Columbia Multi-Document Summarization: Approach and Evaluation

Kathleen R. McKeown Regina Barzilay David Evans
Vasileios Hatzivassiloglou Min Yen Kan Barry Schiffman Simone Teufel

Department of Computer Science
450 Computer Science Building
Columbia University
New York, N.Y. 10027
{kathy, regina, devans, vh, min, bschiff, teufel}@cs.columbia.edu

1 Introduction

Different forms of summarization are useful in different situations, depending on the intended purpose of the summary and on the types of documents summarized. Columbia's multidocument summarization system for DUC builds on this observation. Given that DUC makes no specific assumptions about the task where the summary is to be used [Spärck-Jones 2001] and includes many different document types, we developed a composite summarization system that uses different summarization strategies dependent on the type of documents in the input set. In our system, a router automatically determines the type of the input set of documents and invokes the appropriate summarization subcomponent. The focus of our system is multidocument summarization and for this task, the challenges are to identify similarities and important differences across the input set of documents.

A main focus of our work to date has been on summarization of sets of documents that all describe the same event, as in current news, where repetitions of core information across sources is expected. This strategy can be effective when there is a lot of similar information that needs to be presented concisely, but is not as effective when the input data spans different weakly related events or a long time period. Consequently, we developed an alternate summarization strategy that can be adapted to documents of different types, including biographies and multiple weakly related events.

To summarize documents on the same event, the Columbia summarizer uses an enhanced version of MultiGen [Barzilay *et al.* 1999; Hatzivassiloglou *et al.* 1999; McKeown *et al.* 1999; Barzilay *et al.* 2001; Hatzivassiloglou *et al.* 2001]; for biographical documents, it uses an alternate system, DEMS (Dissimilarity Engine for Multidocument Summarization), tuned to the biographical task; and for sets of loosely similar documents, it uses DEMS with a more general configuration. DEMS incorporates techniques used in the BioGen system, a system developed jointly by Mitre and Columbia [Schiffman *et al.* 2001], as well as techniques we have been developing to identify differences between input articles. While we have spent several years on the development of MultiGen, the alternate summarization strategies used in DEMS were developed in a matter of weeks.

In this paper we discuss Columbia's summarization system for DUC, covering the different component summarizers that handle different document types, the router that decides which summarizer to use, and a preliminary analysis of evaluation results relative to other systems and of factors such as the document types and the model summaries that affect the evaluation. Our analysis shows that Columbia's system consis-

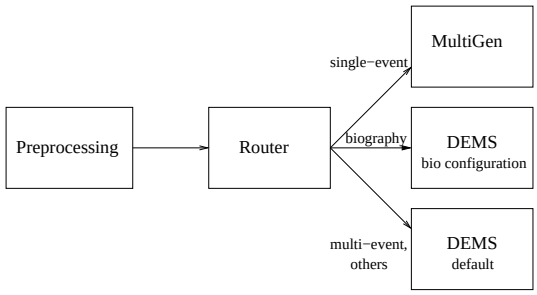

Figure 1: Columbia System Architecture

tently falls within a cluster of five top systems, each of which did better on a number of different metrics.

2 System Overview

Columbia's system follows a pipeline architecture, shown in Figure 1. During the preprocessing stage, the input articles are transformed to a uniform XML representation. Then, the router component of the system determines the type of each input document set as one of three possibilities and directs the input texts to one of the underlying summarizers. One of the document types is processed by MultiGen, which operates over sets of articles describing the same event, while DEMS handles the two other more heterogeneous types, such as "Person-centered" and "Multi-event", with different feature weights depending on the input type.

2.1 Data Preprocessing

We first transform the incoming data into a uniform XML format. The data is originally in several different SGML formats, one for each publisher (e.g., Wall Street Journal, Associated Press). For instance, one format might use the label `DATE` while others use `DT` or `PUBLISHDATE`. This means that the semantics of the SGML tags in the different input formats can only be interpreted using world knowledge. The output of our preprocessing is valid XML conforming to a DTD (document type description) we wrote to describe the semantics of the data. We also clean up the text itself: subtitles are automatically determined and marked and initial datelines and bullet points are recognized.

2.2 Router

The router is a program module that decides the *type* of each document set. Using the training corpus, we manually derived the following typology of document sets:

- *Single-event*: The documents center around one single event happening at one place and at roughly the same time, involving the same agents and actions. A good example for a single-event document set is the set reporting on the eruption of Mount Pinatubo in the training data (D29e).

- *Person-centered* ("biography"): The documents deal with one event concerning one person but include background information about that person, usually in the form of additional events in the past, or in the form of follow-up events of the initial person-centered event. An example in the training data is the document set describing Alan Greenspan's career (D49i). The time covered is typically longer than in the single-event case.

- *Multi-event*: Several events occurring at different places and times, and usually with different protagonists, are reported together. There is a common theme to these events, e.g., a document set might collect many fire incidents on unrelated cruise ships (D21d), or many solar eclipses (D51i). The time span covered is unpredictable, but longer than in the single-event case.

- *Other*: Such document sets contain even more loosely related documents, like the set describing research and conflicting policies concerning the Antarctica (D07b), or the one covering the entire Iraq-Kuwait war (D09b). Of all document set types, we observed the longest time span in this category.

Among the 30 sets in the training corpus, we only found two single-event document sets, but 10 person-centered sets and 7 multi-event sets, whereas 11 sets were so loosely connected that they could only be described as "other".

The router uses the following information when deciding the document set type:

- Overall time span between publication dates. Note that data preprocessing is necessary to bring the publication date information into a uniform and comparable format.
- Proportion of articles published on the same day, in comparison to the number of articles in the set.
- Frequency of capitalized words in order to roughly approximate named entities.
- Frequency of pronouns "he" and "she".

If the overall time span between publication dates is less than 80 days, or if more than 50% of all documents are published on the same day, we hypothesize that one event is predominant in the document set, and the set is routed to MultiGen. We also experimented with the longest and shortest distance between publication dates within a document set, but found the overall time span and the same-day feature to be more robust indicators of document set type. In document sets containing biographies, we found that one capitalized word stood out above all others, given that there were frequent references to one person. As we did not have the time to incorporate a named-entity component in the system, this strategy is an acceptable compromise. Furthermore, the number of personal pronouns was high. Thus, the system routes the document set to DEMS in "biography mode" if the frequency of pronouns is above 0.018, and the frequency of the most frequent capitalized word is above 0.012.[1] Otherwise, DEMS summarizes in default mode, thus classifying the documents as weakly related.

[1] These values were empirically derived during training.

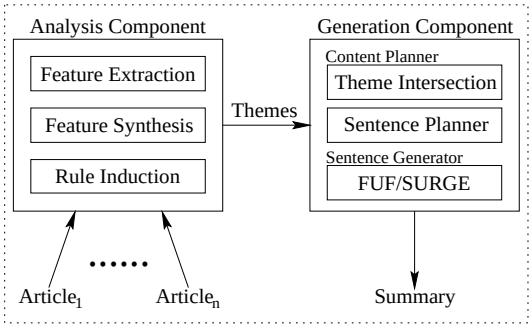

Figure 2: MultiGen Architecture

The router does not further distinguish between the "Multi-event" and "Other" types of document sets, since we did not develop a different strategy for the "Other" category in the limited time available for the DUC evaluation.

