# Leveraging Pyramids

**Terry COPECK**, **Stan SZPAKOWICZ**
School of Information Technology and Engineering
University of Ottawa
800 King Edward Avenue
Ottawa, Ontario, Canada K1N 6N5
{terry, szpak}@site.uottawa.ca

## Abstract

The Pyramid system associates importance ratings with individual sentences in evaluated summaries. Given that many summarization systems compose their output from source document sentences with limited or no editing, it is often possible to propagate Pyramid ratings back to the source document and tag source sentences with one measure of their suitability to a generic summary. In the case of Pyramid evaluations of summaries of twenty topics in DUC 2005, this could be done in sufficiently many instances to make the effort worthwhile: 14% of nineteen thousand source document sentences were annotated with importance ratings.

## 1 Introduction

Consider a corpus in which sentences are ranked according to a publicly-documented measure of their suitability for use in a generic summary. The availability of such a resource would be of use when designing and refining a summarization system. To produce corpora of this sort is however very laborious, so they are not easily come by. Arrival of the Pyramid summary evaluation system (Nenkova and Passonneau, 2004) on the DUC scene in 2005 is an opportunity to develop such a resource. We describe how it was possible with limited effort to annotate a substantial percentage of the sentences in source documents with their Pyramid rankings by leveraging the significant investment already present in judgments of semantic equivalence in that system's materials.

Previous measures of summary quality used in DUC—in particular ROUGE and earlier manual evaluation facilitated by SEE, as well as manual assessments on various qualitative dimensions—do not assign explicit values to particular elements of the summaries under study. What is noteworthy about the new arrival is that it makes it possible to single out individual sentences and say of them, "this sentence contributes more to the summary than that one does".

As we understand it, the Pyramid system works in the following way. A group of manually-written reference summaries based on a collection of documents on a topic are edited by hand to produce a set of Summary Content Units (SCUs) which convey the main concepts of the topic. In syntactic terms, SCUs approximate simple declarative sentences or phrases. *The Court rejected Libya's plea concerning the extradition on the Lockerbie bombing suspects* and *The Peace Palace is the World Court's seat* are examples of SCUs. A set of SCUs and the manual summaries on which it is based together constitute a pyramid, whose representation is stored in XML in a .pyr file that can be accessed using the viewer which is part of the system.

Candidate summaries of the topic are evaluated in

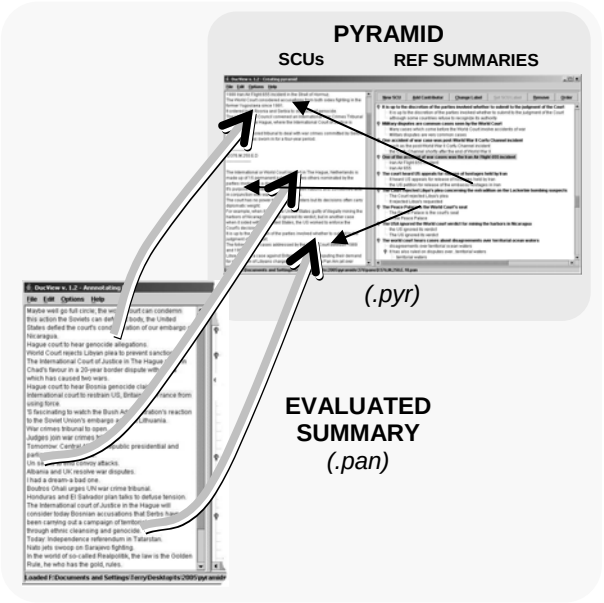


Figure 1: The Pyramid System

terms of the topic pyramid. Functions in the viewer let the user manually annotate a summary's component phrases and clauses and link them to those SCUs in the pyramid which are their semantic equivalents. Individual summary evaluations are stored together with the associated pyramids in XML .pan files. Figure 1 illustrates these elements of the Pyramid system.

