87
ROUGE-2					0.98

Table 7: Pearson's correlation between the different evaluation metrics used in DUC 2005

.95) as are two variants on the ROUGE scores, ROUGE-SU4 and ROUGE-2. We see, however, that the two human scores on responsiveness, while highly correlated, are not mutually substitutable. This suggests that responsiveness cannot be used as a gold standard, since depending on which human answered the question, systems are ranked differently. Similarly, the two pyramid variants are highly correlated with responsiveness, although they are not mutually substitutable. The correlation between ROUGE and pyramid scores are also in the same range (from .80 to .90). And while ROUGE correlates better with the primary set of responsiveness judgements, its correlation with the secondary responsiveness judgements is lower, while the pyramid correlations with secondary responsiveness annotations are higher. Most important, it should be noted that significance testing reveals that there is no significant difference between any of the correlations shown in this table.

5 Interannotator Agreement

5.1 Preliminaries

Krippendorff (1980) proposed that agreement of .67 should be sufficient for many tasks, and this number has often been cited as the necessary threshold for many annotation tasks. However, he further argues that the target agreement value should be empirically determined with respect to the costs pertaining to the use of the data. Frequently, annotation projects in CL lack a means of measuring the cost of variations in interannotator agreement, because annotations are often collected independently of their use, and are used in many ways. For the DUC 2005 data, however, we can directly investigate the relationship between interannotator agreeement and the scores associated with the multiply annotated peers.

For six of the 20 document sets, we have two peer annotations to compare for interannotator agreement. For one of the sets, a third annotator was provided with a reduced pyramid eliminating SCUs of W=1.

To compare pairs of annotations of peers against the same pyramid, we record for each SCU in the pyramid whether the annotator found it in the peer summary, and how often. This yields an agreement matrix where each $\text{cell}_{i,j}$ indicates how often annotator_i finds SCU_j in the peer. Thus in contrast to comparing pyramid annotations, peer annotations will always have the same number of coding units, making it possible to apply Cohen's κ. Because of its familiarity to the CL community, we computed the six κ scores, finding an average value of .57, without accounting for partial matches.

Annotators	Setid	α	Dice	α_{Dice}
102,218	324	0.59	0.71	0.67
108,120	400	0.45	0.72	0.53
109,122	407	0.41	0.59	0.49
112,126	426	0.54	0.74	0.63
116,124	633	0.58	0.87	0.68
121,125	695	0.51	0.75	0.61
102,123	*324*	*0.60*	*0.82*	*0.69*
218,123	*324*	*0.49*	*0.66*	*0.56*

Table 8: Correlation and agreement coefficients

Different annotators will typically find a different number of SCUs in the same summary, as well as a different selection of SCUs. In addition, the same SCU can occur more than once in a peer, particularly for a summary that repeats the same sentences or phrases. Cell values in the matrix thus range from 0 to the number of repetitions of the SCU that occurs most often in a peer.

Krippendorff's (1980) α coefficient handles partial agreement. It differs from other agreement coefficients in that it is computed by comparing disagreements, although the range and directionality of its values is the same. As shown in (Passonneau, 1997), in their most general form (i.e., independent of the method of estimating probability), κ and α are equivalent.

α can be weighted to capture partial matches, but we need an appropriate distance metric (δ) to compare values. For nominal data, when any pair of values from independent coders are compared, say r and s, δrs 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. We use this δ for our first of three comparisons.

For illustrative purposes, we also use the Dice coefficient, which is often used as a similarity measure in document processing tasks. The Dice coefficient is computed from 2x2 contingency tables that indicate where annotators agree and disagree, and is equivalent here to F-measure. Let a represent the number of times both annotators agree that an SCU appears in the current peer sum-

mary, b the number of times that the first annotator finds SCUs that the second annotator does not, and c the number of times that the second annotator finds SCUs that the first does not. The formula for Dice is then: $\frac{(2a)}{(2a+b+c)}$. In our example where annotator A finds an SCU three times in a peer, and annotator B finds the same SCU twice, $a = 2$, $b = 1$, $c = 0$, and Dice equals .8. In our third comparison, we use α with (1-Dice) as the distance metric for comparing coding values from distinct annotators. In our example, (1-Dice)= .2, so is more representative of the true agreement than the nominal distance metric.

5.2 Agreement Results

Table 8 presents our three measures of comparison for the eight pairs of doubly annotated peers. The three rightmost columns correspond to unweighted α, the Dice coefficient, and α where $\delta = (1 - Dice)$. The rows represent pairs of annotators on a particular document set and pyramid. The italicized rows include annotator 123 in the pair, who received a reduced pyramid (see above). The cell values represent averages across the 27 peers in each set. Analysis of variance with each measure in turn as the dependent variable, and annotator pair, set, and peer as factors, shows no sigificant difference in variance on agreement, so it is fair to report average agreement.

The column labeled α in Table 8 indicates that not counting partial matches, agreement between annotators on sets 324 and 633 turn out to be about equally good, and closer to perfect agreement (1) than to perfectly random (0), and the overall average (.52) is very close to the corresponding κ (.57). Because Dice is a ratio of agreements to disagreements, the Dice values in Table 8 tell us that by and large, different pairs of annotators agree. The pair who did set 407 had relatively less agreement, and the pair who did set 633 had relatively more. However, by factoring out chance agreement, α reveals more than Dice, namely that 324 and 633 were equally good.

Because it accounts for partial matches, α_{Dice} gives values that are higher than those for α. For the non-italicized rows, values for α_{Dice} range from .49, about halfway between chance and perfect agreement, to .68 for sets 324 and 633, this being the magic threshold identified by Krippendorff, independent of annotation task. One could speculate as to whether the annotators who did sets 324 and 633 were more careful than the other annotators, or whether these sets were less problematic for annotators. However, a much more direct means of evaluating whether this threshold is *good enough* is to compare the scores derived from the different pairs of annotators.

