# Vocabulary Usage in Newswire Summaries

**Terry COPECK**
School of IT and Engineering
University of Ottawa
Ottawa, Ontario Canada
terry@site.uottawa.ca

**Stan SZPAKOWICZ**
School of IT and Engineering
University of Ottawa
Ottawa, Ontario Canada
szpak@site.uottawa.ca

**Abstract**

Analysis of 9000 manually written summaries of newswire stories used in four Document Understanding Conferences indicates that approximately 40% of their lexical items do not occur in the source document. A further comparison of different summaries of the same document shows agreement on 28% of their vocabulary. It can be argued that these relationships establish a performance ceiling for automated summarization systems which do not perform syntactic and semantic analysis on the source document.

## 1 Introduction

Automatic summarization systems rely on manually prepared summaries for training data, heuristics and evaluation. Generic summaries are notoriously hard to standardize; biased summaries, even in a most restricted task or application, also tend to vary between authors. It is unrealistic to expect one perfect model summary, and the presence of many, potentially quite diverse, models introduces considerable uncertainty into the summarization process. In addition, many summarization systems tacitly assume that model summaries are somehow close to the source documents.

We investigate this assumption, and study the variability of manually produced summaries. We first describe the collection of documents with summaries which has been accumulated over several years of participation in the Document Understanding Conference (DUC) evaluation exercises sponsored by the National Institute of Science and Technology (NIST). We then present our methodology, discuss the rather pessimistic results, and finally draw a few simple conclusions.

## 2 The Corpus

### 2.1 General Organization

The authors have assembled a corpus of manually written summaries of texts from their archive of materials provided to participants in the DUC conferences, held annually since 2001. It is available at the DUC Web site to readers who are qualified to access the DUC document sets on application to NIST. To help interested parties assess it for their purposes we provide more detail than usual on its organization and contents.

Most summaries in the corpus are *abstracts*, written by human readers of the source document to best express its content without restriction in any manner save length (words or characters). One method of performing automatic summarization is to construct the desired amount of output by concatenating representative sentences from the source document, which reduces the task to one of determining most adequately what 'representative' means. Such summaries are called *extracts*. In 2002, recognizing that many participants summarize by extraction, NIST produced versions of documents divided into individual sentences and asked its author volunteers to compose their summaries similarly. Because we use a sentence-extraction technique in our summarization system, this data is of particular interest to us. It is not included in the corpus being treated here and will be discussed in a separate paper.

The DUC corpus contains 11,867 files organized in a three-level hierarchy of directories totalling 62MB. The top level identifies the source year and exists simply to avoid the name collision which occurs when different years use same-named subdirectories. The middle 291 directories identify the *document clusters*; DUC reuses collections of newswire stories assembled for the TREC and TDT research initiatives which report on a common topic or theme. Directories on the lowest level contain SGML-tagged and untagged versions of 2,781 individual source documents, and between one and five

| | DOCUMENTS | | | | | SUMMARIES | | | | | D : S |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **10** | **50** | **100** | **200** | **?** | **10** | **50** | **100** | **200** | **?** | |
| **2001** | | 28 | 316 | 56 | **400** | | 84 | 946 | 168 | **1198** | 1 : 3 |
| **2002** | 59 | 59 | 626 | 59 | **803** | 116 | 116 | 1228 | 116 | **1576** | 1 : 2 |
| **2003** | 624 | | 90 | | **714** | 2496 | | 360 | | **2856** | 1 : 4 |
| **2004** | 740 | | 124 | | **864** | 2960 | | 496 | | **3455** | 1 : 4 |
| **?** | **1423** | **87** | **1156** | **115** | **2781** | **5572** | **200** | **3030** | **284** | **9086** | 1 : 3 |

Table 1: Number of documents and summaries by size and by year, and ratios

summaries of each, 9,086 summaries in total. In most cases the document involved is just that: a single news report originally published in a newspaper. 552 directories, approximately 20% of the corpus, represent *multi-document* summaries—ones which the author has based on all the files in a cluster of related documents. For these summaries a source document against which to compare them has been constructed by concatenating the individual documents in a cluster into one file. Concatenation is done in directory order, though the order of documents does not matter here.

