Evaluating Content Selection in Human- or Machine-Generated Summaries: The Pyramid Scoring Method

Rebecca J. Passonneau   Ani Nenkova

September 3, 2003

1 Introduction

From the outset of automated generation of summaries, the difficulty of evaluation has been widely discussed (e.g., [10] [25]). Despite many promising attempts, we believe it remains an unsolved problem. Here we present a method for scoring the content of summaries of any length against a weighted inventory of content units, which we refer to as a pyramid. Our method is derived from empirical analysis of human-generated summaries, and provides an informative metric for human or machine-generated summaries. It is particularly suited for naturalistic summaries of multiple documents. By *naturalistic* summaries, we mean human- or machine-generated summaries that select information from source texts and re-present the information in concise, fluent text.

What differentiates the pyramid method from other approaches is that we take the frequently made observation that there is no best model summary as a foundation, rather than an obstacle. As we discuss later, we think the use of multiple models, while preferable to a single gold standard, can only be an indirect way of approximating what a pyramid addresses directly. Previous work has shown that despite the observed variations in content across human summaries, certain information will be consistently selected from a source text ([3]). Our inventory-based approach assigns differential weights to content units, depending on their cognitive importance as reflected by their frequency within a pool of summaries. Given a pyramid inventory, there can exist multiple configurations of content units that would be assigned the same score. As a consequence, our method predicts multiple, equally good summaries. Our score is calibrated to observed distributions of content units in human summaries, rather than to highly variable human scores (in contrast to the approach taken in [7]), because it is difficult to elicit robust ratings from humans. To construct the pyramid, we abstract from a repository of human-written summaries, although we believe this is only one method for constructing a pyramid. We believe alternative methods, such as eliciting human judgments about the content in the source texts ([22]), might lead to equally useful weighted inventories.

After presenting our method in idealized form, we illustrate its application to three sets of human and machine summaries from DUC 2003. For the purpose of comparison, we contrast our method with the DUC 2003 scoring procedure, which also attempts to identify the underlying content units expressed in summaries. Our ultimate goal is to use the manual scoring presented here as a foundation for developing an automated method, as noted in the conclusion. However, the main purpose of this report is to present our scoring method, to explain how we derived it from analysis of the content of multiple summaries of the same document sets, and to demonstrate that the scores it leads to are meaningful and robust.

In brief, the pyramid scoring method derives from the observation that different humans will summarize the same textual sources by selecting partly matching and partly distinct units of information. We explain here how we identify different expressions of the same Summary Content Units (SCUs) within a set of summaries. We weight SCUs according to their *coverage*, or, how many distinct summaries they appear in. In principle, the pyramid method depends on a vertical rank-ordering of the disjoint coverage sets for a particular document set, placing the largest set (lowest coverage) on the bottom and the smallest (highest coverage) on top, as shown in Figure 5. Each tier in a pyramid corresponds to a weighting factor. Our empirical task is to populate the tiers of a given pyramid with appropriately weighted SCUs. Figure 5, with four tiers, illustrates an empirical fact about the distribution of SCUs: the greater the coverage associated with a tier, the fewer SCUs it will contain. In Figure 4, a Venn diagram is overlaid on a pyramid to show it predicts multiple, equally relevant summaries. Both figures are discussed below. At present, we use a pyramid only to evaluate summary content: note that we treat each tier of SCUs as if it were an unordered set; further, the only constraint on selection is the proportion of highly weighted content units overall. These, of course, are simplifications; in future work, we hope to take into account ordering information and other interdependencies among content units.

Our presentation is organized as follows. In Section 2, we briefly discuss related work in order to motivate the design principles of our method; in particular, the need to capture the variation seen in human summaries and the advantages of creating a weighted inventory of information, rather than relying on a model summary, as in the DUC 2003 scoring method. In Section 3, we present the formula with a limited number of weighting factors, then in Section 4 we illustrate how we create an inventory of Summarization Content Units (SCUs) to populate the tiers of a pyramid for scoring summaries of a particular document set. Section 5 gives a brief overview of the DUC 2003 scoring procedure. In Section 6 we apply the pyramid method to the human-generated summaries from DUC 2003, and in Section 7 we apply it to the machine-generated ones. In both sections, we compare the scores we assign to the DUC 2003 summary dataset with the DUC scoring method for both human- and machine-generated summaries to argue that our method is more informative, more reliable, and more robust. In Section 8, we discuss limitations and obstacles and our current efforts to address them.

2 Related Work

Our main focus is on previous work on evaluation of naturalistic summaries, but we also discuss evaluation of summaries produced by sentence extraction in order to cover a variety of techniques. We argue that a prerequisite for a consistent, informative, and robust metric is a better understanding of the similarities and differences across human summaries. In particular we contrast pattern-matching evaluation methods, e.g., those that measure ngram overlap of automated summaries with human-generated text (e.g., one or more model summaries), with evaluation methods similar to the approach used in past and current DUC efforts, based on human identification of overlap between abstract content units (Model Units, or MUs; Elementary Discourse Units, or EDUs). Because our method depends on creating a similarly abstract representation of content units, we also discuss previous work on annotating such units with respect to inter-annotator reliability.

2.1 Variation in Human Summaries and Summary Ratings

A major issue in the field of automatic summarization is the observation that when different humans produce summaries for the same document, each individual choses different information to include in a summary. Research as early as [23] reports that extracts selected by six different human judges for 10 articles from Scientific American had only 8% overlap on average (the 6 subjects agreed upon an average of only 1.6 sentences per article out of 20 sentences selected for each article). Also, judges agreed with their own previous judgments only 55% of the time. In their experiments measuring overlap was straightforward, and also very conservative, since the entire sentence had to be an exact match.

The lack of overlap in the content of human summaries has been discussed on numerous occasions (cf. [9], [24], [8], [7], [3]). Halteren and Teufel [3], for example, collected 50 100-word summaries of the same 600-word text, and report that no consensus emerges for the ideal content of a 100 word summary, nor do the summaries fall into distinct homogenous groups. (We discuss their work in more detail in Section 3.1.) Jing et al. [5] report that for single document summarization good agreement can be achieved when the compression rate is high, thus indicating that humans can consistantly point out *the most important information*, but disagree on the ranking of less important fragments. This conjecture was sustained by the fact that agreement deteriorated when the same humans had to produce longer summaries.

It has often been suggested that any automatically assigned evaluation metric must be calibrated to a human evaluation [8] [7]. This approach has been successful for evaluation of machine translation, notably with the Bleu [14] approach. However, it presupposes eliciting a reliable human evaluation, which has proved problematic for summarization. For example, it has been observed that human ratings of summaries are not very stable [7] [12], and we see this in the current DUC 2003 results, as reported below (Section 6). Human ratings of how important a given sentence is to a summary are sometimes inconsistent for the same rater at different times [7] [19], although this might depend on the specifics of the task, given that results on a similar task appear more stable in [21] [22].

For DUC 2003, a procedure was developed for eliciting human judgments about the degree of semantic overlap between a so-called *peer* summary and a model summary (see Section 5). Judges compare automatically identified Elementary Discourse Units (EDUs) in the model against sentences in the peer; the comprehensive score for the summary is an average of the individual judgments for a given summary. As we illustrate in Section 6, the DUC 2003 results are problematic in that different humans assign different scores to the same summary, and summaries from the same human are judged to be different in quality.

The reported variations across human summaries, and in human evaluation of summaries, would seem to present a confusing, and ultimately discouraging, view of the feasibility of automated evaluation. We counter this view by attempting to present a more coherent picture of the human summarization process, and by demonstrating that given an appropriate annotation procedure for content units, and an appropriate scoring method, humans can apply a DUC-like evaluation of content that is both robust and meaningful.

2.2 Ngram Evaluation Applied to Summarization

There have been several attempts to apply the Bleu method [14] developed for machine translation to summarization. In essence, the Bleu method counts ngram overlap of a machine translation with a repository of model translations, or with a translation of a single text, if it is sufficiently large. The scoring in [14] applies a modified recall metric and a brevity penalty; most importantly, its consistency with human evaluation has been demonstrated. Apparently, the idea motivating the adaptation of this metric is that where translation maps from one human language to another, summarization can also be viewed as a translation of sorts: a more expansive text is *translated* (converted) to a less expansive one. For translation, the limiting condition is that there are many possible *good* translations of the same source language phrases. However, the goal of a good translation is relatively well defined: to convey in the target language L_T a close approximation to what was conveyed in the source language L_S. In summarization, what counts as a good summary is not as well defined; there are many more degrees of freedom with respect to selection of what to say, how much to say, in what order, and to what rhetorical purpose.

In [19], the Bleu method is applied to the sentence extraction type of summarization. The authors note that there is low human agreement on ranking sentences to include in an extract. They conclude from a series of experiments that the Bleu method leads to consistent rankings of four systems only when multiple reference summaries are used. A reference summary is constructed from assessments of three judgments of sentences on a relevance scale of 1-10 (the same data as in [22]). We find the results inconclusive, and difficult to generalize to naturalistic summaries. Lin and Hovy [7], who also apply a pattern matching method analogous to Bleu, earlier reached the same conclusion about the need for multiple reference summaries. In contrast to [19], they deal with naturalistic summarization, and explicitly discuss the problematic issue of lack of human agreement both across humans, and for the same human, on the degree of semantic overlap of a summary sentence with a model one. Their solution is to compare the evaluation matrices produced by human and automated evaluation; that is, they apparently assume that the variation in scores assigned by humans is a valid target, rather than an artifact of a poor elicitation procedure or false assumptions about what humans understand the evaluation task to be. They get a Spearman correlation coefficient of .70 using three reference summaries on a multi-document task.

2.3 Annotation of Abstract Content Units

Previous work suggests that it is possible for humans to reliably annotate semantic units in text. For example, Halteren and Teufel [3] claim that factoid identification is *more objective* than DUC-style judgments. Annotators proceed by identifying similarities and differences across texts, presumably a more objective judgment process than assigning a percentage to the degree of information overlap between an EDU and a sentence. They report 96% recall and 97% precision of an individual annotator's representation, using a consensus annotation as the standard. However, note that high recall and precision do not necessarily correspond to high interrater reliability [16].

In preparing materials for assessing how the rhetorical and argumentative characteristics of a text correlate with student readers' comprehension, Beck et al. [2] apply a narrative analysis procedure originally developed by Omanson [13]. Their goal is to compare students' recall of original versus revised passages, which means they need to quantify the overlap in content. Omanson's original procedure identified content units at the clause level, but when applied in [2], the authors found a need to create units *often smaller than a clause* (note that we do as well in our SCU annotation). They report an interrater reliability on a large sample of text coded by two raters of .92.

