Abstract

Since 2001, the Document Understanding Conferences have been the forum for researchers in automatic text summarization to compare methods and results on common test sets. Over the years, several types of summarization tasks have been addressed—single document summarization, multi-document summarization, summarization focused by question, and headline generation. This paper is an overview of the achieved results in the different types of summarization tasks. We compare both the broader classes of baselines, systems and humans, as well as individual pairs of summarizers (both human and automatic). An analysis of variance model is fitted, with summarizer and input set as independent variables, and the coverage score as the dependent variable, and simulation-based multiple comparisons were performed. The results document the progress in the field as a whole, rather than focusing on a single system, and thus can serve as a future reference on the work done up to date, as well as a starting point in the formulation of future tasks.

Methodology and analysis

In order to study the differences between summarizers, we fit a two-way analysis of variance model with the systems and the input as factors and the coverage score as dependent variable. The simple main effect model we use is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

(1)

$Y_{ij}$ is the coverage score for system $i$ for input $j$, and it is equal to $\mu$, a grand mean coverage for any summary, adjusted for the effect of the summarizer ($\alpha_i$) and the effect of the input ($\beta_j$) and some random noise $\epsilon_{ij}$. Other possible sources of variation such as evaluator, or system/evaluator interactions and the like, are not included in the model, since they have been controlled for in the evaluation set up as discussed in (Harman & Over 2004).

For all tasks, both main effects are significant with $p = 0$, which indicates that significant differences between summarizers exists, as well as that some sets or documents are easier to summarize than others (and summarizers get higher coverage scores for them). It is of interest to be able to compare each two summarizers against each other, total...
better than systems

<table>
<thead>
<tr>
<th>sys</th>
<th>better than systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 15 16 17 18 23 25 30</td>
</tr>
<tr>
<td>A</td>
<td>1 15 16 17 18 21 23 25 31</td>
</tr>
<tr>
<td>B</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31</td>
</tr>
<tr>
<td>C</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31</td>
</tr>
<tr>
<td>D</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31</td>
</tr>
<tr>
<td>E</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31</td>
</tr>
<tr>
<td>F</td>
<td>1 15 16 17 18 23 25 29 31</td>
</tr>
<tr>
<td>G</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31</td>
</tr>
<tr>
<td>H</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31 A F J</td>
</tr>
<tr>
<td>I</td>
<td>1 15 16 17 18 19 21 23 25 27 28 29 31</td>
</tr>
<tr>
<td>J</td>
<td>1 15 16 17 18 23 25 29 31</td>
</tr>
<tr>
<td>15</td>
<td>17 25 30</td>
</tr>
<tr>
<td>16</td>
<td>17 30</td>
</tr>
<tr>
<td>18</td>
<td>17 30</td>
</tr>
<tr>
<td>19</td>
<td>15 16 17 18 23 25 30</td>
</tr>
<tr>
<td>21</td>
<td>15 16 17 18 25 30</td>
</tr>
<tr>
<td>23</td>
<td>17 25 30</td>
</tr>
<tr>
<td>25</td>
<td>17 30</td>
</tr>
<tr>
<td>27</td>
<td>15 16 17 18 23 25 30</td>
</tr>
<tr>
<td>28</td>
<td>15 16 17 18 23 25 30</td>
</tr>
<tr>
<td>29</td>
<td>16 17 25 30</td>
</tr>
<tr>
<td>31</td>
<td>16 17 18 25 30</td>
</tr>
</tbody>
</table>

Table 1: Significant differences based on 95% confidence intervals for single document summarizers in 2002. The second column lists the summarizers that the summarizer in the first column is significantly better than.

of \( \binom{N}{2} \) comparisons when \( N \) is the number of summarizers. Thus, the number of pairwise comparisons to be made is quite large, since each year more than 10 automatic summarizers were tested and several human summarizers wrote summaries for comparison purposes. In order to control the probability of declaring a difference between two systems to be significant when it is actually not (Type I error), we use a simulation based multiple comparisons procedure that controls for the overall experiment-wise probability of error. 95% confidence intervals for the differences between system means are computed with simulation size \( = 12616 \). Pairs of summarizers where the confidence interval for the difference between the two means excludes zero can be declared to be significantly different.

**Summarizer codes conventions**

In all years and tasks, human summarizers have letter codes A to K. Baselines have codes 1 to 5. The remaining codes are those for participating systems. Each year, non-model human summarizers were available for each set/document, but no human wrote summaries for all the sets. This means that the average scores for humans are estimated from a smaller sample than those of system averages.