### 2.2 The Corpus in Detail

The Document Understanding Conference has evolved over the four years represented in our corpus, and this is reflected in the materials which are available for our purposes. Table 1 classifies these files by year and by target size of summary; the rightmost column indicates the ratio of summaries to source documents, that is, the average number of summaries per document. Totals appear in bold. The following factors of interest can be identified in its data:

- **Size**. Initially DUC targeted summaries of 50, 100 and 200 words. The following year 10-word summaries were added, and in 2003 only 10- and 100-word summaries were produced;
- **Growth**. Despite the high cost of producing manual summaries, the number of documents under consideration has *doubled* over the four years under study while the number of summaries has *tripled*;
- **Ratio**. On average, *three* manual summaries are available for each source document;
- **Formation**. While longer summaries are routinely composed of well-formed sentences, subsentential constructs such as *headlines* are acceptable 10-word summaries, as are *lists* of key words and phrases.
- **Author**. Although the 2004 DUC source documents include *machine translations* of foreign language news stories, in each case a parallel human translation was available. Only source documents written or translated by human beings appear in the corpus.

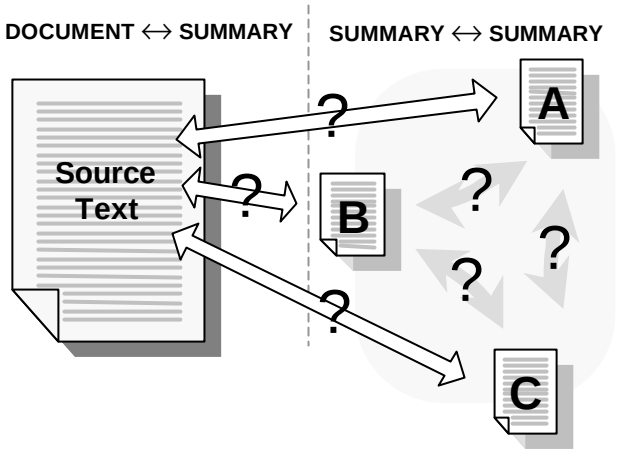


Figure 1: Files and relationships investigated

## 3 The Evaluation Model

Figure 1 shows the typical contents of a third-level source document directory. Relations we wish to investigate are marked by arrows. There are two: the relationship between the vocabulary used in the source document and summaries of it, and that among the vocabulary used in summaries themselves. The first is marked by white arrows, the second by grey.

The number of document-summary relations in the corpus is determined by the larger cardinality set involved, which here is the number of summaries: thus 9,086 instances. For every document with N summaries, we consider all C(N, 2) pairs of summaries. In total there are 11,441 summary-summary relationships.

We ask two questions: *to what degree do summaries use words appearing in the source document?* and, *to what degree do different summaries use the same vocabulary?*

### 3.1 Measures

To answer our two questions we decided to compute statistics on two types of elements of each pair of test documents: their *phrases*, and ultimately, their

individual *tokens*. Phrases were extracted by applying a 987-item stop list developed by the authors (Copeck and Szpakowicz 2003) to the test documents. Each collocation separated by stop words is taken as a phrase[1]. Test documents were tokenized by breaking the text on white space and trimming off punctuation external to the token. Instances of each sort of item were recorded in a hash table and written to file.

Tokens are an obvious and unambiguous baseline for lexical agreement, one used by such summary evaluation systems as ROUGE (Lin and Hovy, 2003). On the other hand, it is important to explain what we mean by units we call *phrases*; they should not be confused with syntactically correct constituents such as noun phrases or verb phrases. Our units often are not syntactically well-formed. Adjacent constituents not separated by a stop word are *unified*, single constituents are *divided* on any embedded stop word, and those composed entirely of stop words are simply *missed*.