The high interrater reliability found by Beck et al. [2] and Halteren and Teufel [3] led us to believe that a reliable procedure for coding SCUs was a reasonable goal, despite the apparent lack of agreement on coding overlap of content found in [7] and in the DUC 2003 data. It should be noted that the narrative analysis method in [2] was developed and reported on over several years. For other types of semantic annotation of discourse (e.g., dialogue), it has also been shown that iteration over a set of instructions with explicit procedures can yield good interrater reliability [4] [11].

3 Pyramid Scoring Method

3.1 Background

The key observation that motivates the pyramid scoring method is that almost all of the information contained in a human-written summary is contained in the source texts,[1] though expressed in different words, and that while no two humans will write the same summary, the information in the source text can be prioritized in terms of how likely it is to appear in any one person's summary. Our goal is to develop a general method to abstract and prioritize units of information from source texts in a manner that replicates human behavior. This will enhance our understanding of the human summarization process, and therefore lead to better metrics for evaluating automatically generated naturalistic summaries.

As noted in the previous section, Halteren and Teufel [3] discuss their collection of a large set of 100-word summaries of the same 600 word document for which they analyzed the distribution of distinct content units they refer to as *factoids*. Their Figure 1 (adapted here as Figure 1; each point on the X axis represents the mean whereas their original figure indicates the range) plots the growth in number of distinct factoids (Y axis) against the number of summaries, ranging from 1 to 40. They note that the number of distinct factoids in a single summary ranges from 32 to 55, and that this number increases (apparently logarithmically) with the number of summaries examined. They conclude that *there is a difference in the perceived importance between the various factoids*, reflected in the likelihood that a factoid occurs in multiple summaries. We strongly concur, and believe that an evaluation metric for content selection should therefore assign a quantitative value to *perceived importance,* with frequency across summaries being **one (indirect) source** of evidence. (In Section 6.2, we discuss alternative sources of evidence.) Here we use the DUC 2003 summaries in which four distinct human summaries were collected for thirty document sets to illustrate the pyramid metric. Although this is relatively few summaries per document set, it affords the opportunity to apply our SCU annotation and scoring to distinct document sets (in contrast to [3] which looks at single document summarization for a single document).

While the DUC summaries, like the Halteren and Teufel summaries, are 100 words in length, they contrast with respect to compression rate and complexity. The DUC 2003 task involved summarization from multiple source texts (N=10; avg. document length = roughly 500 words, or 5,000 words per set). In the DUC summaries, the degree of compression is thus far greater. Because the source text consists of multiple documents, the selection task is more complex. It requires the additional semantic task of integrating information across documents, which presumably requires more inferencing (e.g., regarding sources of evidence), and

[1] Halteren and Teufel [3] mention that rarely, a summarizer includes material based on unjustified inference; in unpublished work on this topic, Richard Gerrig observed that with semantically loaded material, summarizers are more or less likely to include unsupported material.

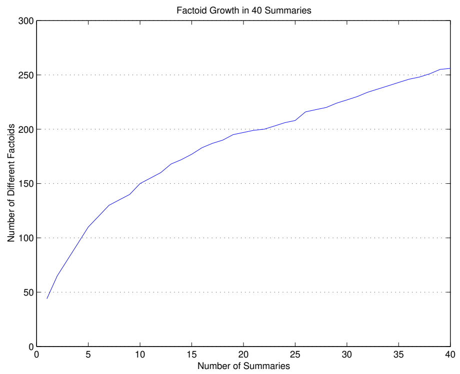

Figure 1: Halteren and Teufel's Figure 1 (slightly modified). The x axis gives the average number of summaries, the y axis is the number of distinct factoids in that number of summaries.

the additional pragmatic task of relating rhetorical structure across documents.

In lieu of *factoid* we use the term Summarization Content Unit (SCU; cf. Section 2). In part, this is to avoid the danger that our interpretation of the notion of factoid in [3] is imperfect, but more significantly, it is motivated by our belief that it is impossible to arrive at a single, best, complete and consistent representation of a text in a formal semantics. For example, selecting the formal predicates to represent natural language lexical items is necessarily somewhat arbitrary, and inextricably related to the kinds of inferences the formal language is intended to support (cf. [1] [18]). As noted in [3], although factoids are an approximation of a first order predicate logic style of semantics, the semantics of a single factoid is motivated not by the inherent atomicity of the meaning, but by the appearance of essentially the same content across multiple summaries. They note that if two or more atomic propositions (eg., F1 and F2) are always expressed together (e.g., in a single phrase), they treat them as a single factoid.

In order to address content selection, we must define what we mean by a Summarization Content Unit, or SCU. At present, we define it indirectly in terms of what we see in human summaries. In part, this is because we doubt there can be an absolute definition that specifies what level of granularity of concepts an annotator should use. Rather than attempting to provide a semantic or functional characterisation of what an SCU is, our annotation procedure defines how to compare summaries to locate the same or different SCUs. On the one hand, we believe restricting annotation to such comparison leads to more reliable results [6]; on the other hand, we also believe that natural language semantics is flexible, and we hope to develop a dynamic rather than static representation of SCUs (similar in spirit to the notion of generativity in [20]).

Figure 2 presents one set of sample summaries from the DUC 2003 data set. The topic pertains to a financial crisis experienced by Philippine Airlines (PAL). The sentences in the summaries have been numbered consecutively. As

Summary A

A.1 Philippine Airlines (PAL) experienced a crisis in 1998.

A.2 **Unable to make payments on a $2.1 billion debt,** it was faced by a pilot's strike in June and the region's currency problems which reduced passenger numbers and inflated costs.

A.3 On September 23 PAL shut down after the ground crew union turned down a settlement which it accepted two weeks later.

A.4 PAL resumed domestic flights on October 7 and international flights on October 26.

A.5 Resolution of the basic financial problems was elusive, however, and as of December 18 PAL was still $2.2 billion in debt and losing close to $1 million a day.

Summary H

H.1 Starting in May 1998, Philippine Airlines (PAL) laid off 5000 of its 13,000 workers.

H.2 A 3-week pilots' strike in June and a currency crisis that reduced passenger numbers **made payments on PAL's $2 billion debt impossible.**

H.3 President Estrada brokered an agreement to suspend collective bargaining for 10 years in exchange for 20% of PAL stock and union seats on its board.

H.4 The large ground crew union initially voted no.

H.5 After PAL shut down operations for 13 days starting Sept. 23rd, leaving much of the country without air service and foreign carriers flying some domestic routes, 61% voted yes.

H.6 Unions agreed to some employee cuts with separation benefits.

Summary I

I.1 Philippines Airlines (PAL), Asia's oldest airline, devastated in 1998 by pilot and ground worker strikes and **with a rising $2.1 billion debt,** stopped all operations for 13 days.

I.2 A 10-year, no-strike agreement was finally ratified when the government began using foreign carriers for domestic flights.

I.3 PAL resumed domestic flights on October 7 and international flights slowly over the next weeks.

I.4 This shutdown had been another blow to the debt ridden national flag carrier.

I.5 Following the strike, efforts to revive the airlines met more obstacles.

I.6 Cathay Pacific Airways would only help if the payroll, including 200 pilots hired during the pilot strike, were slashed.

Summary J

J.1 The fate of Asia's oldest airline, PAL, is uncertain.

J.2 Negotiations with Cathay Pacific Airways to infuse $100 million dollars into the company collapsed when PAL Chairman Tan refused to agree to major job cuts and to relinquish management control to Cathay.

J.3 **PAL is buried under a $2.2 billion dollar debt it cannot repay** and $1 million a day losses.

J.4 A replacement investor is unlikely and President Estrada, though supported by Tan in the election, has rejected a government bailout.

J.5 PAL's financial troubles were exacerbated by a two-week shutdown in September due to a dispute with its largest union.

Figure 2: Phrases contributing to an SCU with Maximal Coverage (i.e., all 4 summaries; boldface)

A.2	**Unable to make payments** on a $2.1 billion debt,
H.2	**made payments** on PAL's $2 billion debt **impossible**
I.1	with a rising $2.1 billion debt,
J.3	PAL is buried under a $2.2 billion dollar debt **it cannot repay**

SCU 1:	Coverage=4	PAL has a debt of over $2 billion
SCU 2:	Coverage=3	PAL cannot make its debt payments

Figure 3: Candidate for one SCU with coverage 4 split into two with coverages 4 and 3

shown by the boldface lines in the figure, all four summaries mention that the airline is over 2 billion dollars in debt (A.2, H.2, I.1 and J.3); in some cases it is mentioned in the first sentence, in other cases it is mentioned later. As noted in the introduction, our scoring method does not deal with ordering issues, which are undoubtedly important; here we address content selection only.

The four boldface phrases in Figure 2 illustrate a candidate SCU, namely the PAL debt. However, note that three of the four summaries mention another factor in connection with the debt: PAL's inability to repay it. Thus what first appears to be a single SCU with maximal coverage (4 summaries, cf. Section 1), becomes two SCUs. Figure 3 illustrates the two SCUs, and gives the coverage. Although we have not yet investigated the interrelation of the semantics of SCUs and their coverage, it is probably not an accident that the SCU in Figure 3 with the lower coverage is semantically dependent on (presupposes) the higher coverage SCU: that is, being unable to repay a debt (SCU 2) implies being in debt (SCU 1).

In the next two subsections, we present the general formula for scoring summaries, then we illustrate a specific SCU annotation for one of the DUC sets of human-generated summaries.

3.2 Formula

In our preliminary SCU annotations of DUC summaries, a 100 word summary consists on average of 12 SCUs. In the process of refining our annotation method, we find we have moved towards a slightly higher number of SCUs, thus the example pyramid included below in Figure 6 has 35 and the associated summaries have 15 on average. In Halteren and Teufel's larger set of 40 summaries, they find a range of 32 to 55 factoids per summary, and a total of 256 factoids overall. In considering the different ranges of factoids versus SCUs, it is important to note that their number depends very much on the number of summaries and what they contain. Any new summary can potentially yield a candidate SCU having only partial identity with one already in the inventory; this could result in the splitting of the original SCU into two new units, as illustrated in Figure 3 above.