The next subsections describe the different summarization systems used for summarizing documents once their type has been determined by the router.

2.3 MultiGen

MultiGen summarizes a specific type of input: news articles presenting different descriptions of the same event. Repeated information about the event is a good indicator of its importance to the event, and can be used for summary generation. Our approach is unique in its integration of machine learning and statistical techniques to identify similar paragraphs, intersection of similar phrases within paragraphs, and language generation to reformulate the wording of the summary.

MultiGen follows a pipeline architecture as shown in Figure 2. The analysis component of the system breaks documents into smaller text units and then computes a similarity metric across text units, regardless of the source document. Once similar paragraphs are identified, they are passed to the generation component which further identifies and selects information to be reformulated as coherent text.

The analysis, or similarity computation component, takes as input a set of articles. It breaks the article into sentence-sized units for comparison, and then extracts a set of linguistic and posi-

tional features, which serve as input into the similarity algorithm. These features include primitive features such as word, stem and WordNet overlap as well as composite features, which aim to capture matches on the syntactic level such as subject-verb and verb-object relations. We construct a vector for each pair of sentences, representing matches on each of the different features. We use a log-linear regression model to convert the evidence from the various features to a single similarity value. The model was trained on a set of 10,535 pairs of paragraphs which were manually marked for similarity. The output of the model is a listing of real-value similarity values on sentence pairs. These similarity values are fed into a clustering algorithm, that partitions the text units into clusters of closely related ones. The clustering is performed using a non-hierarchical clustering technique, the *exchange method* [Späth 1985], producing clusters of closely related sentences that we term *themes*.

The generation component consists of an ordering component, an intersection component, and a sentence generator. The goal of the ordering component is to order themes into coherent text which respects the chronological order of the main events. To implement this strategy in MultiGen, we select for each theme the sentence which has the earliest publication day (*theme time stamp*), and order themes according to their time stamps. To increase the coherence of the output text, we identify blocks of themes which talk about the same event and apply chronological ordering on blocks of themes. Once the themes are ordered, the content planner (i.e., intersection component) identifies phrases within the sentences of each theme that are close enough to other phrases in the theme that they can be included in the summary. It does this by producing a predicate-argument structure for each sentence in each input sentence, comparing arguments to determine phrases that are similar. The sentence planner then determines which phrases should be combined into a single, more complex sentence, looking again at constraints from the input document as well as common references between phrases. Finally, the constituent structure produced by these two stages is mapped to the linearizer which generates an English sentence for each theme with a non-empty intersection.

MultiGen was developed prior to DUC, and its architecture was not significantly altered for the competition, except for a modification needed in order to produce summaries of different length. By default, MultiGen produces very concise summaries where 50 or more related news articles are reduced to one or two paragraphs describing the salient sub-events. To meet the target lengths of DUC with the relatively sparse input provided (approximately 10 documents in each set), we had to adjust internal intersection parameters for more verbose output and rank the different themes. This ranking is based on theme size, similarity score and significance. The first two of these scores are produced by the similarity component, and the significance score of the theme is computed using *lexical chains* [Barzilay and Elhadad 1997], as the sum of lexical chain scores of theme sentences computed from the text in which a sentence originally belongs. Lexical chains, sequences of semantically related words, are tightly connected to the lexical cohesive structure of the text and have been shown to be useful for determining which sentences are important for single document summarization. Here, a theme which has many sentences which have been ranked by lexical chains to be important for a single document summary, is, in turn, given a higher significance score for the multi-document summary.

2.4 DEMS: Sentence Extraction for Biographies and Differences

One of our goals in the development of DEMS was to provide a robust alternate summarizer within a short development time, so that document sets that do not conform to single event descriptions can still be summarized. Some of the sets provided in the training and evaluation sets cover many events related only loosely; for example, one set in the evaluation corpus contained 10 articles covering 9 unrelated political murders that occurred in places all around the world, spanning the years from 1979 to 1994:

- the Palestine Liberation Army loses its second in command
- Communist rebels kill a local police chief in the Philippines
- a presidential candidate is assassinated in Colombia (two articles)
- Egypt's parliamentary speaker is assassinated
- a Jordanian diplomat is killed
- a presidential candidate in Mexico is killed
- news on the killing of a Korean presidential candidate 15 years before
- No. 2 official in Mexico's ruling party is killed
- India's former prime minister is assassinated

Such a set does not provide enough coherence for a summary driven by validating information from one document against another. Indeed, it seems that the best summary for such document sets might be a list enumerating the different events. Therefore, the data we encountered in DUC raises a number of important issues: How will systems choose which topics to include in the short summaries? How will a summary be coherent without explaining why this odd list is presented? What glue holds this group of documents together? Rather than attempting to fully address these questions, we have designed DEMS as an alternative summarizer that can handle even these very loosely related documents.

Because of the time limitations involved leading up to the DUC evaluation in July, a decision was made early on to use sentence extraction in DEMS. DEMS uses four categories of features to determine which sentences to extract to form the summary. One of these categories is relevant specifically to biographies, while the others are relevant to determining important differences and are used in both configurations.

Summarization Features for Loosely Related Documents DEMS uses four classes of features for determining which sentences to include in the final summary. The most innovative features are included in the class that attempts to measure the importance of the words in a unit of text. One of these features is an importance measure derived from the analysis of a large corpus of news. We used a lexicon of key terms, nouns, verbs and adjectives that were more likely to appear in the first paragraph of a news article than in the entire article. This tells us when sentences use terms that are likely to be considered important, since journalists tend to include important information in the lead paragraph. Another feature in the class used to measure informativeness is based on a study of verbs that was done for the BioGen [Schiffman *et al.* 2001] system. The idea is that a verb associated with a large number of subjects is not likely to express important content by itself. For example, the verb "arrest" is strongly associated with the subject "police," but not with a large number of other nouns. Thus, the verb "arrest" conveys some contextual information that a verb like "happen" would not.

The second class used information to weight higher sentences which contained words related to the semantic themes covered in all documents of the input set. Thus, instead of computing word frequency, we counted semantic groups in the entire set being summarized. We also counted the semantic groups within each member document in the set in order to weight higher semantic themes within a document that were unique points of that document. We used semantic groups derived from WordNet, putting together synonyms, hypernyms, and hyponyms but excluding words that had more than five senses. The frequency count for a specific semantic group was incremented each time we encountered a word from that group.

The structure of the articles in the news domain prompted a third group of features, namely the date of publication and the location of the sentence within the article. We weighted the sentences giving more importance to sentences occurring near the beginning of an article than

the end. The document sets spanned a rather long time in many cases and use of the date allowed us to guard against losing the most current news. Sentences from articles with more recent dates were weighted more heavily.

A fourth group of features were based on syntax and style, including the presence and location of pronouns and the length of the sentence. We found that very short sentences were usually cryptic, while overly long sentences contained extraneous information. We set the length of an ideal sentence at 20 words and computed the absolute value of the actual length to the ideal length and used it as a negative value. The presence of pronouns was also weighted negatively, to avoid dangling references.