Our phrases, however, are not n-grams. A 10-word summary has precisely 9 bigrams but, in this study, only 3.4 phrases on average (Table 2). On the continuum of grammaticality these units can thus be seen as lying somewhere between generated blindly n-grams and syntactically well-formed phrasal constituents. We judge them to be weakly syntactically motivated[2] and only roughly analogous to the *factoids* identified by van Halteren and Teufel (2003) in the sense that they also express semantic constructs. Where van Halteren and Teufel identified factoids in 50 summaries, we sacrificed accuracy for automation in order to process 9000.

We then assessed the degree to which a pair of documents for comparison shared vocabulary in terms of these units. This was done by counting matches between the phrases. Six different kinds of match were identified and are listed here in what we deem to be decreasing order of stringency. While the match types are labelled and described in terms of summary and source document for clarity, they apply equally to summary pairs. Candidate phrases are underlined and matching elements tinted in the examples; headings used in the results table (Table 2) appear in SMALL CAPS.

- **Exact match**. The most demanding, requires candidates agree in all respects. EXACT

  `after Mayo Clinic stay` ↔
  `Mayo Clinic group`

- **Case-insensitive exact match** relaxes the requirement for agreement in case. EXACT CI

  `concerning bilateral relations` ↔
  `Bilateral relations with`

- **Head of summary phrase in document** requires only that the head of the candidate appear in the source document phrase. The head is the rightmost word in a phrase. HEAD DOC

  `calls Sharon disaster` ↔
  `deemed tantamount to disaster`

- **Head of document phrase in summary** is the previous test in reverse. HEAD SUM

- **Summary phrase is substring of document phrase.** True if the summary phrase appears anywhere in the document phrase. SUB DOC

  `has identified Iraqi agent as` ↔
  `the Iraqi agent defection`

- **Document phrase is substring of summary phrase** reverses the previous test. SUB SUM

Tests for matches between the tokens of two documents are more limited because only single lexical items are involved. Exact match can be supplemented by case insensitivity and by stemming to identify any common root shared by two tokens. The Porter stemmer was used.

The objective of all these tests is to capture any sort of meaningful resemblance between the vocabularies employed in two texts. Without question, additional measures can and should be identified and implemented to correct, expand, and refine the analysis.

### 3.2 Methodology

The study was carried out in three stages. A *pre-study* determined the "lie of the land"—what the general character of results was likely to be, the most appropriate methodology to realize them, and so on. In particular this initial investigation alerted us to the fact that so few phrases in any two texts under study matched exactly as to provide little useful data, leading us to add more relaxed measures of lexical agreement. This initial investigation made it clear that there was no point in attempting to find a subset of vocabulary used in a number of summaries—it would be vanishingly small—and we therefore confined ourselves in the main study to pairwise comparisons. The pre-study also suggested that summary size would be a significant factor in lexical agreement while source document size

[1] When analysis of a summary indicated that it was a list of comma- or semicolon-delimited phrases, the phrasing provided by the summary author was adopted, including any stopwords present. *Turkey attacks Kurds in Iraq, warns Syria, accusations fuel tensions, Mubarak intercedes* is thus split into four phrases with the first retaining the stopword *in*. There are 453 such summaries.

[2] While the lexical units in question might be more accurately labelled *syntactically motivated ngrams*, for simplicity we use *phrase* in the discussion.

```
AFA19981230.1000.0058:  X <> W  exact: 2, exactCI: 2, partSum2: 2, partSum1 2, token-
    Match: 6
    X: Jordanian King Hussein to meet with Clinton concerning bilateral relations
    W: King Hussein to meet with Clinton after visiting Mayo Clinic
  2 exact:  meet,Clinton
  2 exactCI:  meet,clinton
  2 headSum1:  clinton,meet
  2 headSum2:  meet,clinton
  6 tokMatch:  hussein,meet,clinton,to,king,with
```

Figure 2: Text and matches for two summaries of AFA19981230.1000.0058

would be less so, indications which were not entirely borne out by the strength of the results ultimately observed.