We believe that in principle, we could directly derive SCUs from a set of

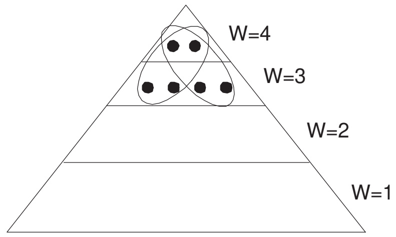

Figure 4: Two of six ideally informative summaries of size 4

source texts, given a better understanding of how human readers and writers select, prioritize, and rhetorically present linguistic meaning. Let us assume for the sake of argument that we had a principled method of doing so that yielded SCUs at a rate similar to our annotation, which we discuss in detail in Section 4, say 30 SCUs per 100 words. At this rate, a set of ten source texts consisting of 500 words each, as in the DUC 2003 sets, would yield 1500 SCUs. With sufficient data, such as a large pool of summaries (as in [3]), or relevance judgments on each sentence (as in [21]), a pyramid would have many more tiers than the four illustrated here.

It is an empirical question what the ideal pyramid size might be for a given set of texts, assuming a very large number of summaries. As we discuss later, we believe that the ranking of SCUs reflected in our pyramids is only indirectly a function of frequency within a summary pool, and that there are other methods to determine how to prioritize the SCUs from a source text set. However, given that we have four human summaries per document set from the DUC 2003 effort, we illustrate our method using a 4-level pyramid, and present the equation for computing scores based on this.

In essence, we score a summary by computing a ratio $\mathcal{P}$ of the observed distribution of weighted SCUs found in the summary ($\mathcal{D}$), divided by the ideal distribution of weighted SCUs, or maximum possible score (Max).

$$\mathcal{P} = \frac{D}{Max} \tag{1}$$

The idea behind this ratio is that a given summary will have an observed number of SCUs, which is its size X, and the greater the proportion of *maximally weighted* SCUs in X, the closer $\mathcal{P}$ will be to one. There can be multiple ideal summaries for a given SCU size, as shown in Figure 4. Here we illustrate how Max and $\mathcal{D}$ are computed.

A pyramid has n tiers T consisting of disjoint sets of differentially weighted SCUs. Each tier T_i, where i ranges from 1 to n, has a cardinality $|T_i|$ and a weight w_{T_i} (corresponding to its observed coverage). If we index the tiers bottom up, then in a pyramid with four tiers (as we will have for the DUC sets we use here), the top tier is T_4. Each tier provides a weighting factor equal to its height, thus the highest T_4 has a weight $w_{T_4} = 4$, and so on down to T_1 with $w_{T_1} = 1$. In an optimal summary, the SCUs are distributed from the top of the pyramid down until reaching the total number of SCUs to be expressed. Thus the maximum possible score (Max) for a summary of size X requires that no SCUs from a given tier T_j can be selected until all the SCUs from the next higher tier (T_{j+1}) have been exhausted. We multiply the number of SCUs from a given tier by the weight. Thus, the maximum score for a summary whose size X is equal to the size of the pyramid would have the ideal distribution shown in (2).

$$\text{Max}_{\text{ideal}} = \sum_{i=1}^{n} w_{T_i} \times |T_i| \tag{2}$$

Where X is not equal to the size of the pyramid, the value of Max depends on calculating for each tier T_i how many SCUs the summary should contain, based on the preference for more highly weighted SCUs:

$$\text{Max} = \sum_{i=j+1}^{n} w_{T_i} \times |T_i| + w_{T_j} \times (X - \sum_{i=j+1}^{n} |T_i|)$$
$$\text{where } j = \max_i (\sum_{t=i}^{n} |T_i| \geq X) \tag{3}$$

In the equation above, j is equal to the index of the lowest tier a perfectly informative summary will draw units from. This tier is the first one top down such that the sum of its cardinality and the cardinalities of tiers above it is greater than or equal to the summary size. If X is less than the cardinality of the most highly weighted tier, then Max is simply $X \times w_{T_n}$ (the product of X and the highest weighting factor). If X is less than the total number of SCUs in the top two tiers, then Max is the sum of $w_{T_n} \times |T_n|$ and $w_{T_{n-1}} \times (X - |T_n|)$. If X is greater than the total number of SCUs in the pyramid, then there are various possibilities for weighting the SCUS that appear in the summary and not the pyramid; we have not yet dealt with this situation.

Now that we see how Max is computed, we can compare the ideal distribution predicted for a given summary with its observed distribution, D. We determine the actual distribution of X into disjoint sets by finding the intersection D_i of the summary SCUs with each tier T_i and computing the cardinality.

$$\mathcal{D} = \sum_{i=1}^{n} w_{T_i} \times D_i \tag{4}$$

Then the pyramid score $\mathcal{P}$ is the ratio of D to Max:

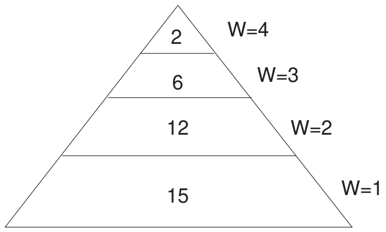

Figure 5: Distribution of SCUs in the pyramid inventory

$$\mathcal{P} = \frac{\sum_{i=1}^{n} w_{T_i} \times D_i}{\sum_{i=j+1}^{n} w_{T_i} \times |T_i| + w_{T_j} \times (X - \sum_{i=j+1}^{n} |T_i|)} \tag{5}$$

An ideal pyramid would represent all information in the source text, much of which would have a very low weight (potentially zero). The more SCUs drawn from a pyramid, the closer the new text is to a revision than to a summary. We define the pyramid in this manner in part because we don't believe there is a single ideal size for a summary, and in part because we believe the line between a long summary and a revision is blurred. For a summary whose size fits the pyramid, the score $\mathcal{P}$ carries an implicit penalty for irrelevant information; SCUs missing from the pyramid have weight *zero*.[2]

Because $\mathcal{P}$ compares the actual distribution of SCUs to an empirically determined ideal, it provides a more informative measure of whether the content of a summary reflects the way a sample of readers would prioritize the information in the text, as reflected in human summaries. We refer to the pyramid as an idealization because we believe that what humans put into a summary is an indirect reflection of what they understand to be the relative importance of the content of the source texts.

4 SCU Annotation

Figure 6 illustrates our consensus SCU annotation for the summaries in Figure 2. As illustrated, a few SCUs appear in all 4 summaries (with weight $w_{T_4} = 4$), a larger number appear in 3 ($w_{T_3} = 3$), and so on. The size of each tier in the pyramid corresponding to this annotation is illustrated in Figure 5. Scores assigned to specific summaries using this pyramid are deferred to Section 6.

One issue in developing an evaluation based on SCUs (or similar semantic units) is reliability [6]. Is the identification of SCUs

[2] One can imagine adding a negative weight for incorrect information, but although humans occasionally do this, it rarely occurs in current automated methods.

$\mathbf{w_{T_i}} = 4$ ($N = 2$: 12 PAs)
SCU-1: PAL has \$2.1 billion debt (2 w_{T_4}: 6 PAs)
SCU-2: PAL enforced a shutdown (2 w_{T_4}: 6 PAs)

$\mathbf{w_{T_3}} = 3$ (N=6: 18 PAs)
SCU-3: PAL in crisis (1 w_{T_4}/1 w_{T_3}: 3 PAs)
SCU-4: PAL unable to repay debt (2 w_{T_3}: 3 PAs)
SCU-5: PAL experienced pilots' strike (2 w_{T_3}: 3 PAs)
SCU-6: this PAL crisis occurred in 1988 (1 w_{T_3}/1 unsplit: 3 PAs)
SCU-7: shutdown began in September (1 w_{T_3}/1 unsplit: 3 PAs)
SCU-8: shutdown lasted two weeks (2 unsplit: 3 PAs)

$\mathbf{w_{T_2}} = 2$ (N=12: 20 PAs)
SCU-9: to compensate for shutdown, foreign carriers flew domestic routes (2 w_{T_2}: 2 PAs)
SCU-10: PAL resumed domestic flights 10/07, international 2 weeks later (2 w_{T_2}: 2 PAs)
SCU-11: the pilot strikes were in June (2 w_{T_2}: 2 PAs)
SCU-12: region experienced a currency crisis (2 w_{T_2}: 2 PAs)
SCU-13: currency crisis resulted in reduced passenger numbers (2 w_{T_2}: 2 PAs)
SCU-14: PAL is oldest Asian airline (1 unsplit/1 w_{T_2}: 2 PAs)
SCU-15: PAL has losses of \$1million/day (2 w_{T_2}: 2 PAs)
SCU-16: PAL negotiated with Cathay for help (conflict: 1 PA)
SCU-17: 10-year no-strike agreement struck (2 w_{T_2}: 2 PAs)
SCU-18: ground crew union first rejected settlement (1 w_{T_3}/1w_{T_2} : 2 PAs)
SCU-19: agreement accepted (1 w_{T_3}/1 w_{T_1}, 1 conflict: 0 PAs)
SCU-20: in aftermath, PAL finances still shaky (1 w_{T_2}/1 w_{T_1}: 1 PA)

$\mathbf{w_{T_1}} = 1$ (N=15: 15 PAs)
SCU-21: [the strikes] inflated costs. (2 w_{T_1}: 1 PA)
SCU-22: as of May 1998, Philippine Airlines (PAL) laid off 5000 (2 w_{T_1}: 1 PA)
SCU-23: 3-week [pilots' strike in June] (2 w_{T_1}: 1 PA)
SCU-24: in exchange for 20% of PAL stock and union seats on its board (2 w_{T_1}: 1 PA)
SCU-25: much of the country without air service (2 w_{T_1}: 1 PA)
SCU-26: [there were] ground worker [strikes] (2 w_{T_1}: 1 PA)
SCU-27: 200 pilots hired during the pilot strike, (1 w_{T_1}/1 unsplit: 1 PA)
SCU-28: as of December 18 PAL was still \$2.2 billion in debt (1 w_{T_1}/1 unsplit: 1 PA)
SCU-29: PAL asked CATHAY for \$100 million dollars (2 w_{T_1}: 1 PA)
SCU-30: negotiations re: relinquishing management control to Cathay (2 w_{T_1}: 1 PA)
SCU-31: A replacement investor is unlikely (2 w_{T_1}: 1 PA)
SCU-32: Tan supported Estrada in the election (1 w_{T_1}/1 unsplit: 1 PA)
SCU-33: President Estrada has rejected a government bailout (1O/1 unsplit: 1 PA)
SCU-34: PAL's financial troubles were exacerbated (1 w_{T_1}/1 unsplit: 1 PA)
SCU-35: PAL Chairman Tan refused to agree to major job cuts (1 w_{T_1}/1 unsplit: 1 PA)

Figure 6: SCU Pyramid Generated from the Summaries in Figure 2

- stable: will the same coder find the same SCUs upon recoding the data at a later time?
- reproducible: will distinct coders find the same SCUs?
- accurate: is there a *correct* SCU analysis that any specific coding closely approximates?