Biographies The last group of features targets the biographical document sets, which were those that covered a sequence of events surrounding one individual. These sets had a stronger focus than many of the general sets and needed some special attention. It was clear that for these document sets, sentences mentioning the subject of the biography by name were good candidates for inclusion in summaries. The main feature here was a binary value reflecting whether or not the target individual is found in the sentence, and a related feature of whether or not another individual is found in the sentence. Without an accurate way of resolving nominal references, we ignored anaphora.

We found the target's name by the same method used in determining which sets were biographical—by the relative prominence of one capitalized word among all others in the set of articles. For example, "Thomas" represents 0.0168 of all capitalized words in the development set on the nomination of Clarence Thomas to the Supreme Court, while the next most frequent, "Judge" (which, when capitalized, is often used as a title) is 0.007. However, this prominent capitalized noun could refer to a place or organization, and we need another method of distinguishing persons from other proper nouns, i.e. of confirming that we are indeed dealing with a biographical set. The simple solution we use is to count personal pronouns that could refer to a human being (e.g.,"he" and "her", but not "it"). The personal pronoun counts match the most prominent name in the biographical sets, but not in the others.

A final feature measured how prominent an occurrence of the most prominent capitalized noun was for each sentence where it occurred. We estimated this prominence factor as the relative position of the capitalized noun within the sentence.

The additional biographical features were used on only three sets in the evaluation run[2], but there were 10 such biographical sets in the development material. The determination of whether a set is biographical or not is made automatically by the router.

Feature Combination Machine learning of feature weights was not possible given that the training data contained summaries that were considerably rewritten from the document sentences. We determined weights by testing what seemed like reasonable combinations of features, checking to see if our summaries were moving closer to the model summaries or further away. We began with arbitrary weights and begin changing them by hand and then reviewing the summaries produced by hand to see if they captured what seemed to be the thrust of the documents. In essence, we ignored the details of the training summaries, and sought to find what we considered the most important information in the clusters of documents and to ignore irrelevant side issues and descriptions. We found that the summaries did not appear to be improving much after six cycles of this process of adjusting the weights and reviewing the results.

3 Evaluation

In this section we present an early analysis of the evaluation of our summarization system, based on comparisons between system and human-generated summaries, as organized by NIST.

[2]One of these sets, set 31, was not evaluated by the DUC assessors.

We first summarize the evaluation framework, then introduce evaluation measures we use and present evaluation results on these measures. We also analyze the overall results of the DUC competition, ranking different factors that affect the evaluation scores and comparing our system to other competitors.

3.1 Evaluation Background

The DUC multidocument summarization evaluation involved 30 document sets.[3] For each test data set and each target summary size (50, 100, 200, and 400 words), one automatically generated summary was submitted from each participating site, and one gold-standard summary was created by humans. Comparisons for each data set and target summary size involved the human-created summary (*model summary*) versus the summaries automatically produced by competing systems and, in some cases, one additional summary created by humans (*peer summaries*). Although multiple model summaries were available for some document set and summary size combinations, only one comparison was performed for each peer, document set, and summary size. This is due to the human involvement in the comparison process, explained below.

For each data set and target summary size, the author of the model summary assessed the degree of match between that model summary and various peer summaries. First, qualitative measures pertaining to the comparison as a whole were reported on a scale between 0 and 4. These measures were grammaticality, cohesion, and organization.

To calculate quantitative measures of overlap between system- and human-created summaries, the human-created summaries were segmented by hand into *model units* (MUs), which are informational units that should express one self-contained fact in the ideal case. These units might be sentence clauses, however, they are often sentences. Summaries compared to the human-created summaries (both system-generated and summaries created by a different human) were automatically segmented into *peer units* (PUs), which are always sentences. Subsequently, the assessor located the PU(s) that covered the content of each MU, if any, and assigned an estimate of the degree of match, between 1 and 4.[4] Excess units in the peer summary were rated as a whole on whether they should, may, or should not be present in an ideal summary.

3.2 Evaluation Metrics

3.2.1 Evaluating Content: Precision, Recall and Excess Peer Units

From the above judgments we were able to create a variety of evaluation metrics. The first of these are quantitative measures of precision and recall. These summarize the per-PU or per-MU decisions of the evaluators in a single number.

Precision is calculated for each peer summary as the number of PUs matching some MU divided by the number of PUs in the peer summary:

$$\text{Precision} = \frac{\text{Number of distinct PUs matched to one or more MUs}}{\text{Total number of PUs in the peer summary}}$$

This is a conservative estimate of precision; we could increase the number of PUs that are considered correct by using information about the number of PUs not assigned to MUs. However, this is not currently possible since the data on PUs not assigned to MUs is qualitative in nature ("most", "some", etc.) rather than a count. Furthermore, we could more accurately measure precision by using weights reflecting the degree of match between a PU and MU. Again, the data that was collected does not currently allow this. Because a given PU may match multiple MUs, and multiple PUs may be recorded as covering a single MU, we cannot use the information about the degree of match between a MU and set of PUs to appropriately weigh the PUs that are correctly present in the peer summary.

[3] Not all 30 document sets were fully evaluated for each participating system.

[4] Matching grades were at least 1, since otherwise no PUs were reported for that MU.

We can perform more accurate analyses of recall given the data that was collected by NIST. Each MU is either not matched to any PU or covered by one or more PUs, with the collective degree of coverage reported. Therefore, we can set thresholds corresponding to degrees of match (1 to 4), and define a "covered" MU as one which matches PUs at or above the specified threshold. In the extremes, a very lenient recall measure counts as successfully covered MUs even those for which "little" content was covered (threshold of 1), and a very strict recall measure counts as successes only the MUs that were completely covered in the peer summary (threshold of 4). In other words,

$$\text{Recall}_t = \frac{\text{Number of MUs matched at or above } t}{\text{Total number of MUs in the model summary}}$$

One way to combine the four recall measures is to treat the degree of match reported by the assessors as a ratio rather than ordinal value (i.e., assume that a match of 2 is twice as good as a match of 1, and half as good as a match of 4). Under that assumption, we can average the degrees of match over the MUs and report the resulting measure as the *average degree of match*. This measure captures the relative weight of the different matching degrees in a way that the Recall_t measures cannot. It can reveal cases where recall at low thresholds is high but a lot of the matches are at a low level of content overlap.

We have calculated the above six measures for our and other peer systems (including the baselines and the human-created summaries compared against other models). Table 1 shows the macro-averaged values of the recall measures that depend on a threshold choice, while Table 2 shows micro- and macro-average values of the precision and average degree of match measures that do not depend on a threshold. These tables are based on all evaluated summaries; additional tables specific to each target summary size (and micro-averaged results corresponding to Table 1) are available at `www.cs.columbia.edu/~vh/DUC/analysis/`.

In addition to precision and recall, we also evaluated content by averaging two of the scores provided for the excess PUs that were unused during the matching process. We calculate the average rating for excess PUs that should have been included in an ideal summary (high values are better, indicating a system has relatively many PUs that should have been included in the model summary), and the average rating for the PUs that are definitely extraneous (low values are better). We also calculate the difference between these two ratings. These results are shown in Table 3. Additional results specific to target summary sizes are available at `www.cs.columbia.edu/~vh/DUC/analysis/`.