The main *study* proceeded in two phases. After the corpus had been organized as described in Section 2 and untagged versions of the source documents produced for the analysis program to work on, that process traversed the directory tree, decomposing each text file into its phrases and tokens. These were stored in hash tables and written to file to provide an audit point on the process. The hash tables were then used to test each pair of test documents for matches—the source document to each summary, and all combinations of summaries. The resulting counts for all comparisons together with other data were then written to a file with results, one line per source document in a comma-delimited format suitable for importation to a spreadsheet program.

The second phase of the main study involved organizing the spreadsheet data into a format permitting the calculation of statistics on various categorizations of documents they describe. Because the source document record was variable-length in itself and also contained a varying number of variable-length sub-records of document pair comparisons, this was a fairly time-consuming clerical task. It did however provide the counts and averages presented in Table 2 and subsequently allowed the user to recategorize the data fairly easily.

A *post-study* was then conducted to validate the computation of measures by reporting these to the user for individual document sets, and applied to a

**DOCUMENT - SUMMARY**

| | SUMMARY | | | PHRASES | | | | | | TOKENS | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | COUNT | TOKENS | PHRASES | EXACT | EXACT CI | HEAD DOC | HEAD SUM | SUB DOC | SUB SUM | EXACT | STEM CI |
| 10 | 5572 | 10.0 | 3.4 | 0.8 | 1.0 | 1.4 | 0.9 | 2.3 | 2.7 | 5.4 | 6.3 |
| 50 | 200 | 47.4 | 15.5 | 5.5 | 5.7 | 8.8 | 4.9 | 11.8 | 12.0 | 30.6 | 32.6 |
| 100 | 3030 | 95.6 | 30.5 | 12.1 | 12.5 | 14.9 | 10.1 | 22.3 | 20.5 | 52.7 | 54.8 |
| 200 | 284 | 157.5 | 48.6 | 19.7 | 20.4 | 28.3 | 17.1 | 38.4 | 35.3 | 82.9 | 85.8 |
| ALL | **9086** | **44.0** | **14.1** | **5.2** | **5.5** | **6.9** | **8.4** | **10.3** | **28.2** | **24.2** | **25.5** |
| 10 | | | | 22% | 29% | 43% | 27% | 69% | 79% | 55% | 63% |
| 50 | | | | 35% | 37% | 57% | 31% | 76% | 77% | 65% | 69% |
| 100 | | | | 39% | 41% | 49% | 34% | 78% | 74% | 55% | 58% |
| 200 | | | | 40% | 42% | 56% | 35% | 79% | 73% | 51% | 53% |
| ALL | | | | **37%** | **39%** | **49%** | **33%** | **73%** | **70%** | **55%** | **58%** |

**SUMMARY - SUMMARY**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 8241 | 10.0 | 3.4 | 0.17 | 0.21 | 0.24 | 0.24 | | | 2.82 | 3.13 |
| 50 | 141 | 47.4 | 15.5 | 0.71 | 0.84 | 1.09 | 1.06 | | | 10.89 | 11.77 |
| 100 | 2834 | 95.6 | 30.5 | 4.21 | 4.39 | 4.76 | 4.82 | | | 28.16 | 29.66 |
| 200 | 225 | 157.5 | 48.6 | 4.26 | 4.52 | 6.24 | 5.93 | | | 35.16 | 37.14 |
| ALL | **11441** | **44.0** | **14.1** | **1.26** | **1.34** | **1.5** | **1.5** | | | **9.8** | **10.48** |
| 10 | | | | 5% | 6% | 7% | 7% | | | 28% | 31% |
| 50 | | | | 5% | 5% | 7% | 7% | | | 23% | 25% |
| 100 | | | | 14% | 14% | 16% | 16% | | | 29% | 31% |
| 200 | | | | 9% | 9% | 13% | 12% | | | 22% | 24% |
| ALL | | | | **9%** | **10%** | **11%** | **11%** | | | **22%** | **24%** |