**Generic single-document summarization**

The generic single document summarization task was addressed in DUC 2001 and 2002, with target summary length of 100 words. The baseline in both years, with summarizer code 1, was the same: taking the first 100 words of the input document. Space constraints would not permit to list all 507 pairwise comparisons and the respective 95% confidence interval, so table 1 lists only the pairs of summarizers where a significant difference could be declared in 2002. The trends are similar in 2001, with humans performing better than systems and the baselines, and systems not outperforming the baseline. The peer averages for 2001 are not listed since the coverage evaluation was done on a different scale from the one used in following years and thus results across years cannot be compared to track progress. The peer averages for 2002 are given in table 2.

Both years, none of the systems outperforms the baseline (and the systems as a group do not outperform the baseline) and in fact the baseline has better coverage than most of the automatic systems (see the first row in table 1). It has often been noted that this baseline is indeed quite strong, due to journalistic convention for putting the most important part of an article in the initial paragraphs. But the fact that human summarizers (with the exception of F and J) significantly outperform the baseline shows that the task is meaningful and that better-than-baseline performance is possible. The difference between the worst human and the baseline is 0.1 while the difference between the best system and the baseline is 0.01. Seven out of the nine human summarizers are significantly better than all the systems, but humans F, J and A are not significantly better than all systems. This suggests that while humans as a group are much better than automatic summarizers, specific human summarization strategies are not so good and do not lead to significant advantage over some automatic systems. These facts indicate that while the single-document task has not been kept in later years, the task remains an open problem.

**Generic multi-document summarization**

Generic multi-document summaries were produced in 2001, 2002 and 2004, with target lengths of 50, 100, 200 and 400 words in 2001 and 50, 100, and 200 words in 2002 and 100 words only in 2004. The two baselines used are 1) the first \( n \) words of the most recent document (corresponding to system code 1 in 2001 and code 2 in 2002) and 2) the first sentence from each document in chronological order until the length requirement is reached (codes 2 and 3 in 2001 and 2002 respectively). Note that the first baseline is equivalent to what is currently used in news aggregation systems such as Google News. Automatic summarizers in general perform better than this baseline as can be seen in table 3 and table 4. The peer averages for 2002 and 2004 are given in tables 5 and 6 respectively.

When the target summary size is added as a factor in the analysis of variance model, it is highly significant with \( p < 0.0001 \) and multiple comparisons between the between the four lengths in 2001 show that scores significantly increase as the summary size increases. This indicates that the summary size impacts coverage scores and thus when performing evaluation one should control for the size of summaries that are being compared. If size is not controlled for, it can confound the results in coverage scores comparisons. An interesting questions arises of what would happen...
Table 2: Summarizer code, coverage and number of summaries for the single document summarization task at DUC 2002

| summarizer | 1 15 16 17 18 19 21 23 25 27 28 29 |
| coverage  | .37 .33 .30 .08 .32 .39 .37 .34 .29 .38 .38 .36 |
| sums      | 291 295 295 292 294 295 295 289 294 292 295 294 |

| summarizer | 30 31 | A | B | C | D | E | F | G | H | I | J |
| coverage  | .06   | .36 | .47 | .51 | .46 | .52 | .49 | .47 | .54 | .57 | .53 | .47 |
| sums      | 294   | 292 | 30  | 30  | 24  | 30  | 25  | 25  | 30  | 29  | 30  | 30  |

Table 3: Significant differences for the 2001 multi-document task. The summary length has been added as a factor in the analysis of variance model.

1. Only one system significantly outperforms the baseline of selecting first sentences from the input articles (Id=2).
2. Not all humans significantly outperform all automatic systems. This is very positive and indicates that on a set of 30 input clusters, humans do not perform significantly better.
3. Eight (out of 12) systems outperform baseline 1 (selecting the first \( N \) words of one of the articles in the cluster). This approach is used by news aggregation sites, such as Google news. The DUC results show that overall, specifically developed multi-document summarization systems perform better than such a baseline.
4. In the comparisons between different humans, only two are significant (E and F are both significantly better than H). The fact that there are significant differences between human summarizers suggests that a cognitive study of the human summarization techniques can be useful in shedding light for the development of automatic summarizers.

Figure 1 shows a pair of summaries written on for the same input by humans F and H. The examples give an idea why the unusual summarization strategy of H can lead to worse summaries. Summarizer H also happens to be the human that does not outperform significantly most of the automatic systems.