NIST collected a third qualitative variable on excess PUs, asking assessors to collectively rate the extent that the unmarked peer units are neither irrelevant nor ones that should have been included in the summary. We feel that this measure is defined vaguely and does not add much as a gauge of overall summary quality, especially since the assessor does not go back to the original texts when making this determination.

3.2.2 Evaluating Style: Qualitative Measures

Additional qualitative assessments are directly based on the grammaticality, organization, and coherence scores assigned by the evaluators. The averages of these scores across all document sets are shown in Table 4. Additional results per target summary size are available at `www.cs.columbia.edu/~vh/DUC/analysis/`.

3.3 Comparison Between Peers

We have performed comparisons between each pair of peers (baselines, humans, and automated systems) on each of the six recall/precision measures and the six qualitative measures discussed in the previous section. For each performance measure and each target summary size, we compute a table indicating the results of comparing all pairs of peers across all document sets for which both members of the pair were eval-

Peer	Observations	$t = 1$	$t = 2$	$t = 3$	$t = 4$
1	116	18.86	13.34	3.45	1.47
2	114	30.68	21.45	8.01	2.95
A	24	59.06	40.21	13.76	4.94
B	23	57.02	44.47	26.15	9.25
C	24	55.63	48.80	20.29	6.60
D	23	40.20	32.94	9.21	4.69
E	24	61.08	53.21	19.19	7.29
F	24	59.70	55.25	30.32	16.12
G	19	51.23	44.27	15.37	3.88
H	20	29.32	21.80	16.35	7.78
I	24	53.34	46.53	20.93	7.58
K	24	59.01	48.75	17.74	7.41
L	115	33.74 (5)	25.38 (5)	8.92 (3)	4.01 (3)
M	108	27.27 (8)	19.56 (8)	5.42 (9)	2.63 (9)
N	115	37.49 (2)	29.53 (1)	10.86 (2)	4.49 (2)
O	115	26.09 (9)	18.30 (9)	6.59 (7)	3.80 (4)
P	116	33.94 (4)	26.20 (4)	8.78 (4)	3.76 (5)
R	114	27.48 (7)	20.90 (7)	5.92 (8)	3.03 (7)
S	112	30.55 (6)	21.69 (6)	7.22 (6)	3.49 (6)
T	116	35.53 (3)	28.82 (3)	15.03 (1)	7.42 (1)
U	115	21.39 (11)	12.53 (12)	2.62 (12)	1.21 (12)
W	115	20.18 (12)	13.36 (11)	4.97 (11)	2.57 (11)
Y	116	41.58 (1)	29.14 (2)	7.54 (5)	2.61 (10)
Z	116	23.21 (10)	15.60 (10)	5.27 (10)	2.67 (8)

Table 1: Macro-averaged recall scores, dependent on threshold t, for all summary sizes. Relative rankings between automated systems are in parentheses.

uated.[5] The table indicates by the letters "l" or "u" whether the system on the left part of the table (corresponding to the current row) or the system on the upper part of the table (corresponding to the current column) is better.

We base this comparison on the Wilcoxon signed rank statistic, which compares the differences in the observed ranks of the scores of the two systems on different i.i.d. samples to what is expected under the null hypothesis that the median of the difference in performance between the two systems is 0 [Conover 1980]. The different document sets provide independent samples for each system. A normal approximation is used when more than 25 paired samples are available or when ties in the ranks exist [Lehmann 1975]. Because of the correlation expected across different length summaries from the same document set, the analysis cannot be extended to the entire set of summaries produced by each peer.

We note cases where the Wilcoxon signed rank statistic is significant at the 5% level by capitalizing the corresponding "l" or "u" letter in the table. Note that we chose to use a non-parametric statistic, sacrificing some power rather than making additional normality assumptions (e.g., as required for a t-test). The small sample size may also make detection of existing differences between peers harder.

We also mark with a dot (".") the diagonal, with an equal sign ("=") cases where the two

[5]This is generally less than the 30 document sets in the test data, since not all systems were evaluated on all document sets and summary sizes.

Peer	Observations	Average degree of match		Precision	
		Macro	Micro	Macro	Micro
1	116	0.3712	0.3683	32.92	32.85
2	114	0.6309	0.6756	47.42	51.80
A	24	1.1797	1.2930	66.96	70.11
B	23	1.3690	1.4508	57.68	58.04
C	24	1.3132	1.3245	58.70	57.00
D	23	0.8704	0.9567	55.49	59.59
E	24	1.4078	1.3207	55.67	57.29
F	24	1.6140	1.6240	58.93	60.39
G	19	1.1476	1.1838	65.16	66.48
H	20	0.7526	0.8521	47.40	47.03
I	24	1.2839	1.2508	60.65	55.09
K	24	1.3291	1.3586	59.43	55.25
L	115	0.7205 (5)	0.7491 (5)	51.71 (3)	51.13 (4)
M	108	0.5488 (8)	0.6158 (7)	43.39 (7)	45.76 (7)
N	115	0.8237 (2)	0.8625 (2)	58.74 (1)	60.58 (1)
O	115	0.5479 (9)	0.5699 (8)	44.14 (6)	45.79 (6)
P	116	0.7267 (4)	0.7882 (3)	49.56 (4)	52.26 (3)
R	114	0.5732 (7)	0.5437 (10)	42.00 (8)	43.02 (8)
S	112	0.6295 (6)	0.7163 (6)	52.99 (2)	55.44 (2)
T	116	0.8680 (1)	0.8841 (1)	48.96 (5)	50.76 (5)
U	115	0.3774 (12)	0.4407 (11)	23.65 (12)	26.06 (12)
W	115	0.4108 (11)	0.4405 (12)	29.47 (11)	31.33 (11)
Y	116	0.8088 (3)	0.7793 (4)	41.51 (9)	42.58 (9)
Z	116	0.4675 (10)	0.5542 (9)	37.98 (10)	42.54 (10)

Table 2: Average degree of match and precision scores (macro- and micro-averages), for all summary sizes. Relative rankings between automated systems are in parentheses.

peers exhibit equal performance across all common samples on the basis of their signed rank statistic, and with an "x" cases where there are no common document sets where both peers have been evaluated. The latter happens only in comparisons involving two human-constructed summaries.

Tables 5 and 6 show the output of this process for precision on 50 and 100 word summaries respectively, while Table 7 shows the output for the average degree of match on 200 word summaries. All tables are available online at `www.cs.columbia.edu/~vh/DUC/analysis/`.