Table 2: Counts and percentages of vocabulary agreement, by size and total

small random sample of text pairs. Figure 2 shows the comparison of two summaries of source document AFA19981230.1000.0058. A secondary objective of the post-study was to inspect the actual data. Were there factors in play in the data that had escaped us? None were made evident beyond the all-too-familiar demonstration of the wide variety of language use in play. The log file of document phrase hash tables provided an additional snapshot of the kind of materials with which the automated computation had been working.

## 4 Results

### 4.1 Data Averages

Table 2 illustrates the degree to which summaries in the DUC corpus employ the same vocabulary as the source documents on which they are based and the degree to which they resemble each other in wording. The table, actually a stack of four tables which share common headings, presents data on the document-summary relationship followed by inter-summary data, giving counts and then percentages for each relationship. Statistics on the given relationship appear in the first three columns on the left; counts and averages are classified by summary size. The central group of six columns presents from left to right, in decreasing order of strictness, the average number of phrase matches found for the size category. The final two columns on the right present parallel match data for tokens. Thus for example the column entitled STEM CI shows the average number of stemmed, case-insensitive token matches in a pair of test documents of the size category indicated. Each table in the stack ends with a boldface row that averages statistics across all size categories.

Inspection of the results in Table 2 leads to these general observations:

- With the exception of 200-word summaries falling somewhat short (157 words), each category approaches its target size quite closely;
- Phrases average three tokens in length regardless of summary size;
- The objective of relaxing match criteria in the main study was achieved. With few exceptions, each less strict match type produces more hits than its more stringent neighbors;

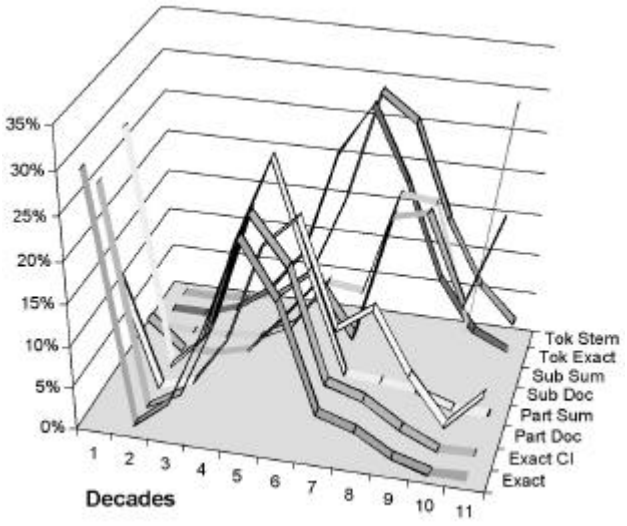


Figure 3: Percentages of summary vocabulary agreement for all source documents, by measure

- The much smaller size of the now discontinued 50- and 200-word categories argues against investing much confidence in their data;
- Finally, while no effect was found for source document size (and results for that categorization are therefore not presented), the percentage tables suggest summary size has some limited impact on vocabulary agreement. This effect occurs solely on the phrasal level, most strongly on its strictest measures; token values are effectively flat.

We are uncertain why this last situation is so. Consider only the well-populated 10-word and 100-word summary classes. The effect cannot be accounted for a preponderance of multiple-document summaries in either class which might provide more opportunities for matches. Despite many more of these being among the 100-word summaries than the 10-word (1974 single : 1056 multi, versus 116 single : 5456 multi), the percentage of exact phrasal matches is essentially the same in each subcategorization of these classes.

We speculate that authors may compose the sentences in 100-word summaries in terms of phrases from the source document, while 10-word summaries, which more closely resemble terse headlines, cannot be composed by direct reuse of source document phrases. 50- and 200-word summaries are also composed of sentences. Their exact match percentages approach those of 100-word summaries, lending support to this interpretation.