Though we have not yet completed a full-scale reliability analysis, we are confident that the method is stable and reproducible in principle, given our procedure for annotating SCUs, our preliminary results, and claims made elsewhere that such coding can be done consistently by two or more researchers (cf. Section 2.3). As noted above, we do not believe correctness applies.

To arrive at the consensus SCU annotation illustrated in Figure 6, the two co-authors created independent SCU annotations, using a provisional set of instructions. We achieved what we believe was rather good consistency, and we expect this to improve as we refine the instructions. Our separate annotations contained 33 versus 37 SCUs, and we created a consensus annotation consisting of 35 SCUs. The symbols next to each consensus SCU represent a comparison of the two original annotations:

- 2 w_{T_i}: both annotators agreed on the SCU, and on its weight ($N = 21$; number M of all pairwise agreements=45)
- 1 w_{T_i}/1 w_{T_j}: both annotators agreed on the SCU, but differed on its weight ($N = 4$); typically, the delta was one ($N = 3$)
- 1 unsplit: both annotators agreed on the weight of an SCU, and partially on the SCU, but one found an additional SCU (N=10)
- 2 unsplit: both annotators agreed on the weight of a partial SCU, but found a single SCU instead of two; the new SCU would necessarily be assigned a lower weight (N=1)
- conflict: annotators differed on the semantics and membership of an SCU (N=2)

To summarize our comparison of the two original annotations, we had roughly the same number of SCUs, and in most cases assigned them the same weight. We had perfect agreement on 23 SCUs: 2 of 2 where $w_{T_4} = 4$, 2 of 6 where $w_{T_3} = 3$, 8 of 12 where $w_{T_2} = 2$, 9 of 15 where $w_{T_1} = 1$. Most disagreements were not conflicts on what counted as an SCU, but on whether to split an SCU.[3] Other annotation differences arose primarily from whether to create a singleton. That is, almost all the differences between the two annotators were due to how

[3] For example, SCU-6 and SCU-7 differences were due to one annotator failing to split the year and month from SCU-2; SCU-8 was not represented in the original annotations, but was discovered during the consensus review and similarly resulted from a failure to separate out the duration information mentioned in connection to SCU-2.

	A	H	C	J
consensus SCUs	.95	.89	.85	.76
annot1 SCUs	.97	.87	.83	.82
annot2 SCUs	.94	.87	.84	.74

Table 1: Scores for the human summaries based on the two seperate annotations versus the consensus annotation. No significant difference in scores is observed.

inclusive to make an SCU. There were only 2 conflicts, i.e., cases where two phrases were grouped together by one annotator, and each phrase was assigned to a different SCU by the other annotator.

In sum, we find complete consistency on SCU identification and weighting for 45 pairwise agreements (PAs), with 20 additional agreements where the annotators disagreed on weight (typically by one). Perfect agreement on the consensus model would involve 69 pairwise agreements, compared to the observed 65. We are currently considering how best to quantify the comparison of distinct SCU annotations.

One of the strongest arguments we can offer for the *reliability* of SCU annotation is that the two original annotations are similar enough not to affect the scores significantly. Table 1 gives three sets of pyramid scores for the PAL summaries illustrated in Figure 2. The first row gives the scores using the consensus annotation from Figure 6, and the next two rows give the scores for the original annotations. In the next section, we explain how to interpret the scores; here we note simply that there is no significant difference in the scores assigned across the three SCU annotations (between subjects ANOVA=0.11, p=0.90).

5 DUC Scoring Method

Within the Document Understanding Conference, different aspects of summarization have been studied: the generation of abstracts and extracts of different length varying between 50 and 400 words, single- and multi-document summaries, very short summaries and summaries focused by a topic or oriented by opinion. The evaluation of summaries is based on the comparison of a summary (machine-generated, produced by a human or a baseline) by comparing its content to a gold standard summary produced by a human, and called a model. Over the years, different numbers of models were produced by NIST to be used during the evaluation. In 2001, there were multiple models for some document sets and some summary lengths used to study the different factors that influence a summary score. In 2002, there were two abstracts and two extracts produced for each document set. In 2003 only generic summaries of length 100 words were produced; NIST provided four human summaries, any one of which could serve as the model. Thus in general, a model summary is simply a summary produced by a human with no attempt to choose the *best* model among the human summaries available. In 2003, a partial experiment was performed on summaries focused by opinion to see if varying the model summary produces different scores. The results for each system turned out to be very close, regardless of which human summary was used as a model. No attempt was ever made to use multiple model summaries for evaluation, but the multiple human summaries for the 2003 docsets made our study possible.

The procedure used for evaluating summaries in DUC is the following:

1. A human subject reads the entire input set and creates a 100 word summary for it, called a model.
2. The model summary is split into content units, roughly equal to clauses or elementary discourse units (EDUs). This step is performed automatically using a tool for EDU annotation developed at ISI[4].
3. The summary to be evaluated (called a peer summary) is automatically split into sentences. (Thus the content units of the model and the summary to be evaluated are of different granularity—EDUs for the model, and sentences for the peer).
4. Then a human judge evaluates the peer summary against the model. For each content unit in the model:
   - **(a)** Find all peer units that express at least some facts from the model unit and mark them.
   - **(b)** After all such peer units are marked, think about the whole set of marked peer units and answer the question:
   - **(c)** "The marked peer units, taken together, express about k% of the meaning expressed by the current model unit", where k can be equal to 0, 20, 40, 60, 80 and 100.

The overall score for the summary is based on the content unit coverage. In the offcial DUC results tables that NIST gives out, the score for the entire summary is the average of the scores of all the content model units.Some participants use slightly modified versions of the coverage metric, where the proportion of marked peer units to the number of model units is factored in.

The selection of units with the same content is facilitated by the use of the Summary Evaluation Environment (SEE)[5] developed at ISI, which displays the model and peer summary side by side and allows the user to make selections by using a mouse.

6 Application to human summaries

Here we illustrate the application of our scoring method to three sets of human summaries from DUC 2003. In order to make a representative comparison that

[4] http://www.isi.edu/licensed-sw/spade/

[5] http://www.isi.edu/c̃yl/SEE

highlights the differences between the DUC metric and the pyramid scores, we selected sets assigned very high and very low scores by the DUC scoring method. Our sample includes D30042, (referred to here as **Lockerbie**), which had the highest average score in the DUC method, and the two sets that the DUC method assigned the lowest scores, D31050 (**China Democracy**) and D31041 (**PAL**—for Philippine Airlines). For the text of these summaries, see Appendix A; note that the **PAL** set is also shown in Figure 2.

As with many other scoring methods, including the DUC method we compare our scores with, we believe a global score of summary content selection is desirable in order to have a uniform metric of comparison across summaries, whether written by humans or generated by machines. When we look at the scores assigned by the DUC method to the three sets examined, here, we will review in more detail some of the drawbacks regarding the difficulty in interpreting the DUC scores. In brief, the drawbacks are that the method is asymmetrical, in that scores depend on the choice of model summary to score against; scores vary widely depending on factors other than the content contained in the summary, which hinders reliable conclusions based on the score.[6] In addition, some of the variablity we see with this scoring method is due to arbitrariness in the decisions a scorer must make. In contrast to the DUC scoring method, we aim for a *symmetrical* score, i.e., one which does not depend on the choice of a model; one that quantifies the content selection in a manner that supports meaningful conclusions; and one that can be applied reliably by different evaluators. A well-defined procedure for arriving at judgments on summary content will have another advantage apart from reliability, namely a basis for determining the feasibility of an automated method to perform a similar procedure.

Table 2 presents the DUC scores assigned to the three sets of summaries. As described above, to apply this method requires selecting a model summary; column 1 of the table gives the model for each set. Note that no metric can be assigned to the *model,* so there is no way to determine whether one model (e.g., A for **Lockerbie**) is better than another (e.g., H for **PAL**). A second drawback in choosing a designated model is that the resulting scores are asymmetric, that is, the scores necessarily change depending on which summary is selected to be the model.

In considering Table 2, consider the following additional problematic distributions. First, the variation across the high scoring set (**Lockerbie**) would suggest, assuming a *meaningful* metric, that these four humans have extreme variation in their summarization skills, with summarizer C being the least adept summarizer (C has the Minimum score of .54), and summarizer B (with the Maximum score of .82) being the most adept. In comparison, we see much less variation within the other two sets in the table. Note also that summarizer D seems to have produced a relatively good summary for **Lockerbie** and a relatively poor one for **China**, as if the same summarizer happened to vary widely on different document sets.

[6]For example, we suspect it has an undesirable sensitivity to the thematic similarity of the source texts.

Lockerbie			
A (Model)	B	C	D
n.a.	.82	.54	.74
PAL			
H (Model)	A	I	J
n.a.	.30	.30	.10
China Democracy			
C (Model)	D	E	F
n.a.	.28	.27	.13

Table 2: DUC Scores for the Three Sets of Human Summaries

Summarizer	Mean	Std. Dev.
H	.61	.14
D	.53	.12
A	.49	.18
B	.49	.12
C	.48	.08
I	.46	.13
E	.43	.14
G	.43	.12
J	.39	.13
F	.34	.18

Table 3: Means DUC Scores and Deviations for 10 Human Summarizers

In fact, the variation for the three **Lockerbie** scores reflects a general distributional inconsistency: among the 10 humans who participated in the summarization task, the individual and group variation is extremely high, and apparently random. There were a total of 90 human summaries scored on the DUC method: the average score was .47, with a high standard deviation of .16 (Min=.10; Max=.82), meaning any individual's score is likely to be 33% higher or lower than the average. Table 3 gives the averages and standard deviations for each human; H had the highest average (Avg=.61; SDev=.14), and F the lowest (Avg=.34; SDev=.18, i.e., any score by F is likely to be 50% higher or lower than this average); most of the summarizers were close to the average but showed spreads from 17% to 53% around the mean.