Other factors such as length being equal, it is reasonable to maintain that a sentence in a summary whose constituents link to many SCUs contributes more to the summary than does a sentence with a single or no link to a SCU. In such a case the first sentence accordingly merits a higher ranking than does the second, and a highly-rated summary would thus have most or all of its syntactic constituents linked to SCUs.

Our system composes summaries by selecting suitable sentences from the document collection. We therefore wondered if it would be feasible to work backward from summaries which had been evaluated in the Pyramid system. We wanted to transfer the rankings assigned to their sentences back to the original sentences in the source document collection. This would result in a certain number of the sentences in a collection of source documents being characterized according to the Pyramid measure on their suitability to be in a generic summary of the collection topic without reference to any particular summary. Figure 2 illustrates the concepts involved in this plan.

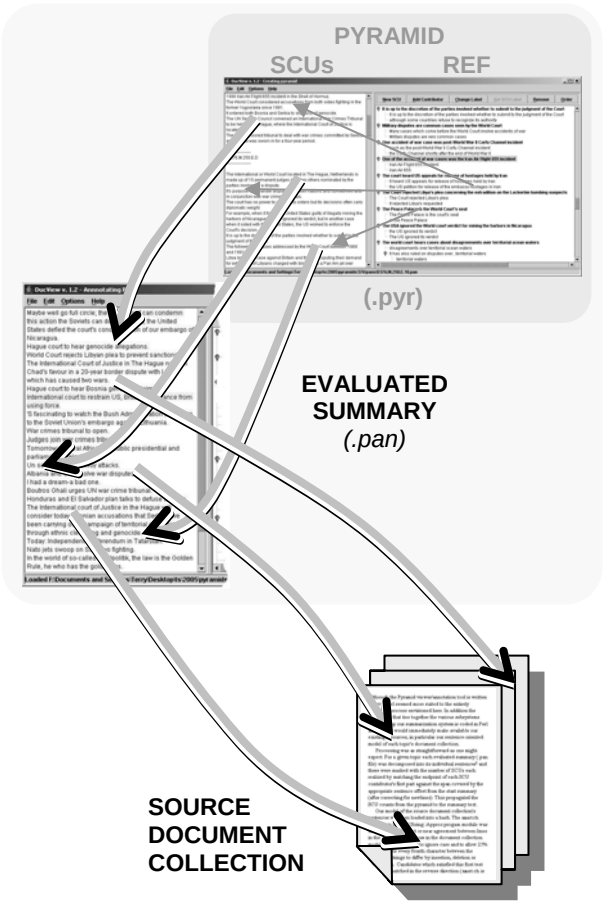


Figure 2 : Leveraging the System

The undertaking is premised on one key assumption. For a source document collection to be annotated to a sufficient degree to be of use to anyone, it is crucial that enough participants in DUC besides ourselves construct their summaries from elements that can be identified with sentences in source documents. We are happy to be able to report that this proved to be the case, and that it was possible to establish Pyramid rankings for 14.2% of the approximately 19,000 sentences in the twenty document collections which were rated by participants in DUC 2005.

We use *sentence* and *line* interchangeably from this point on in the discussion. Without question non-sentential material, text and otherwise, is present in source documents. Equally without question is the greater part of the text of those documents composed of well-formed sentences.

## 2 Methodology

Although the Pyramid viewer/annotation tool is written in Java, Perl seemed more suited to the entirely automatic process envisioned here. In addition the framework that ties together the various subsystems which make up our summarization system is written in Perl. Using that language immediately makes available our existing resources, in particular our sentence-oriented model of each topic's document collection (Copeck and Szpakowicz, 2003).

The actual process involved was as straightforward as one might expect. For a given topic, each evaluated summary (.pan file) was decomposed into its individual sentences and each of these was marked with the number of SCUs it realized. For three of the 25 DUC participants, <LINE> tags were manually inserted into the summary text body to break it into the individual lines necessary to make an attempt at matching

136:2(2) Former Norwalk City Administrator William H. Kraus was sentenced in federal court in San Diego to five years' probation and fined $1,000 for his part in a land-fraud scheme that bilked investers out of more than $3 million.