### 4.2 Data Variance

Whether count or percentage, exclusively average data is presented in Table 2. While measures of central tendency are an important dimension of any population, a full statistical description also requires some indication of measures of variance. These appear in Figure 3 which shows, for each of the six phrasal and two token measures, what percentage of the total number of summaries falls into each tenth of the range of possible values. For example, a summary in which 40% of the phrases were exactly matched in the source document would be represented in the figure by the vertical position of the frontmost band over the extent of the decade labeled '4'—24%. The figure's three-dimensional aspect allows the viewer to track which decades have the greatest number of instances as measures move from more strict to more relaxed, front to back.

However, the most striking message communicated by Figure 3 is that large numbers of summaries have zero values for the stricter measures, EXACT, EXACT CI and PART SUM in particular and PART DOC to a lesser degree. These same measures have their most frequent values around the 50% decade, with troughs both before and after. To understand why this is so requires some explanation. Suppose a summary contains two phrases. If none are matched in the source its score is 0%. If one is matched its score is 50%; if both, 100%. A summary with three phrases has four possible percentage values: 0%, 33%, 66% and 100%. The 'hump' of partial matching is thus around the fifty percent level because most summaries are ten words, and have only 1 or 2 candidates to be matched. The ranges involved in the stricter measures are not large.

That acknowledged, we can see that the modal or most frequent decade does indeed tend in an irregular way to move from left to right, from zero to 100 percent, as measures become less strict. In making this observation, note that the two backmost bands represent measures on tokens, a different syntactic element than the phrase. The information about the distribution of summary measures shown in this figure is not unexpected.

### 4.3 Key Findings

The central fact that these data communicate quite clearly is that summaries do not employ many of the same phrases their source documents do, and even fewer than do other summaries. In particular, on average only 37% of summary phrases appear in the source document, while summaries share only 9% of their phrases. This becomes more understandable when we note that on average only 55% of the individual words used in summaries, both common vocabulary terms and proper names, appear in the source document; and between summaries, on average only 22% are found in both.

It may be argued that the lower counts for inter-summary vocabulary agreement can be explained thus: since a summary is so much smaller than its source document, lower counts should result. One reply to that argument is that, while acknowledging that synonymy, generalization and specialization would augment the values found, the essence of a generic summary is to report the pith, the gist, the central points, of a document and that these key elements should not vary so widely from one summary to the next.

## 5 Pertinent Research

Previous research addressing summary vocabulary is limited, and most has been undertaken in connection with another issue: either with the problem of evaluating summary quality (Mani, 2001; Lin and Hovy, 2002) or to assess sentence element suitability for use in a summary (Jing and McKeown, 1999). In such a case results arise as a by-product of the main line of research and conclusions about vocabulary must be inferred from other findings.

Mani (2001) reports that "previous studies, most of which have focused on extracts, have shown evidence of low agreement among humans as to which sentences are good summary sentences." Lin and Hovy's (2002) discovery of low inter-rater agreement in single (~40%) and multiple (~29%) summary evaluation may also pertain to our findings. It stands to reason that individuals who disagree on sentence pertinence or do not rate the same summary highly are not likely to use the same words to write the summary. In the very overt rating situation they describe, Lin and Hovy were also able to identify human error and quantify it as a significant factor in rater performance. This reality may introduce variance as a consequence of suboptimal performance: a writer may simply fail to use the *mot juste*.

In contrast, Jing, McKeown, Barzilay and Elhadad (1998) found human summarizers to be 'quite consistent' as to what should be included, a result they acknowledge to be 'surprisingly high'. Jing *et al.* note that agreement drops off with summary

length, that their experience is somewhat at variance with that of other researchers, and that this may be accounted for in part by regularity in the structure of the documents summarized.