In Figure 7, a scatterplot of the scores of the 10 human summarizers by summary, the summaries have been numbered from 1 to 30, with each human scored on 9 of these summaries (where the 10th summary is the model). This figure illustrates the striking lack of regularity in the way the scores vary: no two humans have the same pattern of increases and decreases, and no two document sets have the same ordering or grouping. Rather than assuming that the skills of the human summarizers vary as extremely as these scores suggest, or change from text set to text set independently of how other humans do on the same

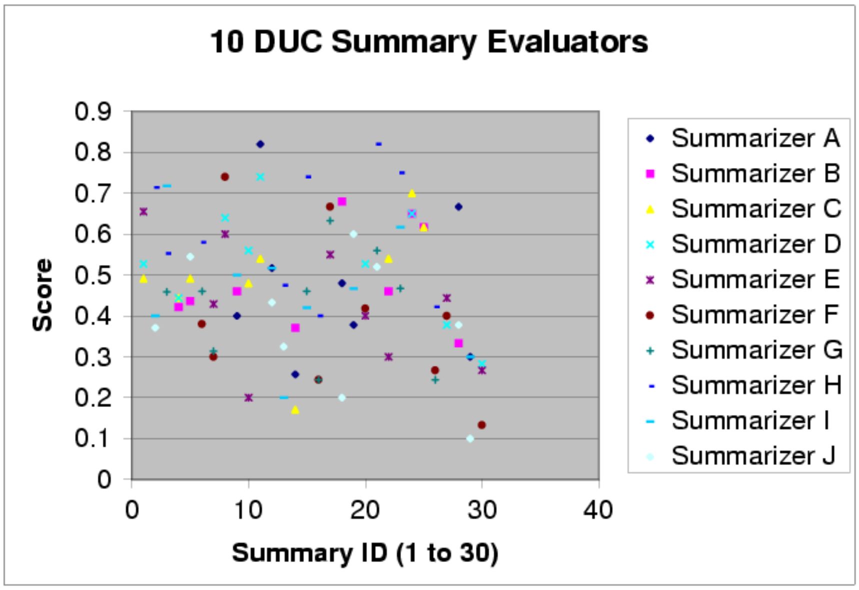

Figure 7: Scatterplot of Human Summarizer Scores for DUC 2003 Human Summaries

text sets, we aim for a metric that is both more robust, meaning that any differences between humans are more stable, and that offers a more informative comparison.

There is a second problematic fact about the distribution in Table 2. The scores on the **Lockerbie** set are much higher than the scores on the **PAL** and **China Democracy** sets, which would suggest a qualitative difference between the two groups of summarizers, or between the two sets of source texts. However, no obvious characteristic of the summaries distinguishes the apparently "poor" summaries from the "good" ones.

There are many possible explanations for the disparity in the DUC scores on the **Lockerbie** versus **PAL** and **China Democracy** summaries, but the one addressed by the pyramid scoring method is that certain pieces of information in a document set are going to be more relevant to more people, and should thus be assigned a higher weight than less relevant information. Since the DUC method treats all EDUs as having an equivalent weight, it is unable to distinguish between summaries that have a high proportion of information that a large sample humans would find important.

In the following section, we will explain how the pyramid scoring method eliminates two of the problematic distributional facts we see in the DUC scores, inexplicably high variance in human performance, and inexplicably great variation between summary sets. Like the DUC method, we assign a single score between 0 and 1 to a summary, based on the content, but our method avoids the three weaknesses discussed above:

- requirement for a designated "model" summary
- equal weight assigned to all content units (EDUs in DUC)
- insufficient reliability of the scoring method

6.1 Pyramid Scores of the Three Summary Sets

For the three sets of summaries discussed here, the two co-authors created a consensus SCU annotation, labelling each SCU with a numeric index, a semantic description, and the number of summaries the SCU appears in. Figure 6, for example, gives the SCU annotation for the PAL set. The two raters (the co-authors) assigned 33 versus 37 SCUs, and arrived at a consensus annotation with 35 SCUs for which we had 65 out of 69 pairwise agreements on the dual attributes of SCU identity and rank.

Each annotation generates a pyramid, which is a rank-ordered partition of SCUs. Table 4 presents the pyramid counts for the three sets of summaries. Each column shows the distribution of SCUs across the four weights.

Table 5 presents the scores assigned by the pyramid method to the three sets of summaries. Let's compare it with Table 2. First, note that because there is no designated model summary, and we use the same inventory of SCUs for scoring any summary, we can assign independent scores to all four. We have arranged the table so that the first column contains the summaries used as the *model* for

w	Lockerbie	PAL	China Democracy
4	5	2	3
3	9	6	6
2	9	12	6
1	11	15	24
TOTAL	34	35	39

Table 4: Pyramids for the three sets of summaries

Table 5: Pyramid Scores for the Three Sets of Human Summaries

Lockerbie			
B	A	C	D
1.0	.88	.73	.97
PAL			
A	H	C	J
.95	.89	.85	.76
China Democracy			
C	D	E	F
.88	.80	.79	.73

the DUC method. Second, note that the scores for the three sets are much more comparable than the DUC method, and much higher. This suggests that our methods treats summaries more equivalently, independent of which human has written the summary, and somewhat independent of the source text (of course, some texts may be inherently more coherent, and easier to summarize).

Most important for a meaningful evaluation metric, the pyramid scores can be given a very specific interpretation. The highest scoring summary, summary B for **Lockerbie** received a score of 1: this means that whatever the number of SCUs in this summary, we know they are distributed from top to bottom in the pyramid, with the higher ranking SCUs being fully represented before any lower ranking SCUs appear. Similarly, the interpretation of the low scores is that these summaries contain a lower proportion of high-ranking SCUs, and a higher proportion of low-ranking SCUs. Furthermore, we can meaningfully compare rows and columns of Table 5. **Lockerbie** summary C has the lowest score, meaning it has a higher proportion of SCUs from the bottom of the pyramid; in fact, it is equivalently weak in relevant content as **China** summary F.

A global score for content selection is desirable in order to compare systems, but its utility is limited by the meaningfulness of the conclusions that can be drawn. Our first goal in developing a method to evaluate machine generated summaries is thus to create a meaningful score for human summaries given the observations noted in the preceding section, namely that summaries can vary in length, that SCUs differ in their informational status with some being more necessary to an adequate summary, and that SCUs are semantic abstractions whose realization (and identification) in text exhibits a great deal of individual variation, as well as being dependent on human cognition and on socio-cultural assumptions. Because our long-term goal is to assign appropriate weights to most of the information from the source texts, the pyramid metric can, in principle, score a summary of any length. Given an incomplete pyramid, a very long summary containing information might be unfairly penalized (cf. [5], where it is noted that length can affect results significantly, which they cite as a problem with designating a single *model* summary).

6.2 Second Order Summaries

If we could identify suitable features to assign to document sets and the SCUs derived from them, it ought to be possible to discover generalizations about how humans decide what content to include in a summary—given a sizable enough training set—using statistical or machine learning techniques. However, it's likely such a training set would need to be quite large, with a larger number of summaries per source text as the compression rate and/or number of source documents increases. For example, from Figure 1 [3], we see that the total number of distinct factoids in a set of 100-word summaries of a single 600-word document continues to grow steeply even after 10 summaries. In their data, the compression rate is 1/6, from a single source text. This strongly suggests that the DUC data, consisting of four 100-word summaries per document set, where each document set consists of ten news articles of about 500 words each (compression rate of 1/15), would not be an ideally large training set; still, it might yield useful information about the characteristics of differentially weighted SCUs, and we hope to investigate whether this is possible.

Given the cost of creating a large training set, it would be extremely useful to develop other less costly methods for collecting our pyramid data. Here we present preliminary results of an alternative method we examined. We elicited what we call second-order summaries from subjects, in which people were asked to write 100-word summaries of DUC document sets after consulting only the original 4 summaries, or the summaries plus source texts. Although this method needs further development, it offers strong evidence that SCU frequency across summaries is only an indirect reflection of a more general cognitive process.

We take the distributional facts of SCUs within a sample of human-written summaries of the same material to be an indirect reflection of each individual's linguistic and cognitive processes, and of the group's collective consciousness as to which units of meaning in the documents are more important, more representative, or more useful.[7] On this assumption, frequency of a given SCU within a corpus of summaries of the same document is an artifact of a more primary cognitive phenomenon that we might be able to study in other ways. In [21], annotators ranked sentences within a document on a 10-point scale of

[7] We also assume that one individual might write different summaries (cf. [25]), depending on whether he or she understands the task to involve a presentation of a relatively *objective* encapsulation of key information, or in terms of a more specific, and perhaps loaded, rhetorical or information-seeking purpose.

Summarizer	DUC Score	Pyramid Score
Lockerbie: Brief		
M	.40	.97
N	.58	.86
O	MODEL	.82
P	.67	.91
Lockerbie: Full		
M	.58	.86
Q	.44	.90
R	.58	.70
S	.79	.88
China Democracy: Brief		
M	.60	.72
N	.67	.93
O	MODEL	.84
P	.55	.80
China Democracy: Full		
M	.38	.62
Q	.54	.71
R	.49	.74
S	.73	.80
T	.44	.79

Table 6: Second Order Summaries: DUC versus Pyramid Evaluation

relevance to the topic of the document, and also assigned entailment relations to sentences. Such relevance-ranking probably reflects judgments similar to the human process of content selection for summarization. In addition, the combination of sentence ranking and entailments in [21] resembles the decisions annotators must make in creating an SCU annotation.

In the alternative source of evidence we discuss here, we asked humans to write 100-word summaries from sets of DUC summaries. We had two conditions for second order summaries, which we collected for the **Lockerbie** and **China Democracy** sets. In the **full** condition, subjects read the summaries along with the ten source texts. In this condition, the summaries presumably serve as *navigational* aids in reading the source texts, as well as providing reinforcement during the summarization process. In the **brief** condition, subjects read only the original summaries. Table 6 compares DUC and pyramid scores on the second order summaries. The DUC scores represent the average of the two authors' scores. As we can see, the DUC scores were much higher for the second order summaries than for the first order ones, although the comparison is only valid if the authors' execution of the DUC method was consistent with the DUC evaluators. Again, the scores from the pyramid method are much higher than the DUC method; they are not significantly different from the pyramid scores of the first order summaries.

The pyramid scores for the **brief** condition appear higher than for the **full** condition, although the small sample prevents us from placing too much emphasis on this observation. The pyramid average for **brief** is .89, and for **full** it is .87. We expected lower scores on the **full** condition, because we expected that if summarizers could consult the original texts, it was more likely that they would select information that had not appeared in any other summary. This indeed happened, with scores for the second order summaries dropping, but still they remained in the range above .70. Though we did not have subjects write new, original summaries (i.e, without access to the previously written summaries), we believe such summaries would also receive reasonably high scores, given the results under the **full** condition.

Some of the more interesting observations we find in the second order summaries depend on qualitative analysis. For example, we found that occasionally, an SCU would appear in all second stage summaries that had appeared in only a single first stage summary, which suggests it had accidentally low coverage in the initial summaries.