147:2(1) Special Assistant U.S. Atty. Michael R. Pent, the prosecutor in the case, said Kraus received a relatively light sentence because he played a limited role in the fraud ring and because he agreed to cooperate with prosecutors.

170:3(1) A federal judge, bluntly rejecting claims that the government's prosecution was politically motivated, sentenced political extremist Lyndon H. LaRouche Jr. to 15 years in prison on mail and tax fraud convictions Friday and refused to release him on bail while he appeals.

Figure 3 : Source Document Sentences Annotated with SCU Information

possible. In almost every instance the usual cues of punctuation and capitalization gave a clear and obvious indication where to insert the break.

The marking of SCUs was accomplished by reassembling the summary text from its individual sentences and recording each sentence's offset from the beginning of that text after correcting for newlines. The link between a summary and a SCU is represented in the .pyr file by the following XML element

```
...
<contributor label="a 20-year dispute">
  <part label="a 20-year dispute"
   start="308" end="325" />
  ...
  </contributor>
```

whose `end` attribute indicates the offset of the end of the `part` substring in the text body of the summary counted from its beginning. We identified the summary sentence by determining which one spanned the endpoint of a SCU contributor's first part. This first stage of processing propagated the SCU counts from the pyramid to the summary text.

Our model of the source document collection sentences was then loaded into a hash table. The `amatch` function in the Perl `String::Approx` program module was used to ascertain exact or nearly exact agreement between lines in the summary which were linked to a SCU and those in the document collection model. `amatch` was set to ignore case and to allow 25% editing, that is, to allow up to one-quarter of the characters in the compared strings to vary by insertion, deletion or substitution. Because the `amatch` function is directed, candidates which satisfied this first test were then required to match in the reverse direction. This ensured that short summary lines were not matching small substrings of lines in the document collection. Manual review of the results showed that the fairly relaxed setting of allowing 25% variation was not being defeated by the data; in every case save one[1], a single hit for each summary sentence was found in the collection, including source sentences which had punctuation and discourse particles elided before being added to the summary. The appearance of such edited sentences in summaries argues for leaving settings as loose as possible, so long as accuracy is not compromised.

This approximate matching operation propagated summary sentence SCU counts back to the source document, which was our objective. Processing concluded by writing out results to file: a listing of all sentence strings in the document collection linked to a SCU together with their index in the collection sequence of sentences, the number of SCUs each realized, and the number of summaries which employed the sentence. Figure 3 shows an example of this output. Its first line indicates that sentence #136 in the document collection (for topic #D695, not indicated here) realizes two SCUs and was employed in two summaries.

SCU weights are the bracketed numbers that appear at the beginning of each SCU label as displayed in the Pyramid viewer. They indicate how many instances in reference summaries support identification of that particular SCU. At this time neither SCU weights nor SCU IDs have been carried back to the source collection sentences, though it would not be difficult to

[1] In the one test case where a summary sentence had two matches in the document collection, the matches appeared in successive reports from the same periodical in which the author of the second report reproduced an earlier statement almost unchanged from its first publication.

do so. The final stage of processing computed the subtotals and grand totals reported in tables below.

## 3 Data

Following the submission of summaries to DUC 2005, participants were asked if they would be willing to use the Pyramid evaluation system to assess the set of summaries on a single topic. In return their summaries would be included in the sets provided to others to evaluate. Most volunteered to perform this task, 25 out of 32 participants. Two manually-written summaries were included in each set as controls, and their authors appear each to have also rated a set of summaries. The number of topics this year is fifty, by chance exactly double the number of Pyramid volunteers.

A number of topics were rated by more than one participant: five were rated twice and one three times. If this multiple rating was meant as a check on rater consistency, results are mixed but generally positive—27% of 208 rating pair SCU counts in the raw data agree, 40% differ by one SCU, and the remaining 1/3 differ by 2 or more SCUs.

In the event, 27 different summaries on each of twenty different topics were rated, 729 ratings of 540 unique summaries. This constitutes about one-third of the total 32 summaries on each of fifty topics submitted to DUC by participants this year.