3.4 Performance of Columbia's Summarizer

In this section, we provide several comparative views of the evaluation data. First, we examine Columbia's performance on summary content; this involves ranking its results on precision, recall and analyzing the assessors' rankings of excess PUs that are missing from or totally unrelated to the model summary. We provide an aggregate view of the different systems, showing for each system the number of systems it ranks above. Next, we examine Columbia's performance on style characteristics, including grammaticality, coherence, and organization. Finally, we also look at an aggregate analysis that ranks each system in comparison to every other sys-

Peer	Observations	Good excess PUs	Bad excess PUs	Difference
1	116	0.2155	0.6034	−0.3879
2	114	0.1053	0.5614	−0.4561
A	24	0.9167	0.0833	0.8333
B	23	0.3913	0.0870	0.3043
C	24	0.4583	0.4167	0.0417
D	23	0.6087	0.0000	0.6087
E	24	0.2500	0.2917	−0.0417
F	24	0.8750	0.0833	0.7917
G	19	0.3684	0.1579	0.2105
H	20	1.0000	0.4000	0.6000
I	24	0.6250	0.0833	0.5417
K	24	0.5417	0.5000	0.0417
L	115	0.3391 (3)	0.2783 (2)	0.0609 (2)
M	108	0.2963 (4)	0.5926 (10)	−0.2963 (9)
N	115	0.2783 (6)	0.2435 (1)	0.0348 (3)
O	115	0.2957 (5)	0.3043 (4)	−0.0087 (4)
P	116	0.1724 (10)	0.3276 (5)	−0.1552 (6)
R	114	0.3772 (2)	0.4123 (7)	−0.0351 (5)
S	112	0.1607 (11)	0.3929 (6)	−0.2321 (8)
T	116	0.3879 (1)	0.3017 (3)	0.0862 (1)
U	115	0.2261 (8)	0.9043 (11)	−0.6783 (11)
W	115	0.1217 (12)	1.2522 (12)	−1.1304 (12)
Y	116	0.2414 (7)	0.5690 (9)	−0.3276 (10)
Z	116	0.2069 (9)	0.4138 (8)	−0.2069 (7)

Table 3: Average ratings for PUs that should be added in the summary, that are definitely extraneous to the summary, and their difference. Relative rankings between automated systems are in parentheses.

tem, counting how many times differences between systems were statistically significant.

Evaluation of Content: Precision, Recall and Ratings of Excess PUs Based on the many analyses, Columbia's system (System L) performs well on summary content when compared to other systems. It is typically ranked third or fourth, with different systems ranked ahead of it for each analysis. For example, there are two systems that had a better precision than Columbia's system (N, S), and two systems that had a better recall at high thresholds (N, T). In general, at the different levels of recall, Columbia's system ranks within the top five. Evaluation of excess PUs shows that Columbia ranks third on producing the greatest number "good" excess PUs (after T,R), second on producing the lowest number of "bad" excess PUs (after N) and second on difference between the two (after T). These analyses create a group of four top systems (N, S, T, L) that consistently do better than others. System P also follows closely our system in many scores, although scoring less than our system in most cases.

From Table 1, we observe that, on average, our system ranks third on recall at high threshold levels (more strict matches) and fifth on recall at low threshold levels (lenient matches). System N appears to be best on recall across different thresholds, while system T also outperforms our system obtaining the highest scores at high

Peer	Observations	Grammaticality	Cohesion	Organization
1	116	3.1810	2.6293	2.8017
2	114	3.2719	1.7193	1.6491
A	24	3.7083	2.2500	3.7500
B	23	3.5217	2.6087	3.1304
C	24	3.6667	2.7500	2.9583
D	23	3.8696	2.8261	2.9565
E	24	3.7917	3.0833	3.1250
F	24	4.0000	2.1667	3.4583
G	19	3.6316	2.5263	2.7895
H	20	3.6500	2.9000	3.0500
I	24	3.7917	3.0000	3.3333
K	24	3.7083	3.2500	3.1667
L	115	3.7217 (2)	1.8435 (8)	1.9130 (9)
M	108	3.5370 (7)	2.1759 (2)	2.3981 (3)
N	115	3.6609 (5)	2.0087 (5)	2.2261 (5)
O	115	3.7913 (1)	2.1565 (4)	2.3217 (4)
P	116	3.6724 (3)	1.9310 (7)	2.1724 (6)
R	114	3.6140 (6)	2.1754 (3)	2.4561 (2)
S	112	3.6696 (4)	1.9375 (6)	2.0536 (7)
T	116	3.5086 (8)	2.3362 (1)	2.6121 (1)
U	115	3.2696 (10)	1.3043 (12)	1.0870 (12)
W	115	3.1217 (11)	1.4609 (11)	1.2522 (11)
Y	116	2.4483 (12)	1.7328 (10)	1.7672 (10)
Z	116	3.2759 (9)	1.8017 (9)	1.9397 (8)

Table 4: Average grammaticality, cohesion, and organization over all summary sizes. Relative rankings between automated systems are in parentheses.

thresholds. Our system has approximately the same scores as system P (generally ranking 4th or 5th). System Y achieves impressive scores on low thresholds, but worse than our system or the other top performers on high thresholds. There are several other groups of similarly performing systems: S follows P and our system; M, O, and R form a group further down; and finally Z outperforms the group of U and W that obtain the worst recall scores. A similar picture is revealed from the average degree of match (weighted recall) measure in Table 2. T obtains the highest scores, followed by N, then Y, then P and our system very close together, then S, then M, O, and R together, then Z, and finally U and W together.

On precision (Table 2), we note that our system takes the third place when using macro-averaging and fourth place when using micro-averaging. System N achieves the best scores, followed by S, our system, P, and T. Further down the precision list we see systems M and O close together, then R and Y also close, then Z, and with a large difference W and finally U.

Most systems have consistent rankings in both the precision and recall dimensions: N is near the top in both categories, our system and P score generally in the top four, M, O, and R obtain middle-level scores, and Z, W, and U offer the worst scores for both types of measures. However, T and Y score much better on recall than precision, while the reverse holds for S.

Content can also be measured based on the perceived quality of PUs produced by a peer but

```
  12ABCDEFGHIKLMNOPRSTUWYZ
1 .uuuUuUuuuUUUuUuUuUUllUu
2 l.uuuuuuuuUuUlUuuuuuLlul
A ll.x=xxllx==ululululll=l
B llx.xxx=l=x=ululll=ullll
C Lllx.=l=xx==lll=lll=llll
D llxxl.l=x==xulu=llllllll
E Llxx==.==lxlllllllll=Llll
F ll=l=ll.xxxxull=ll=llllL
G ll=lxxlx.x=xllullllullll
H llx=xl=xx.==u=ul==uulu==
I LL=x=lxxll.=llllLllLlll
K Lllllx=xxll.lllllllllLLLL
L LLllulululuu.lullllllLLLL
M luuuuuuuu=uuu.uuuuuullul
N LLllululluulluull.llLllLLlL
O lluu==u=uuuuulu.llulLLll
P Llluuuuuu=uuuluu.lulLlll
R lluuuuuuu=UuulUuu.uullul
S Lll=uuu=lluuululll.lLLlL
T Llul=u=ulluuuluuulu.Llll
U uUu=uuUuluUUUuUUUuUU.uUu
W uuuuuuuuuluUUuUUuuUul.uu
Y Ll=luuluu=uUUluuuluuLl.l
Z luuuuuuUu=uUUuUuuuUullu.
```