Observing that "expert summarizers often reuse the text in the original document to produce a summary" Jing and McKeown (1999) analyzed 300 human written summaries of news articles and found that "a significant portion (78%) of summary sentences produced by humans are based on cut-and-paste", where 'cut-and-paste' indicates vocabulary agreement. This suggests that 22% of summary sentences are not produced in this way; and the authors report that 315 (19%) sentences do not match any sentence in the document.

In their 2002 paper, Lin and Hovy examine the use of multiple gold standard summaries for summarization evaluation, and conclude "we need more than one model summary although we cannot estimate how many model summaries are required to achieve reliable automated summary evaluation".

Attempting to answer that question, van Halteren and Teufel (2003) conclude that 30 to 40 manual summaries should be sufficient to establish a stable consensus model summary. Their research, which directly explores the differences and similarities between various human summaries to establish a basis for such an estimate, finds great variation in summary content as reflected in *factoids*[3]. This variation does not fall off with the number of summaries and accordingly no two summaries correlate highly. Although factoid measures did not correlate highly with those of unigrams (tokens), the former did clearly demonstrate an importance hierarchy which is an essential condition if a consensus model summary is to be constructed. Our work can thus be seen as confirming that, in large measure, van Halteren and Teufel's findings apply to the DUC corpus of manual summaries.

## 6 Discussion

We began this study to test two hypotheses. The first is this: automatic summarization is made difficult to the degree that manually-written summaries do not limit themselves to the vocabulary of the source document. For a summarization system to incorporate words which do not appear in the source document requires at a minimum that it has a capacity to substitute a synonym of some word in the text, and some justification for doing so. More likely it would involve constructing a representation of the text's meaning and reasoning (generalization, inferencing) on the content of that representation. The latter are extremely hard tasks.

Our second hypothesis is that *automatic summarization is made difficult to the degree that manually written summaries do not agree among themselves.* While the variety of possible disagreements are multifarious, the use of different vocabulary is a fundamental measure of semantic heterogeneity. Authors cannot easily talk of the same things if they do not use words in common.

Unfortunately, our study of the DUC manual summaries and their source documents provides substantial evidence that summarization of these documents remains difficult indeed.

## 7 Conclusion

Previous research on the degree of agreement between documents and summaries, and between summaries, has generally indicated that there are significant differences in the vocabulary used by authors of summaries and the source document. Our study extends the investigation to a corpus currently popular in the text summarization research community and finds the majority opinion to be borne out there. In addition, our data suggests that summaries resemble the source document more closely than they do each other. The limited number of summaries available for any individual source document prevents us from learning any characteristics of the population of possible summaries. Would more summaries distribute themselves evenly throughout the semantic space defined by the source document's vocabulary? Would clumps and clusters show themselves, or a single cluster as van Halteren and Teufel suggest? If the latter, such a grouping would have a good claim to call itself a consensus summary of the document and a true gold standard would be revealed.

[3] A factoid is an atomic semantic unit corresponding to an expression in first-order predicate logic. As already noted we approximate phrases to factoids.

## References

Copeck, Terry and Stan Szpakowicz. 2003. Picking phrases, picking sentences. In *DUC Workshop at HLT/NAACL-2003 Workshop on Automatic Summarization.*

Jing, Hongyan, Regina Barzilay, Kathleen McKeown and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *1998 AAAI Spring Symposium on Intelligent Text Summarization, AAAI Technical Report SS-98-06.*

Jing, Hongyan. and Kathleen McKeown. 1999. The decomposition of human-written summary sentences. *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99).*

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003).*

Lin, Chin-Yew and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. *Proceedings of Workshop on Automatic Summarization, 2002 ACL (WAS/ACL-02).*

Mani, Inderjeet. 2001. Summarization evaluation: An overview. *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization.*

Van Halteren, Hans, and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. *Proceedings of Workshop on Automatic Summarization, 2003 Language Technology Conference (WAS/HLT-NAACL-2003).*