## 4 Results

Table 1 recapitulates results for the twenty unique topics rated using the Pyramid system. Averages appear for those topics rated more than once. Its first row shows that for topic D311I, the twenty-seven summaries totalled 228 unique sentences, 185 of which were associated with one or more SCUs. 145 of these linked sentences were identified in the 1012 sentences in its document collection.

The left-hand TOTAL column shows the effect of a 250-word ceiling. Regardless of the number of sentences in the document collection, the set of summaries does not amount to more than 230~270 sentences. The count of sentences annotated with one or more SCUs gives some indication of the difficulty of summarizing the particular topic (or perhaps the pertinence of the documents chosen to communicate it). SCU-marked sentences range from 21% (D671G) to 81% (D311I) of all summary sentences. There is much less variability in the degree to which summary sentences with SCUs could be identified in the source document collection: MATCH ranges from 64% (D671G) to 85% (D413A). Note that results in this column are entirely at the mercy of the algorithm used to identify sentence breaks. The `amatch` function does not span lines, so the number of matches found will fall off to the extent that our sentence boundaries differ from those used by any other DUC participant.

There are two offsetting reasons why these figures misrepresent the actual number of source document sentences of interest. First, the figures reported here are counts of matching instances found in the summaries, rather than of unique sentences in source documents, which overstates the number of individual matching sentences. Many summaries use the same sentences (see the first line of Figure 3, where two summaries

| TOPIC | SENTENCES | | | |
|---|---|---|---|---|
| | SUMMARIES | | | DOCS |
| | TOTAL | SCUs | MATCH | TOTAL |
| D311I | 228 | 185 | 145 | 1012 |
| D324E | 258 | 204 | 168 | 359 |
| D345J | 240 | 144 | 110 | 577 |
| D366I | 263 | 104 | 81 | 1793 |
| D376E | 266 | 210 | 163 | 474 |
| D391H | 269 | 141 | 110 | 1833 |
| D393F | 236 | 116 | 95 | 937 |
| D400B | 247 | 126 | 103 | 1057 |
| D407B | 256 | 130 | 101 | 1032 |
| D413A | 252 | 144 | 122 | 607 |
| D422C | 253 | 90 | 66 | 476 |
| D426A | 253 | 180 | 148 | 1279 |
| D431H | 258 | 107 | 74 | 1092 |
| D435F | 242 | 148 | 114 | 1171 |
| D632I | 271 | 63 | 50 | 826 |
| D633G | 262 | 102 | 75 | 869 |
| D654F | 274 | 133 | 103 | 1219 |
| D671G | 254 | 53 | 34 | 624 |
| D683J | 246 | 162 | 128 | 698 |
| D695C | 248 | 115 | 87 | 1038 |
| **TOTAL** | **5076** | **2655** | **2075** | **18973** |
| ***%*** | | | ***10.9%*** | |

Table 1: Instances from all Summaries by Topic: Total Sentences, Sentences Linked to SCUs, Linked Sentences Found in Source

| | SUMMARY SENTENCES | | |
|---|---|---|---|
| # | TOTAL | SCUs | MATCH |
| AUTO | 9.2 | 4.7 | 4.3 |
| MANUAL | 12.1 | 10.3 | 0.2 |
| *%* | *\ /* | *\ /* | |
| *AUTO* | *51%* | *92%* | *47%* |
| *MANUAL* | *85%* | *2%* | *1%* |

Table 2: Counts and Percentages, by Authorship

share the first sentence). Second, though an argument could be made to do so from the outset in order to provide negative examples, initially we did not try to identify source sentences which match no SCU. Since on average only about half of summary sentences do match a SCU, we expected the total number of instance matches would approximately double when all summary sentences were sought in the source document collections. As it turned out, the actual total was 20.3%.

Average sentence counts and percentages for the 25 automated systems and the two manually-written summaries are broken out in the upper section of Table 2. The data shows manually-written summaries tend to be longer than ones produced automatically and, as one would expect, tend to have notably more SCU annotations. Human authors compose their summary texts in a way informed by their understanding of the ideas involved in the topic rather than by selecting appropriate sentences from the document collection, so we expect to find few if any of their summary sentences in source documents. The very small number of hits that do occur are false positives.