Table 5: Peer-to-peer comparison on precision, 50 word summaries.

```
  12ABCDEFGHIKLMNOPRSTUWYZ
1 .Uu=llluuuuuUUUuUuUuLlul
2 L.Ulluuuluu=llulll=lLLLL
A lL.llxxl=xlllL=LLllLlLLL
B Ll=.xxx=l=x=u=ulu=lullll
C ll=x.===xxl=ulllullllllll
D llxxl.==x==x=lullLllLLll
E llxxl=.===x=llllLllLLLl
F ll=llll.xxxxllllLulLllL
G llllxxlx.x=x=llllllllLll
H llxlxllxx.=llllllll=llll
I ll=x=lxxll.lllullllLlll
K l==llxlxx==.ll=llllLlll
L Luull=lu==lu.lUlululLLlL
M LuU=ulluuullu.UlululLLLL
N Ll=lllluuul=LL.LlLllLLLL
O luUllluu=uuluuU.uluuLlll
P LuUlluluuuulllul.lulLLLL
R luu=uU=UuuluuuUuu.uuLlll
S L=uullllluuuullulll.lLLLL
T luUllullu=uluuululu.LLlL
U UUlllU=Ulu==UUUUUUUU.uUu
W uU=uu=UuUullUUUuUuUUl.ul
Y lUUlll=l=lluuUUuUuUuLl.l
Z uUUlullUuuulUUUuUuUUluu.
```

Table 6: Peer-to-peer comparison on precision, 100 word summaries.

not included in the model summary. These measures should reward a system that improves upon the model summary in some respect and punish one that produces digressions. Therefore, we look for high scores in the first column of Table 3, low scores in the second column, and a large positive difference between the two. Our system received the third best score on "good" extra PUs, the second best score on "bad" extra PUs, and the second overall score on their difference. Again, it varied which systems performed ahead of Columbia. For "good" extra PUs, systems T and R ranked first and second, for "bad" extra PUs, system N ranked first, and for overall score, system T ranked first. Columbia was one of only three automated systems with a positive difference between good and bad extra PUs, indicating that it suggests a larger number of useful than extraneous sentences among those that are not present in the model summary. The value of this measure for evaluation is further validated from the fact that all except one human have positive differences, and both baselines have negative differences. As expected, there is a strong negative correlation between the ratings on good and bad extra PUs (-0.5325824 among the automated systems).

Evaluation of Style: Qualitative Scores A second comparison between peers is based on the qualitative scores described in Section 3.2.2 that are assigned by the assessors to the entire summary. Columbia's system also ranked well on the grammaticality score, with only one other system (O) performing better (see Table 4). Most systems performed well on the grammaticality score, as most systems performed some form of sentence extraction; system Y is the only one

```
  12ABCDEFGHIKLMNOPRSTUWYZ
1 .U========U=UUUUUUUUuuUu
2 L.Uu=llU=llluLUluluULLul
A LL.l=xx=xx==LLlLLllllLlL
B Ll=.xxx=l=x=lllLlLllLLLL
C LLlx.l==xx==LllLLllLllLL
D Llxx=.=uxllxllllLlLLLll
E Llxxll.=l=xlLLlLlLLLLLLL
F LL=lll=.xxxxLLLLLlLLLLlL
G llx=xx=x.xlxlllllllllll
H llxlx=lxx.=lllllllllllll
I Ll=xl=xx=l.llllLllllLLll
K Lllllx=xx==.LLlllllllLlL
L LlUl=uU==ll=.LuLuluuLLlL
M LUUl=l=U=lu=U.UlUuUUlLUl
N LLllull==ll=lL.LllllLLlL
O LuUUUlUU=l=lUuU.UuUUlLUl
P LlUlUllU=llllLuL.lluLLlL
R Lul=l==l=lllululu.uuLLul
S Lllull=U=luulLuLul.uLLlL
T LLll=U==uullLuLlll.LLlL
U lUu=l==U===lUuUuUUUU.lUu
W lU=Ul=======UUUUUUUUu.Uu
Y Lll=Ul=l=llluLuLuluuLL.L
Z luU==l=U==l=UuUuUuUUllU.
```

Table 7: Peer-to-peer comparison on degree of match (weighted recall), 200 word summaries.

that received an average grammaticality score less than 3. It is notable that human peers received about the same scores on grammaticality as automated systems; our system outperformed six of the ten humans on that measure. Baselines, although following an extraction approach, received lower grammaticality scores than most systems.

On the other hand, we did not perform as well on the cohesion and organization scores. There were 7 systems with better cohesion scores, and 8 with better organization scores. This is partly due to the fact that ordering was not a primary focus during our system design (due to limited time). While MultiGen employs a sophisticated algorithm for ordering information, the MultiGen system was only used on one document set, while the rest were summarized with DEMS, which uses only simple heuristics to order the information.

It is noteworthy that, at least among automated systems, there is an extremely strong correlation between cohesion and organization, indicating that the assessors may not be differentiating between those two measures. At the same time, excluding the four worst performing systems which receive poor scores on all three of grammaticality, cohesion, and organization (U, W, Y, and Z), there is an apparent *negative* correlation between grammaticality and cohesion/organization. So our system receives an excellent score on grammaticality and poor scores on cohesion/organization, while at the other extreme system T ranks eighth on grammaticality and best on cohesion and organization. Table 8 shows the correlations between these three measures for the top eight systems; the overall correlation between cohesion and organization is 0.9910293 when all twelve systems are included.

Aggregating Individual Peer-to-Peer Comparisons In Section 3.3 we described a framework for testing the significance of the difference in performance between two peers, given an evaluation measure and a specific target summary size. We can determine an aggregate ranking by combining information from the multiple tables similar to Tables 5, 6, and 7, by counting how many times a particular peer outperformed another peer, and how many of these differences were statistically significant. This aggregate measure is approximate, since among other things it glosses over the the correlations between summaries of different sizes for the same documents and the differences in the sample size across comparisons. However, the latter is a significant factor only when the comparison involves human-constructed peer summaries (which were evaluated on significantly fewer model summaries than the automated peers). Comparisons between automated peers involve roughly the same number of document sets, and therefore we report separately how many times each automated peer performed better than other automated peers, and how many of those times the difference was statistically significant.

	Grammaticality	Cohesion	Organization
Grammaticality	1.0000	-0.5929	-0.6475
Cohesion	-0.5929	1.0000	0.9713
Organization	-0.6475	0.9713	1.0000

Table 8: Correlations between grammaticality, cohesion, and organization, calculated on the scores of the top eight systems on these measures.

These results are shown in Table 9 for precision, and Table 10 for recall with threshold 3 (i.e., matching with "most" or "all" the content of the model unit covered by one or more peer units).

We observe that under this measure, our system ranks below N and S on precision, with P and T exhibiting similar numbers to ours. On recall, we score behind N and T, with P following our system. These numbers indicate that some systems (including ours, N, and T) exhibit more consistent behavior across different summary sizes, something that the averages of Tables 1 and 2 do not reveal.

Evaluation of Different Summarization Strategies In the testing phase, the breakdown of the sets in our classification scheme was substantially different than what we had in the training phase and the router at the top level of our system found only one single-event set, D04, and three person-centered sets, D13, D24, and D31. Since we did not have the option of reconsidering the parameters, we ran the system as it was. MultiGen handled set D04 and DEMS handled the other three. No results were given for any site's multidocument summaries for the D31 set. On the remaining two person-centered sets and the single-event set, we observed that the behavior of our specialized summarizers relative to the other systems was in most measures and summary sizes somewhat better than what we obtained with the general settings of DEMS (see Table 11). Nevertheless, the small number of cases where the specialized summarization strategies were applied does not allow us to draw conclusions about their performance relative to our general feature-based summarization strategy.