7 Application to machine summaries

Machine generated summaries are first broken down into clauses, that can be divided further if a clause contains more than one SCU from the pyramid inventory. The total number of units obtained in this way is the SCU size of the summary and is used to compute the score for an ideal summary, as described by the scoring formula. The need to divide the machine summaries initially into clauses comes from the fact that machine summaries very often contain entire sentences that have no overlap with any SCU in the pyramid inventory. The sentence could be kept as one unit, but then very long complex sentences contain a lot of information. It is fairer to consider such sentences as contributing several SCUs to the summary size rather than a single one, given our SCU annotation procedure, which directs annotators to split an SCU of weight 1 into distinct SCUs of weight 1, if the SCU member has more than one propositional constituent (see Appendix B)

Information is often unnecessarily repeated in machine generated summaries. Figure 8 gives an example of the phenomenon: e.g., it contains the two phrases *hand over for trial two suspects* and *turn over for trial two other Libyans wanted for ...*. So the next step is to identify units within the summary that contain repeated information. Such units are combined and subsequently scored only once. Among the 100 word generic machine-produced summaries submited for DUC 2003, 30% contained unnecessarily repeated information (this figure can be computed from the counts for quality question number 11, which explicitly asked human evaluators reading the summary to say if there is any repeated information).

From here on the evaluation proceeds as in the case for human produced summaries – each unit that is found in the pyramid inventory is given a score as defined by the rank of the unit in the pyramid, and SCUs that do not appear

African countries voted in June to ignore the U.N. flight ban which was **imposed in 1992 to try and force Libya to hand over for trial two suspects wanted in the 1988 bombing of an American airliner over Lockerbie, Scotland.** The reported jailing of the three officials comes as **Gadhafi is under pressure to accept a plan to turn over for trial two other Libyans wanted for the 1988 bombing of Pan am flight 103 over Lockerbie, Scotland, that led to 270 deaths.** The visit was Farrakhan's fifth to Libya in the past three years. The leader of the U.S.-based Nation of Islam most recently visited in December 1997.

Figure 8: System 19 summary for the Lockerbie docset.

system	DUC score	Pyramid Score
sys10	0.20	0.36
sys11	0.30	0.40
sys12	0.40	0.45
sys13	0.38	0.48
sys14	0.42	0.56
sys16	0.34	0.38
sys17	0.18	0.35
sys18	0.34	0.52
sys22	0.40	0.56
sys23	0.20	0.40

Table 7: Scores for machine summaries on the Lybia set

in the pyramid are given a score of zero. The overall weight of the summary is computed and is then normalized by the weight of the Max summary for the specific size.

There is an immediately noticeable benefit from the application of the pyramid evaluation method. It is more effective in discriminating between the summaries. Summaries that received the exact same score in the DUC evaluation, received noticably different scores when scored with the pyramid method.

We examined closely pairs of such summaries, that were indistinguishable according to the DUC evaluation, but markedly different according to the pyramid evaluation. Figure 9 shows two summaries that changed their ranking in the two methods. Summary 17 and summary 6 for the PAL set had respective DUC scores of 0.25 and 0.10 in DUC and respective pyramid scores of 0.26 and 0.64.

If we look at the human summaries for the same set (PAL), we see that all humans present the stories from the same angle: PAL has financial troubles, different factors make the troubles worse and the company is looking for ways to get out of it. In this sense, summary 6 more closely resembles a human summary; it states that the Philippine Airlines has a big debt and that strikes

Summary from system 6
PAL, Asia's oldest airline, has been unable to make payments on dlrs 2.1 billion in debt after being devastated by a pilots' strike and by Asia's currency crisis. PAL earlier accepted a preliminary investment offer from Cathay Pacific, Ailing Philippine Airlines and prospective investor Cathay Pacific Airways have clashed over the laying off of PAL workers, prompting PAL to revive talks with another foreign airline, an official said Tuesday. Cathay Pacific Airways said Wednesday it had pulled out of talks to buy a stake in ailing Philippine Airlines - making the uncertain future at PAL even cloudier.

Summary from system 16
President Joseph Estrada on Saturday urged militant unionists at Philippine Airlines to accept a vote by workers approving a 10-year no-strike deal to revive the debt-laden airline. President Joseph Estrada said Saturday the financially troubled Philippine Airlines will resume its international flights on Sunday by flying him to Singapore where he will address the World Economic Forum. Philippine Airlines said Thursday it will attempt to rebuild alone after Hong Kong's Cathay Pacific Airlines pulled out of talks on acquiring a stake in the ailing Philippine flag carrier. A strike by employees precipitated the airline's near-death experience in September, when Tan shut down the carrier after its unions re fused to accept a drastic cost-cutting plan.

Summary from system 17
Christmas is a sacred holiday in the Philippines, and nowhere is that more evident than at the headquarters of Philippine Airlines. But Ramos, who was intent on privatizing the economy, opened the industry to competition, licensing rivals like Air Philippines, Cebu Pacific and Grand Air. PAL closed for nearly two weeks on Sept. 23 after failing to persuade its largest union to accept a management-proposed recovery plan under wh ich their collective bargaining agreement would be suspended for 10 years in exchange for a 20 percent share of the company. The union had been sharply split over the proposal, with militant members saying a suspension of the bargaining agreement would violate workers' rights.

Figure 9: Three reranked summaries. The pyramid scores indicates system 6 produced a better summary and this corresponds to human judgment of the content of the two summaries. Systems 17 and 16 got exactly the same score at DUC, but 16 comes out better in the pyramid method.

system	DUC score	Pyramid Score
sys13	0.30	0.79
sys14	0.03	0.24
sys16	0.25	0.51
sys17	0.25	0.26
sys18	0.03	0.17
sys20	0.03	0.20
sys 6	0.10	0.64

Table 8: Scores for machine summaries on the PAL set

in the company and failed negotiations worsen the problems. Mentioning the finacial troubles is semantically necessary in order to motivate and explain the necessity for shutdown and the negotiations with other companies. Summary 17 fails to give such a context and regardless of the fact that it gives information on the PAL shutdown and the bargaining over the recovery plan, the authors feel it is less helpful than summary 6. This intuition is adequately captured by the pyramid scoring method.

The difference between summary 16 and 17 is not as big as between 17 and 6, but is still noticeable. Even though summary 16 has extraneous information on President Estrada, phrases such as *the financially troubled Philippine airlines* and *the airline's near-death experience in September* convey enough of the main topic of the docset. Summary 17 on the other hand is simply a collection of disconnected facts with no main point or focus.

8 Conclusion

We have presented a method for representing the abstract content units of text in order to create a weighted inventory of information content for evaluating summarization. The most obvious conceptual limitation of the pyramid scoring method is that we only address the issue of content selection. On the one hand, we believe content selection is logically distinct from other issues in naturalistic summarization, such as fluency, coherence, interdependencies among content units, rhetorical structure, perspective, and the possibility that sometimes the inclusion of information can be more distracting than helpful (cf. [2]). We hope we can tackle these other issues given our more informative and reliable method for evaluating content selection. For example, in future work, we hope to apply the same scoring method to question-answering, using the semantics of the question to redistribute information within the pyramid.

The other primary drawback to the evaluation method proposed here that we aim to address in subsequent work is that we have not yet determined how feasible it will be to automate. There are two primary tasks that require automatic or semi-automatic methods:

- populating a pyramid with SCUs, given a body of source text

- scoring a summary against a given pyramid

We hope to test a variety of methods on each task. For example, it may be possible to use machine learning methods to populate a pyramid from textual sources, using an appropriate distance metric (e.g., a multi-dimensional one) and clustering techniques, given a training corpus where humans have annotated SCUs. We believe that a combination of deep (knowledge intensive) methods for the exploratory phase followed by statistical and machine learning methods (as in [17]) will establish a baseline and a clear path for performance improvements.

Appendices

A Three Sets of Human Summaries

A.1 The *Lockerbie* Set (D30042)

D30042.A

A.1 In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.
A.2 In 1992 the U. N. voted sanctions against Libya for its refusal to turn over the suspects.
A.3 Suffering from the sanctions, Gadhafi alternated between defiance and acceptance.
A.4 In August of 1998 the U.N. proposed a trial in the Netherlands under Scottish law which Libya accepted in principle, but insisted that any sentence be served in the Netherlands or Libya.
A.5 The U.N. threatened a tightening of sanctions.
A.6 Then, in late November, the U. N. Secretary General hinted that he might broker an agreement.
A.7 Optimism arose, but the issue was still in doubt.

D30042.B

B.1 Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.
B.2 The United Nations imposed sanctions on Libya in 1992 because of their refusal to surrender the suspects.
B.3 The sanctions included a ban on all international flights.
B.4 After six years of sanctions, Libyan leader Moammar Gadhafi, faced with threats of additional sanctions, agreed in principle to hand over the suspects.
B.5 The two suspects would be tried in the Netherlands by Scottish judges under Scottish law.
B.6 Gadhafi wants guarantees, including a promise that the suspects would serve their sentences in the Netherlands or Libya if convicted.

D30042.C

C.1 Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.
C.2 The U.N. imposed international air travel sanctions on Libya to force their extradition.

C.3 The sanctions were honored by all but certain African countries.
C.4 In 1998, a U.S.-Britain brokered compromise offered that their trial to be held in the Netherlands.
C.5 Gadhafi, tired of sanctions and hoping for increased tourism, actively pursued this compromise with UN Chief Kofi Annan.

D30042.D

D.1 In December 1988 a Pan Am jet was blown up over Lockerbie, Scotland, killing 270.
D.2 Two Libyan suspects were indicted in 1991.
D.3 Since 1992 Libya has been under U.N. sanctions in effect until the suspects are turned over to United States or Britain.
D.4 Sanctions include an air embargo.
D.5 African leaders disregard them.
D.6 Libya has fostered tourism to help the damaged economy.
D.7 In August 1998 United States and Britain proposed a Netherlands trial.
D.8 Libya agreed, but then asked for guarantees that the suspects, if convicted, would be incarcerated in Libya.
D.9 This delayed progress.
D.10 Kofi Annan planned a December 1988 Libyan trip to move negotiations.

A.2 The *PAL* Set (D31041)

D31041.A

A.1 Philippine Airlines (PAL) experienced a crisis in 1998.
A.2 Unable to make payments on a $2.1 billion debt, it was faced by a pilot's strike in June and the region's currency problems which reduced passenger numbers and inflated costs.
A.3 On September 23 PAL shut down after the ground crew union turned down a settlement which it accepted two weeks later.
A.4 PAL resumed domestic flights on October 7 and international flights on October 26.
A.5 Resolution of the basic financial problems was elusive, however, and as of December 18 PAL was still $2.2 billion in debt and losing close to $1 million a day.