The lower section of Table 2 shows the percentages of pairs of adjoining columns in the upper section. In its first row, the count of 4.7 SCU-linked sentences is 51% of 9.2 total sentences. The greyed block in the lower right hand corner shows matched sentences as a percentage of total sentences.

Tables 1 and 2 report counts of sentence instances from the perspective of the summary where repeated selection is not an issue. As one would expect, an appreciable number of sentence instances across the set of summaries identify the same source document sentence: more than one system picked the same sentence in the text. The percentages of summary sentences found in the source document collection previously mentioned are thus not very meaningful. Figure 4 details the extent of repetition. This figure shows the number of instance and unique SCU-marked source document sentences selected by one or more of the 27 different summaries rated with Pyramids. Figure 4 indicates that although the majority of individual source document sentences appeared in a single summary, an appreciable number were chosen by two, three, or more systems. In fact when counts are extended for their number of occurrences, only 36% of all SCU-marked summary sentence instances are identifications of a single source document sentence.

Table 3 summarizes results from the perspective of the document collection. It differs from Tables 1 and 2 in counting unique individual sentences in the source document collection rather than instances of sentences in summaries, which often do repeat each other. It also reports figures for summary sentences which do not match a SCU. Its statistics are therefore more useful to any future user of this SCU-marked corpus.

Each pair of the table's middle columns give counts and percentages for the unique number of summary sentences matched in the topic document collection. The left-hand pair records sentences which have a SCU annotation; the right-hand pair, the total number of unique matched sentences. The right-hand pair thus add sentences which were not marked with a SCU, ones explicitly determined by the annotator to be off-topic and therefore plausible negative examples for learning summarization. In passing we might note that the 10.9% of summary sentences matched in source

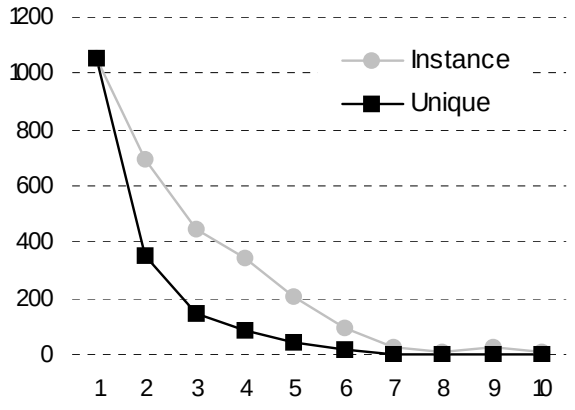


Figure 4: Number of Instance and Unique Sentences Selected Repeatedly, by Repetition

| TOPIC | DOCUMENT SENTENCES | | | | TOTAL |
|---|---|---|---|---|---|
| | WITH SCUs | | ALL MATCHES | | |
| | # | % | # | % | |
| D311I | 106 | 10% | 132 | 13% | 1012 |
| D324E | 71 | 20% | 103 | 29% | 359 |
| D345J | 63 | 11% | 114 | 20% | 577 |
| D366I | 43 | 2% | 148 | 8% | 1793 |
| D376E | 90 | 19% | 134 | 28% | 474 |
| D391H | 77 | 4% | 166 | 9% | 1833 |
| D393F | 51 | 5% | 121 | 13% | 937 |
| D400B | 57 | 5% | 136 | 13% | 1057 |
| D407B | 54 | 5% | 137 | 13% | 1032 |
| D413A | 61 | 10% | 131 | 22% | 607 |
| D422C | 31 | 7% | 106 | 22% | 476 |
| D426A | 94 | 7% | 146 | 11% | 1279 |
| D431H | 51 | 5% | 128 | 12% | 1092 |
| D435F | 68 | 6% | 134 | 11% | 1171 |
| D632I | 30 | 4% | 136 | 16% | 826 |
| D633G | 37 | 4% | 130 | 15% | 869 |
| D654F | 65 | 5% | 164 | 13% | 1219 |
| D671G | 15 | 2% | 120 | 19% | 624 |
| D683J | 87 | 12% | 150 | 21% | 698 |
| D695C | 65 | 6% | 141 | 14% | 1038 |
| **TOTAL** | **1216** | **6.4%** | **2677** | **14.1%** | **18973** |