3.5 A Look at the Overall Evaluation Framework: What Affects the Scores

DUC provided for the first time this year a framework for the quantitative evaluation of multidocument summarization systems on a standalone basis, unconnected with specific application tasks. By casting summarization as a retrieval application, it is possible to calculate measures such as precision and recall and compare different summarization approaches. However, the need for humans to construct model summaries, segment them into minimal units, and perform the comparison between summary units, has limited the number of document sets and the number of model summaries per set. This increases, relative to other evaluations using a comparison framework, the possibility that the evaluation will be influenced by factors other than the performance of the competing summarization systems. As we shall show, the human who constructed the model summary and the document set had a larger effect on the outcome score than the peer system. This is a troublesome result, but one that can be addressed in future evaluations.

In this section, we analyze the relative effects of some of these additional factors in the overall scores that all DUC systems received according to each of the precision/recall measures defined in Section 3.2.1. We consider the distribution of the scores in metrics such as "precision" and "recall with threshold 1", and measure the overall variability of the scores, and how much of that variability can be explained when each of the following factors is considered: the document set, the human who constructs the model summary, the size of the target summary, and the peer summarizer. In other words, we perform a

Peer	Over all peers		Over automated systems	
	Comparisons won	Significant comparisons won	Comparisons won	Significant comparisons won
1	25	3	9	3
2	58	15	33	13
A	67	17	42	15
B	51	4	37	1
C	58	3	43	2
D	61	8	43	8
E	63	12	46	11
F	60	6	43	5
G	64	1	45	1
H	55	0	37	0
I	66	10	46	7
K	63	9	45	7
L	60	17	32	12
M	48	10	23	8
N	80	35	44	29
O	44	8	19	7
P	60	12	32	9
R	37	6	15	5
S	69	20	38	16
T	61	12	30	10
U	12	0	0	0
W	18	0	5	0
Y	44	8	15	6
Z	31	5	11	4

Table 9: Counts of times when a peer performed better than another peer on precision, over all document sets and target summary sizes.

traditional analysis of variance (ANOVA) [Hicks 1982], where we consider each peer system evaluation as an observation with four predictor variables (document set, human modeler, summary size, and peer summarizer) and the quantitative performance measure as the response. We fit a linear model with just one of the predictors, and measure the decrease in uncertainty attributable to that predictor as the sum of squares of the error terms (predicted response minus actual response) with and without that predictor. More formally, given a predictor x, the sum of squares for x is

$$SS(x) = \sum_i (R_i - R_{0i})^2 - \sum_i (R_i - R_{xi})^2$$

where R_i is the actual response for observation i, R_{0i} is the prediction for i obtained with only a constant term in the model, and R_{xi} is the linear predictor based on full knowledge of x plus a constant term and fitted to the entire data in an optimal manner (minimizing $\sum_i (R_i - R_{xi})^2$).

We fit each of the four predictors separately to address potential correlations between the predictors. Also, we would normally use a separate predictor to account for effects due to the human performing the comparison, but this effect is included in the human modeler effect, since the modeler and comparer of the summaries was always the same in this year's DUC. We report sig-

Peer	Over all peers		Over automated systems	
	Comparisons won	Significant comparisons won	Comparisons won	Significant comparisons won
1	9	1	7	1
2	38	14	27	12
A	45	7	34	6
B	58	11	45	9
C	53	8	38	6
D	56	0	42	0
E	62	5	47	4
F	67	14	48	12
G	58	0	44	0
H	60	0	43	0
I	58	0	45	0
K	55	7	38	5
L	38	11	28	9
M	18	2	10	1
N	51	22	39	18
O	30	5	22	4
P	36	15	26	13
R	24	3	18	2
S	29	4	21	3
T	63	40	44	34
U	0	0	0	0
W	21	1	12	1
Y	31	7	22	4
Z	17	5	12	4

Table 10: Counts of times when a peer performed better than another peer on recall with threshold 3 (matching "most" or "all" the content of the model unit), over all common model summaries and target summary sizes.

nificance levels for the F-statistic corresponding to each sum of squares, and also normalize the sum of squares for each predictor by the degrees of freedom that each predictor has; the normalized value is a measure that accounts for the fact that some predictors have a larger effect on the overall score because they are modeled in much more detail.

Table 12 shows the detailed results of this analysis over the 1,832 scores obtained on precision for all peers. Since we are primarily interested in the effect of the peers on summary quality, we also tested a predictor that only distinguishes between the three classes of peers: baseline (peer 1 or 2), human (peer A to K), and automated system (peer L to Z). We note that the most distinguishing factor affecting precision is the human who constructs the model summary (normalized sum of squares of 33,448.99), followed by the broad peer class, then by the document set, then by the target summary size, and finally by the particular peer summarizer. In measuring the overall importance of the predictors, the document set ranks first, followed by the human modeler, then the peer (which subsumes the peer class), and finally the sum-

Document set	Strategy	50 words	100 words	200 words	400 words
D04	MultiGen	4	2	2	1
D13	DEMS/Biography	4	2	5	10
D24	DEMS/Biography	2	4	2	11

Table 11: Ranking of our system relative to other automated systems on degree of match (weighted recall) when specialized summarization strategies are used.

Predictor	DF	Sum of Squares	Mean SS	Mean residual SS	P-value
Document set	28	446,458	15,944.91	670.74	$< 10^{-16}$
Human modeler	9	301,041	33,448.99	743.56	$< 10^{-16}$
Summary size	3	39,530	13,176.74	884.17	$1.38 \cdot 10^{-9}$
Peer class	2	50,797	25,398.42	877.53	$< 10^{-16}$
Peer	23	192,937	8,388.58	809.11	$< 10^{-16}$
None (only constant term)	1,831	1,655,800	904.31	–	–

Table 12: Detailed ANOVA results for different predictors on precision, full data collection.

mary size. Differences between particular peers within the same class account for less than half of the performance differences accountable to different document sets (both of these predictors have roughly the same degrees of freedom), and for two-thirds of the performance differences due to the human modelers (despite the fact that the evaluation had more than twice the number of peers than human modelers). Introducing any one factor in the model is always extremely significant statistically.

This picture remains unchanged when we consider the effects of the different factors on precision only in the subset of the data where the peers are all automated systems (1,382 evaluations, Table 13). However, the importance of the document sets and the human model constructor is reduced somewhat in the various measures of recall, where, for the most lenient definition of recall, the peer overtakes the modeler in importance. The peer class is also more significant for the recall measures, where it generally accounts for more than half of the variance that all the peers contribute. A summary of these results is shown in Table 14, listing the sum of square measure for different predictors and response measures, while detailed tables similar to Tables 12 and 13 are available online from `www.cs.columbia.edu/~vh/DUC/analysis/`.