D31041.H

H.1 Starting in May 1998, Philippine Airlines (PAL) laid off 5000 of its 13,000 workers.
H.2 A 3-week pilots' strike in June and a currency crisis that reduced

passenger numbers made payments on PAL's $2 billion debt impossible.
H.3 President Estrada brokered an agreement to suspend collective bargaining for 10 years in exchange for 20% of PAL stock and union seats on its board.
H.4 The large ground crew union initially voted no.
H.5 After PAL shut down operations for 13 days starting Sept. 23rd, leaving much of the country without air service and foreign carriers flying some domestic routes, 61% voted yes.
H.6 Unions agreed to some employee cuts with separation benefits.

D31041.I

I.1 Philippines Airlines (PAL), Asia's oldest airline, devastated in 1998 by pilot and ground worker strikes and with a rising $2.1 billion debt, stopped all operations for 13 days.
I.2 A 10-year, no-strike agreement was finally ratified when the government began using foreign carriers for domestic flights.
I.3 PAL resumed domestic flights on October 7 and international flights slowly over the next weeks.
I.4 This shutdown had been another blow to the debt ridden national flag carrier.
I.5 Following the strike, efforts to revive the airlines met more obstacles.
I.6 Cathay Pacific Airways would only help if the payroll, including 200 pilots hired during the pilot strike, were slashed.

D31041.J

J.1 The fate of Asia's oldest airline, PAL, is uncertain.
J.2 Negotiations with Cathay Pacific Airways to infuse $100 million dollars into the company collapsed when PAL Chairman Tan refused to agree to major job cuts and to relinquish management control to Cathay.
J.3 PAL is buried under a $2.2 billion dollar debt it cannot repay and $1 million a day losses.
J.4 A replacement investor is unlikely and President Estrada, though supported by Tan in the election, has rejected a government bailout.
J.5 PAL's financial troubles were exacerbated by a two-week shutdown in September due to a dispute with its largest union.

A.3 The *China Democracy* Set (D31050)

D31050.C

C.1 Making obvious their intention to suppress the fledging China Democratic Party, Communist officials arrested three of its most prominent leaders, Xu Wenli, Qin Yongmin and Wang Youcai.
C.2 Colleagues protested and the USA expressed concern, to no avail.
C.3 China, defending its action, brought all three to trial.
C.4 Because potential defense lawyers were harassed, all three had to defend themselves.
C.5 Qin and Wang were charged with inciting subversion, which carries a minimum sentence of five years imprisonment.
C.6 Xu faces the more serious charge of subverting state power which can bring a 1life sentence.
C.7 Chinese President Jiang said multiparty democracy will not be allowed.

D31050.D

D.1 In October 1988 China signed the International Covenant on Civil and Political Rights.
D.2 Perhaps emboldened by that, dissidents pushed for the recognition of a new China Democratic Party.
D.3 The government, citing a long-standing policy, said that it would not tolerate another political party.
D.4 Three prominent dissidents, one of them Wu Xenli, working to establish the new party were arrested.
D.5 They were accused of endangering national security, and faced prison terms of five years.
D.6 Prospective defense lawyers were intimidated.
D.7 The United States criticized the arrests, and an embassy official was rebuffed when he tried to observe a trial.

D31050.E

E.1 With the arrest of political dissidents, including Xu Wenli, Wei Jingsheng, and Qin Yongman, China went on the offensive against individuals seeking to set up political parties in that country.
E.2 In a speech to the Communist Party, Chinese President Jiang Zemin denounced any attempts to set up a multi-party system in China.
E.3 The three prominent dissidents all were arrested on charges of subverting national security.
E.4 With government harassment of lawyers, the dissidents have been unable to secure legan counsel and have had to defend themselves.
E.5 The father of Qin Yongman has sought a delay in his son's trial until legal counsel can be found.

D31050.F

F.1 Chinese officials have finally spoken out on their dealings with dissidents trying to establish the China Democracy Party.
F.2 They claimed that Xu Wenli, was suspected of "activities damaging to national security".
F.3 The US has denounced China's detaining those "peacefully exercising fundamental freedoms".
F.4 Those arrested are expected to be tried under China's vague State Security Law.
F.5 Obtaining lawyers for their trials became virtually impossible forcing some of the dissidents to defend themselves.
F.6 President Zemin said that economic reforms would continue but would not be a prelude to multiparty democracy.
F.7 China did release ailing Liu Nianchun from prison work camp and exiled him to the US.

B Annotation of Summarization Content Units (SCUs)

The following procedure is a step-by-step process for identifying a set of SCUs, along with their coverages (or weights), in a set of summaries of the same text. It is intended to be used by knowledgeable annotators who understand the summarization task, who have more than a passing familiarity with semantics, and who are motivated to do a careful job. It requires patience, particularly at the beginning: the initial decision steps can often seem unclear, and the procedure requires multiple iterations over the evolving annotation. However, as the task nears completion, the decisions become more clear cut, there is less need to reconsider previous decisions, and the annotation moves more rapidly.

B.1 Overview

In principle, the task requires the human annotator to compare all the sentences in all the summaries in order to locate different expressions of the same content; we will exemplify what we mean by *the same* below. In essence, after locating similar expressions, the annotator proposes an SCU, then records the phrase or phrases that express it. During annotation, an SCU is considered provisional until it is *resolved*, a confirmation step we describe below. As the annotation proceeds, less and less of the original text needs to be examined to identify the remaining SCUs. The annotation ends when no text remains to be considered, and all SCUs have been resolved. The stepwise procedure makes use of two work spaces: one to record the evolving SCUs, and one to record what remains in the source text that has not yet been identified as expressing a particular SCU.

B.2 Materials

- A software tool for editing that the annotator is comfortable with and that allows for two distinct screens or windows. Much of the annotation task involves copying or cutting text from one place and pasting it in another.
- A file containing the set of summaries. We indicate how to preprocess the summaries below.
- Recommended: A printed copy of the summaries, after they have been preprocessed as described below.
- Recommended: A copy of the original text the summaries were created from. This supports the preprocessing phase, and can also be useful to resolve vagueness or ambiguity in the summaries.

B.3 Procedure

B.3.1 Step 1: Preprocessing

Indexing. For all summaries S, assign an index to each summary. Then index each sentence. In the examples below, we use the four summaries from DUC 2003 the **China Democracy** set (D31050) with the alphabetic indices they were assigned (C through F), and here we use C1 through C7 for the seven sentences in summary C, and so on. A convenient layout of the summaries is one sentence per line, with the sentence index at the beginning of the line, as in Appendix A.3.

Copy Editing. Summaries often contain misspellings or minor inconsistencies, particularly of proper names, dates, and numbers. It can eliminate confusion to identify these in advance, particularly for annotators who have not read the source texts, and who are not familiar with the content of the summaries. For example, in the **China Democracy** summaries, a prominent Chinese dissident by the name of Xu Wenli is mentioned. In one summary, he was referred to as Wu Xenli, which could lead to confusion. Thus we recommend altering:

D4 Three prominent dissidents, one of them Wu Xenli,
working to establish the new party were arrested.

to:

D4 Three prominent dissidents, one of them Wu Xenli (sic; *Xu Wenli*),
working to establish the new party were arrested.

Similarly, in order to avoid the possibility that an annotator would identify distinct SCUs on the basis of irrelevant differences, these differences can be identified in advance. For example, the dollar amounts in the following examples from the **PAL** summaries (D31041) could be marked in the text.

A2 Unable to make payments on a \$2.1 (round to \$2) billion debt, ...
H2 ... made payments on PAL's \$2 billion debt impossible.
I1 ... with a rising \$2.1 (round to \$2) billion debt, ...
J3 PAL is buried under a \$2.2 (round to \$2) billion dollar debt ...

B.3.2 Step 2: Create Workspaces

Workspace One: Record of Text Examined Put the summaries that have been formatted one sentence per line into the first workspace. This will be used for two purposes:

1. to record what portion of each sentence has already been *consumed*, i.e., has been assigned to a specific SCU;
2. to keep track of the remaining text that needs to be examined

Workspace Two: Record of SCUs and Their Status Start a second workspace in which to record the SCUs as they evolve. Each SCU will have:

1. a label, which may change as the task proceeds;
2. a set of phrases with indices indicating what sentence they came from;
3. a status indicator to record whether the SCU has been fully resolved.

B.3.3 Step 3: Propose an SCU

In this section we illustrate the first proposed SCU in the **China Democracy** summaries. This example will show that during the initial phase of an annotation, a proposed SCU is only provisional; it can change as the annotation proceeds. Here, the proposed SCU1 was initiated with a phrase from sentence C1, yet by the time we completed the annotation, no portion of C1 remained in SCU1. We use this same example to show how, during **Step 4**, the process of *resolving* an SCU can lead to splitting a provisional SCU into two provisional SCUs. (SCU numbers are, of course, arbitrary; they merely provide identifiers for ease of reference.)

If you are just beginning: go to the first sentence in **Workspace One**. Place an open square bracket between the line index and the beginning of the text, to indicate that the current sentence is being considered for the next SCU. Insert **SCU1** at the top of **Workspace Two** and copy the partly bracketed sentence, along with its line index, from **Workspace One** into the space below the **SCU1** header you just created:

Workspace One

C1 [Making obvious their intention to suppress the fledging China Democratic Party, Communist officials arrested three of its most prominent leaders, Xu Wenli, Qin Yongmin and Wang Youcai.
C2 Colleagues protested and the USA expressed concern, to no avail.
… …

Workspace Two
SCU1
C1 [Making obvious their intention to suppress the fledging China Democratic Party, Communist officials arrested three of its most prominent leaders, Xu Wenli, Qin Yongmin and Wang Youcai.