Table 3: Unique Sentences from All Summaries: Counts and Percentages Found in Source, With and Without SCU Annotation, by Topic

documents (Table 1) drops to 6.4% when rendered unique, but increases to 14.1% when negative examples are included. The approximate doubling factor that we hypothesized holds thus both for instances and individual sentences.

Inspection of Table 3 again shows the effect of two limits: of 250 words in a summary, and of only 27 summary subjects to study. While the number of SCU-marked and total matches is generally proportionate to the size of the document collection, it appears bounded in the same way that we earlier saw sentence instances limited to the range of 230~270 sentences.

## 5 Discussion

The most important point about this study is that we do not propose it to be in any way comprehensive. A crucial fact is that it treats only that third of the DUC 2005 summaries which were rated using the Pyramid system. In addition, for those 20 topics that were so rated, it is quite reasonable to believe that assessing 27 summaries per topic does not exhaust all plausible candidate sentences in the various document collections, many of which include well over a thousand sentences. Were more summaries available, more source document sentences would undoubtedly be ranked. Additional evidence for this belief can be found by the indication in Figure 4 that most source document sentences were selected for only one summary. We expect that, if the search space could be explored more fully through the processing of more summaries, greater numbers of suitable sentences would be picked two, three or more times and that additional, currently unchosen, sentences would be picked at least once.

Since any system which composes summaries automatically by selecting document collection sentences is likely to employ this strategy universally, the reader might wonder why eight percent of the sentences in automatically-generated summaries could not be located in the source document collection. We can identify three reasons why this might happen. First, our model of the document collection does not include titles and cannot match instances where a system adopts a document title as a sentence. Second, giving preference to phrasal and shorter elements[2] in summaries is to pick 'sentences' which are less like likely to be identified identically by our sentence boundary detection algorithm. Successful matching should therefore fall off in such cases.

These two factors were in fact observed in the summaries we inspected manually in the course of processing our pyramid topic. A final reason which is less easy to identify is general processing failure: our sentence breaks may not agree with those of other systems, or the sentence in the summary may be edited to such a degree that the `amatch` string matching function cannot ascertain it to be identical with any source document sentence.

Creating a good summary is more complicated than simply stringing together 250 words from the sentences which are given the highest ranking. For example, sentence length is an obvious second factor at play in summary creation—two short sentences may link to more SCUs than a single long one. Accordingly, some

[2] Distinct phrasal elements can appear in a source document when it includes non-sentential material such as tables.

measure of information density is called for. The reader is directed to the Nenkova and Passonneau (2004) paper for a discussion of theoretical issues which the designers of the Pyramid system considered when deciding how to compute a summary score. Use of the ranked corpora we have produced does not however oblige one to adopt any particular scoring formula, and such formulae no doubt are as varied as are summarization systems themselves.

The most obvious use of a corpus of sentences ranked on their suitability to a generic summary is to provide training data for any system that incorporates a learning capability, and for human system designers otherwise. While that is the main use to which we expect to put the corpus, limited additional work in future will involve expanding it by adding those fields identified in the Methodology section as candidates for inclusion.

DUC participants who wish to make use of the SCU-marked corpus described here are invited to contact the authors.

## Acknowledgements

We would like to thank Dr. Diana Inkpen for her helpful comments on the paper. This work has been supported by the Natural Sciences and Engineering Research Council of Canada.

## References

Copeck, Terry and Stan Szpakowicz. 2003. Picking Phrases, Picking Sentences. In Proceedings of the Workshop on Automatic Summarization (DUC 2003), HLT/NAACL-2003.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating Content Selection in Summarization: the Pyramid Method. HLT/NAACL-2004, 145-152.