Finally, the italicized rows represent annotators 102 and 218 from the first row paired with annotator 123, who had the reduced pyramid. On the two α measures, the two pairs {102,218} and {102,123} are both higher and closer in value than the two pairs {102,218} and {218,123}. This means that of the three annotators, 218 is the low outlier. The most positive interpretation is that 102 and 123 are more correct than 218. Note that unless we have advance knowledge about a particular annotator (e.g., from previous reliability studies), comparing only two annotators cannot tell us which one was more careful or consistent, but we know from inspecting the peer annotations that some annotators were much more careful and consistent than others. The three pairs of comparisons on set 324 allow us to see this difference quantitatively. Annotator 123 differed in the pyramid used, so the reliability results also suggest that when an annotator (e.g., 123) uses a much smaller pyramid that excludes the least important SCUs, this does not degrade the resulting peer annotation. It would be useful to repeat the experiment with more annotators, but a complete picture depends on comparing the pyramid scores for these three annotators who did the same set, which we do next.

Annot.	Set id	original scores		modifi ed 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)
102,123	*324*	*.70*	*(.44,.85)*	*.73*	*(.48,.87)*
218,123	*324*	*.60*	*(.29,.80)*	*.77*	*(.55,.89)*

Table 9: Pearson's correlations for original and modified scores of the paired annotations

5.3 Score correlation for distinct annotations

Here we compare the original and modified pyramid scores computed from different annotations of the same peer summaries. Table 9 shows Pearson's correlation on scores, and the confidence intervals, for both sets of scores (p = 0 for all but {218,123}, where p=0.00009). In general, the correlations are high, and the differences between the correlations for the original and modified score are relatively small, with the possible exception of the two rows involving annotator 218. The interannotator agreement tests in the preceding subsection have already suggested that 218 is a less reliable annotator. For the four sets 400, 407, 426 and 695 the correlations are relatively higher (about .9), especially on the modified scores. The two italicized rows involving the reduced pyramid have the lowest correlations and the widest confidence intervals, indicating that the corresponding scores have the most variance. This is also reflected in the relatively high ratios of score variances: 1.6 for {102,123} and 1.4 for {218,123} versus 1.2 for {102,218}.

6 Discussion

Our results indicate that the pyramid method differentiates systems from humans and that the best systems are currently about half as good as humans at identifying information that a sample of humans would find relevant. Use of individual comparisons based on the paired Wilcoxon rank-sum, $\alpha = 0.05$ reveals a group of six systems that consistently outperform others and this group remains the same across original and modified pyramid scores as well as in the responsiveness scores. However, the different metrics used for evaluation, including both manual and automatic, do not consistently pinpoint any one system as performing best. Correlations between different metrics show that different versions of the same metric can substitute for each other (i.e., original for modified pyramid and ROUGE-2 for ROUGE-SU4). Correlations between all pairs of metrics are high even when not mutually substitutable and none of the correlations are significantly different from any other.

While responsiveness has correlations of about .90 with pyramid or ROUGE scores, depending on the human assessor, responsiveness scores do not correlate as highly with each other (.80), indicating responsiveness cannot serve as a gold standard.

The results on interannotator agreement and especially on the correlation of pyramid scores across pairs of annotators on the same peers indicate that participants could perform peer annotation reliably, with no prior exposure to the method. The largest inconsistency between annotators comes from how they chose to break down non-matching SCUs into clauses and thus, if the identification of non-matching SCUs can be eliminated from the task, interannotator agreement would increase.

The high correlations between the original and modified scores suggest that modified scores could be used instead of the original scores, which as noted above, would simplify the peer annotation. Since MAX depends on the average number of SCUs in the models, not on the number of SCUS in the peer, annotators could safely ignore the annotation of content in a peer that does not correspond to content in the pyramid.

7 Future Work

Differences across document sets, pyramids, and amount of score variance are likely related. Analysis of variance with scores as the dependent variable indicated that set was a significant factor ($p = 0$), which could be due to the document sets, the associated pyramids, or more likely, both. A comparison of SCU weight distributions between 2003 and 2005 shows a large difference (2.8 in 2003 versus 1.9 in 2005), which also is probably due in part to differences in the document sets across years, given the disparity in cluster and document size. However, we further hypothesize that the model summary lengths are another factor. In concurrent work, we have found that for a set of eight pyramids using seven 100-word models (the four original 2003 DUC model summaries, plus additional ones we collected), average pyramid scores for sixteen systems have a wider range (δ=.3430 for 2003; .1392 for 2005), along with narrower 95% confidence intervals (ranging from .0178 to .1030, compared with .0649 to .1535). We hope to investigate these interdependencies.

8 Acknowledgements

This work was supported by DARPA DI NBCH1050003. We thank the pyramid and peer annotators for their efforts. We specifically acknowledge the help of Advaith Siddharthan and David Elson for their help in the initial model pyramid annotation. We also thank Stacy President for testing out the peer guidelines and system.

References

Ron Artstein and Massimo Poesio. 2005. Kappa cubed = alpha (or beta). Technical Report NLE Technote 2005-01, University of Essex, Essex.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*.

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage Publications, Beverly Hills, CA.

Ani Nenkova and Rebecca J. Passonneau. 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.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, Pittsburgh, USA.

Rebecca Passonneau. 1997. Applying reliability metrics to co-reference annotation. Technical Report CUCS-017-97, Department of Computer Science, Columbia University, New York.

Rebecca Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portugal.

Natalie Schenker and Jane Gentleman. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186.