Baselines and Human Peers Looking at the various tables of evaluation scores and comparisons between peers, as well as the additional numbers available at `www.cs.columbia.edu/~vh/DUC/analysis/`, we observe that the first of the two baselines (taking the first n words from the chronologically last document in the document set) is beaten by most systems. In particular, our system outperforms this baseline in all 24 combinations of precision/recall measures and summary sizes, and the difference is statistically significant in 22 of these cases. However, the second baseline (taking the lead sentence from each document in the document set) loses to the best automated peers but outperforms several peers that rank low on the evaluation measures. Our system outperforms that baseline in 17 of the 24 precision/recall and summary size combinations. From this analysis, we can conclude that baseline 2 is a harder baseline and thus, perhaps the more valid baseline to use in future evaluations.

We also note that the human peers (marked "A" to "K" on the tables) generally outperformed all automated systems, particularly on recall where differences of as much as 20 percentage points were observed. However, the differences were less pronounced on precision, and for many measures the top automated systems

Predictor	DF	Sum of Squares	Mean SS	Residual mean SS	P-value
Document set	28	377,134	13,469.07	676.56	$< 10^{-16}$
Human modeler	9	238,958	26,550.90	767.91	$< 10^{-16}$
Summary size	3	33,001	11,000.20	914.02	$8.87 \cdot 10^{-8}$
Peer	11	121,096	11,008.71	855.06	$< 10^{-16}$
None (only constant term)	1,381	1,292,525	935.93	–	–

Table 13: Detailed ANOVA results for different predictors on precision, automated systems only.

Predictor	Precision	Degree of match	Recall_1	Recall_2	Recall_3	Recall_4
Document set	446,458	200.92	299,179	274,971	81,781	23,207
Human modeler	301,041	136.79	187,566	185,005	59,775	17,576
Summary size	39,530	20.64	10,992	16,787	17,646	8,101
Peer class	50,797	82.61	119,523	108,868	28,814	4,099
Peer	192,937	133.79	204,751	179,130	50,889	9,612

Table 14: Sum of squares for different factors on various performance measure, from ANOVAs on the entire evaluation data.

had scores that fell in the range of scores of the human peers. It is worth noting, especially for the planning of future evaluations, that the human peers received scores no better than 60–70%, which indicates that the differences between human modelers and the allowable variation between equally valid multidocument summaries need to be captured better in the evaluation metrics.

4 Conclusion and Thoughts for Future Evaluations

Our analysis shows that on content, there were four systems that consistently outperformed others, namely N, T, S, and L. Our analysis of peer comparisons shows that the difference between the individuals within this group is most often not statistically significant. If we count as conclusive comparisons only those that are statistically significant, Columbia's system (L) takes about half of the times the second position, with one other system (varying according to the performance measure examined) outperforming it, and generally one of the top three positions. On style, Columbia did well on grammaticality (second) but most systems did well and humans were sometimes rated worse. Columbia did not fare as well on cohesion (8th) and organization (9th), primarily because we did not address those issues in the DEMS summarizer which handled most of the input document sets. In contrast, this is a problem we had worked on in some depth for MultiGen, but because only one of the input document sets were on a single event, our research on this topic was not evaluated. Given our focus on summarization of events, we hope to see more document sets on a single event in future evaluations.

In addition to analyzing how well Columbia's system did in the evaluation, we also examined factors affecting the validity of the evaluation framework. In particular, we investigated the factors that most influenced variance in the evaluation results. Our analysis showed that the most distinguishing factor affecting precision is the human who constructs the model summary, followed by the broad peer class (whether baseline, human model or peer summary), then by the document set, then by the target summary size, and finally by the particular peer summarizer. In measuring the overall importance of the predictors, the document set ranks first, followed by the human modeler, then the peer (which subsumes the peer class), and finally the summary size. For recall, the peer system overtakes the

human modeler but is still not the primary factor affecting results.

There are two ways to address the high variability of summaries produced by different users. One is to try to formulate more explicit guidelines about what a summary should contain, hoping to directly reduce the variability between human modelers. The second way is indirect: by constructing and evaluating multiple models per document set, we reduce the effects of the human modeler factor on the overall scores. By the same token, increasing the number of document sets will reduce the importance of particular document choices in the evaluation. These steps require significant investments of additional human time for summary construction and comparisons, and therefore it may be impractical to fully carry them out. A tactic that can be used to complement changes in the scope of the evaluation is to better analyze the characteristics of input documents, and perhaps classify them into groups according to their suitability for summarization. Such an analysis can be based on current DUC results, by observing which document sets tend to produce higher summary scores across the board and trying to characterize their properties. A possibility for future evaluations is to focus and score separately particular tracks of specific document types, reducing the variance of the documents within each class.

Finally, our analysis of score distributions revealed a smaller problem in the definitions of the overall qualitative, per-summary rather than per-unit performance measures. Grammaticality scores were high but not perfect, indicating that the assessors were penalizing systems for factors other than grammatical correctness (we assume that most systems were extracting sentences). More importantly, the assessors made no distinctions between cohesion and organization (correlation higher than 0.99), which suggests that the definitions and instructions on these measures could be improved, or at least, these two scores could be combined into one. A final puzzle that merits further analysis is the apparent negative correlation between grammaticality and cohesion/organization.

Acknowledgments

The work reported here was supported in part by the National Science Foundation under STIMULATE grant IRI-96-18797 and by the Defense Advanced Research Projects Agency under TIDES grant NUU01-00-1-8919. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[Barzilay and Elhadad 1997] Regina Barzilay and Michael Elhadad. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, August 1997. Association for Computational Linguistics.

[Barzilay *et al.* 1999] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 550–557, College Park, Maryland, June 1999. Association for Computational Linguistics.

[Barzilay *et al.* 2001] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Sentence Ordering in Multidocument Summarization. In *Proceedings of the 1st Human Language Technology Conference*, San Diego, California, 2001.

[Conover 1980] W. J. Conover. *Practical Nonparametric Statistics.* Wiley, New York, 2nd edition, 1980.

[Hatzivassiloglou *et al.* 1999] Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages

203–212, College Park, Maryland, June 1999. Association for Computational Linguistics.

[Hatzivassiloglou *et al.* 2001] Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. SIMFINDER: A Flexible Clustering Tool for Summarization. In *NAACL Workshop on Automatic Summarization*, pages 41–49. Association for Computational Linguistics, 2001.

[Hicks 1982] Charles R. Hicks. *Fundamental Concepts in the Design of Experiments.* Holt, Rinehart, and Wilson, New York, 3rd edition, 1982.

[Lehmann 1975] E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks.* Holden and Day, San Francisco, 1975.

[McKeown *et al.* 1999] Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. Towards Multidocument Summarization by Reformulation: Progress and Prospects. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-99)*, pages 453–460, Orlando, Florida, July 1999. American Association for Artificial Intelligence.

[Schiffman *et al.* 2001] Barry Schiffman, Inderjeet Mani, and Kristian J. Concepcion. Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 2001.

[Spärck-Jones 2001] Karen Spärck-Jones. Automatic Language and Information Processing: Rethinking Evaluation. *Natural Language Engineering*, **7**(1):29–46, 2001.

[Späth 1985] Helmuth Späth. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples.* Ellis Horwood, Chichester, West Sussex, England, 1985.