Read the summaries to locate textual material in another sentence that has some overlap in content with the sentence in the provisional SCU1 (e.g., C1). Categories of overlap include (for textual examples, see below):

- nearly the same propositional content in any two constituents, tensed or untensed
- same argument fillers to a synonymous verb
- same verb with argument fillers that have an inferential relation as defined in [15]

Here, as we read through the summary D sentence-by-sentence, the first meaningful overlap occurs in D3, concerning the Chinese government's declaration that it would not tolerate *another political party*, which is similar in content to C1 *their intention to suppress the fledgling China Democratic Party.* No other sentence in D has more overlap with this material in the first clause of C1; at this point, three changes are made, one to Workspace One and two to Workspace Two:

1. Workspace One: To record the consumed material from D3, enclose the relevant phrase in D3 within brackets, labeling the rightmost bracket with the SCU number;
2. Workspace Two: To identify the provisional phrase in C1 that contributes to SCU1, add the right bracket indexed with the SCU number–$]_1$–and remove the rest of the sentence;
3. Workspace Two: To record the growth of SCU1, insert **a copy** of the bracketed phrase from D3 in Workspace Two. Note that the referent of the NP *The government* in D3 is the first argument of both the main clause verb *to say* and the gerund *citing.* For the sake of clarity, we show D3 below with the subject of *say* and omit the parenthetical gerundive phrase

The workspaces should now look as follows:

Workspace One

C1 [Making obvious their intention to suppress the fledging China Democratic Party,$]_1$ Communist officials arrested three of its most prominent leaders, Xu Wenli, Qin Yongmin and Wang Youcai.
C2 Colleagues protested and the USA expressed concern, to no avail.
... ...
D3 The government, citing a long-standing policy, [said that it would not tolerate another political party.$]_1$

Workspace Two
SCU1
C1 [Making obvious their intention to suppress the fledging China Democratic Party,]$_1$
D3 [{*the government*}said that it would not tolerate another political party.]$_1$

The three steps shown above establishes that the content of SCU1 will have something to do with the overlap pertaining to a negative attitude of the Chinese government towards an opposition party. The addition of material from D3 has been added to SCU1 in **Workspace Two**, and **Workspace One** has a record of this same material being consumed in the evolution of SCU1.

As we proceed in this fashion linearly through the next two summaries, E and F, we reach a point where we have material from all four summaries associated with the provisional SCU1, as illustrated below:

Workspace Two
SCU1
C1 [Making obvious their intention to suppress the fledging China Democratic Party,]$_1$
D3 [{*the government*}said that it would not tolerate another political party.]$_1$
E3 [{*President Zemin*} denounced any attempts to set up a multi-party system in China.]$_1$
F6 [{President Zemin} said ... {CLAUSE} would not be a prelude to multiparty democracy.]$_1$

B.3.4 Step 4: Attempt to Resolve the Proposed SCU

Resolution of an SCU involves adding a label that expresses its content, and checking the consistency of the *covered* phrases with that content. Any material expressed in a phrase from one of the summaries that is inconsistent with the SCU label, or that has too much additional content in comparison to the other phrases, should either be removed from the brackets in both workspaces, or should be moved to a new SCU in Workspace Two. First we illustrate resolution by example. Then we list the criteria used in the resolution phase.

When we check the contents of the provisional SCU1, we find that the phrases from D3, E3 and F6 are more like each other than they are to C1. For these, we created the label *Chinese government or governmental representative speaks out* *against multiparty system* (cf. new version of **Workspace Two** below). To the right of the label, we add a question mark in square brackets to indicate that the SCU is not yet fully resolved. When we do a second pass over the summaries, we find another sentence in summary C, namely C7, that is closer to the current SCU1 than C1 is. For the moment, we move C1 to a new provisional SCU, and add C7 to SCU1:

> **Workspace Two**
> **SCU1** Chinese government or governmental representative
> speaks out against a multiparty system [?]
> D3 [{*the government*}said that it would not tolerate
> another political party.]$_1$
> E3 [{*President Zemin*} denounced any attempts to set up a
> multi-party system in China.]$_1$
> F6 [{President Zemin} said ...{CLAUSE} would not be a prelude
> to multiparty democracy.]$_1$
> C7 [Chinese President Jiang (sic; Zemin) said
> multiparty democracy will not be allowed.]$_1$

Note that lines C7 and C1 in **Workspace One** should also be updated so that the correct phrases are bracketed, and have the correct indices (not shown here).

We can attempt to fully resolve SCU1 now, or defer it to the final annotation step (Step Five). To do so now, we re-check the consistency of the phrases with the label for SCU; if they are consistent, we can remove the bracketed question mark. In this case, three of the summaries have *President Zemin* as the voice that speaks out against a multiparty system, while the fourth refers only to *the government*. To insure the internal consistency of SCU1, we move the three references to the Chinese president to a new SCU (SCU3). Skipping some of the details here, we label the new SCU3, check its consistency, and stop when we have a fully resolved SCU1 and SCU3, and a provisional SCU2. Here we show the weights assigned to the resolved SCUs (SCU1 and SCU3). We show SCU2 incremented with a phrase from E1, but we do not document the completion of

SCU2 here.

Workspace Two
SCU1 (w=4): Chinese government or governmental representative speaks out against a multiparty system [?]
D3 [{the government...} said that it would not tolerate another political party.]$_1$
E2 [{ a governmental authority} denounced any attempts to set up a multi-party system in China.]$_1$
F6 [{a governmental authority} said... would not be a prelude to multiparty democracy.]$_1$
C7 [{a governmental authority} said multiparty democracy will not be allowed.]$_1$

SCU2
C1 [Making obvious their intention to suppress the fledgling China Democratic Party,]$_2$
E1 [China went on the offensive against... {*individuals seeking to set up political parties*}]$_2$

SCU3 (w=3): The voice of governmental authority (SCU1) is President Zemin
E2 [Chinese President Jiang Zemin {*....denounced....multi-party system*}]$_3$
F6 [President Zemin ... {*said ... would not be a prelude ...*}]$_3$
C7 [Chinese President Jiang {*said multiparty ... not ... allowed*}]$_3$

Criteria for Resolution Step

1. preserve or prefer SCUs with greater coverage; that is, in considering different ways to split a provisional SCU during the resolution step, try to maximize coverage, as long as the semantic decisions remain commensurate

2. SCU members can have minor semantic differences. In the following example of an SCU from the **China Democracy** annotation, note that **C4** has *potential* where **D6** has *prospective* and **E4** has no corresponding adjective; **C4** and **D6** refer to *defense lawyers* while **E4** refers simply to *lawyers*; in **C4** and **E4**, *harrass* is the lexical stem of the action taken against the lawyers whereas in **D6** it is *intimidate*: **SCU9 (w=3):** arrested dissidents' lawyers were harrassed

   C4 [Because potential defense lawyers were harassed,]$_9$

D6 [Prospective defense lawyers were intimidated.]$_9$
E4 [With government harassment of lawyers,]$_9$

3. for an SCU of coverage X that seems internally inconsistent because there is non-overlapping content of a more significant nature than in item 2, then consider creating two SCUs, one with coverage X and one with a coverage less than X so long as resulting SCUs are either resolved, or are closer to being resolved (cf. the creation of SCU3 above)

4. for any singleton SCUs, split propositionally complex sentences or clauses into distinct SCUs (e.g., conjoined clauses, relative clauses, parenthetical clauses, infinitive clauses or phrases, gerundive phrases with one or more argument fillers)

5. when considering alternative ways to resolve an SCU, prefer a resolution that avoids singletons, as long as criteria 1 through 4 are adhered to

6. allow for SCUs based on inferred content **if**
   - this will lead to a bigger coverage SCU, **and**
   - this will lead to resolution of otherwise unresolved SCUs, **and**
   - you are confident that the writer intends the inference, given the rest of the summary (For example, we added F4 to an SCU about Xu Wenli's arrest because we inferred that the writer implicitly referring to the arrest; referring to Appendix A.3, however, you can see that the writer never explicitly mentions the arrest of Xu Wenli)

7. when there are apparent contradictions among the phrases within an SCU, assume the information is incomplete but consistent if there is a way to do so (e.g., the summaries discuss the charges brought against the dissidents as *endangerment of national security* or *subverting state*; whether these are considered to be the same charge may depend on more detailed knowledge than is available about Chinese law

B.3.5 Step 5: Resolve all Unresolved SCUs

Steps 3 and 4 should be repeated until all textual material in all sentences has been identified as contributing to a specific SCU, and recorded as such by enclosing relevant portions of text within brackets labelled with the SCU number in **Workspace One**, and by copying the bracketed phrases to the relevant SCUs in **Workspace Two**. When all the text has been so labelled, there might remain some unresolved SCUs. At this point, apply Step 4 to the remaining SCUs, using the criteria listed above.

References

[1] J. F. Allen. Natural language, knowledge representation and logical form. In R. Weischedel and M. Bates, editors, *Challenges in Natural Language Processing*. Cambridge University Press, 1991. To appear.

[2] Isabel L. Beck, Margaret G. McKeown, Gale M. Sinatra, and Jane A. Loxterman. Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly*, pages 251–276, 1991.

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

[4] Amy Isard and Jean Carletta. Replicability of transaction and action coding in the map task corpus. In *AAAI Spring Symposium: 1Methods in Discourse Interpretation and Generation*, 1995.

[5] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*, 1998.

[6] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA, 1980.

[7] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the Workshop on Automatic Summarization, post conference workshop of ACL 2002*, 2002.

[8] Inderjeet Mani. Summarization evaluation: An overview. In *NAACL 2001 Workshop on Automatic Summarization*, 2001.

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

[10] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Barry Schiffman, and Simone Teufel. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of DUC 2001*, 2001.

[11] Megan Moser and Johanna D. Moore. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the Association for Computational Linguistics*, 1995.

[12] Ani Nenkova and Amit Bagga. Facilitating email thread access by extractive summary generation. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP'03)*, 2003.

[13] R. C. Omanson. An analysis of narratives: Identifying central, supportive and distracting content. *Discourse Processes*, pages 119–224, 1982.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[15] Rebecca J. Passonneau. Protocol for coding discourse referential noun phrases and their antecedents. Technical report, Columbia University, 1994.

[16] Rebecca J. Passonneau. Applying reliability metrics to co-reference annotation. Technical Report CUCS-017-97, Columbia University, Department of Computer Science, June 1997.

[17] Rebecca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, pages 103–139, 1997.

[18] Rebecca J. Passonneau, Carl Weir, Tim Finin, and Marth Palmer. Integrating natural language processing and knowledge based processing. In *Proceedings of the 8th National Conference on Artificial Intelligence*, pages 976–983, 1990.

[19] Katerina Pastra and Horacio Saggion. Colouring summaries bleu. In *EACL 2003*, 2003.

[20] James Pustejovksy. *The Generative Lexicon*. MIT Press, Boston, MA, 1995.

[21] Dragomir Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, 2000.

[22] Dragomir Radev, Simone Teufel, Horacio Saggion, and W. Lam. Evaluation challenges in large-scale multi-document summarization. In *ACL*, 2003.

[23] G. J. Rath, A. Resnick, and R. Savage. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation*, 2(12):139–208, 1961.

[24] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing and Management*, 33(2):193–208, 1997.

[25] Karen Sparck-Jones. Automatic language and information processing: Rethinking evaluation. *Natural Language Engineering*, pages 29–46, 2001.