Table 4 shows the significant pairwise comparisons between multi-document summarizers for 2002. These data also indicate that more pairs are significant when longer summaries are compared. It is again encouraging that several automatic systems are not significantly worse than humans. Table 7 gives the pairs of significantly different summarizers from DUC 2004.

The generic multi-document summarization task is the only one evaluated in the same way across several years. For the 100 word summary in 2002, the best system result is 0.24, while in 2004 it is 0.30, and the average of system coverage was 0.17 and 0.21 respectively. The difference is significant according to a Wilcoxon rank-sum test \((p \text{-value} = 0.02)\). This could be an indication that there is an improvement of systems across the years. Unfortunately, such a conclusion cannot be readily drawn, because as the analysis of variance model shows, the input to a summarizer has a significant impact on the final coverage score and it is not

\(^2\)But given more data, the observed differences could become significant.
all automatic summarizers

all automatic summarizers but 65

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers

all automatic summarizers
### Table 4: Significant differences for multi-document summarizers in 2002 for different summary lengths

<table>
<thead>
<tr>
<th>summarizer</th>
<th>16</th>
<th>19</th>
<th>20</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>28</th>
<th>29</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 wrds</td>
<td>.15</td>
<td>.25</td>
<td>.13</td>
<td>.21</td>
<td>.22</td>
<td>.16</td>
<td>.24</td>
<td>.23</td>
<td>.17</td>
</tr>
<tr>
<td>sums</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>summarizer</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 wrds</td>
<td>.37</td>
<td>.44</td>
<td>.25</td>
<td>.40</td>
<td>.46</td>
<td>.33</td>
<td>.44</td>
<td>.32</td>
<td>.26</td>
<td>.22</td>
</tr>
<tr>
<td>100 wrds</td>
<td>.37</td>
<td>.43</td>
<td>.38</td>
<td>.28</td>
<td>.25</td>
<td>.36</td>
<td>.31</td>
<td>.39</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>200 wrds</td>
<td>.37</td>
<td>.42</td>
<td>.45</td>
<td>.39</td>
<td>.27</td>
<td>.30</td>
<td>.29</td>
<td>.37</td>
<td>.41</td>
<td>.26</td>
</tr>
<tr>
<td>sums</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

### Table 5: Coverage for summarizers at the multi-document task at DUC 2002, for different target summary lengths

<table>
<thead>
<tr>
<th>summarizer</th>
<th>102</th>
<th>11</th>
<th>111</th>
<th>117</th>
<th>120</th>
<th>123</th>
<th>124</th>
<th>138</th>
<th>19</th>
<th>2</th>
<th>27</th>
<th>34</th>
<th>44</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>.24</td>
<td>.22</td>
<td>.05</td>
<td>.16</td>
<td>.24</td>
<td>.17</td>
<td>.26</td>
<td>.17</td>
<td>.23</td>
<td>.20</td>
<td>.17</td>
<td>.22</td>
<td>.26</td>
</tr>
<tr>
<td>sums</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>49</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>summarizer</th>
<th>55</th>
<th>65</th>
<th>81</th>
<th>93</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>.24</td>
<td>.30</td>
<td>.23</td>
<td>.26</td>
<td>.48</td>
<td>.44</td>
<td>.38</td>
<td>.50</td>
<td>.42</td>
<td>.46</td>
<td>.45</td>
<td>.47</td>
</tr>
<tr>
<td>sums</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>

### Table 6: Coverage for the multi-document summarization task in DUC 2004

<table>
<thead>
<tr>
<th>summarizer</th>
<th>16</th>
<th>19</th>
<th>20</th>
<th>25</th>
<th>26</th>
<th>29</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>sums</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

### Table 8: Average coverage for headlines for multiple documents, DUC 2002. The first row contains the system codes, the second the average coverage and the third—the number of summaries produced by the summarizer.

<table>
<thead>
<tr>
<th>summarizer</th>
<th>1</th>
<th>10</th>
<th>13</th>
<th>15</th>
<th>17</th>
<th>18</th>
<th>21</th>
<th>22</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>.48</td>
<td>.15</td>
<td>.25</td>
<td>.17</td>
<td>.40</td>
<td>.36</td>
<td>.32</td>
<td>.31</td>
<td>.24</td>
<td>.29</td>
<td>.38</td>
<td>.3</td>
</tr>
<tr>
<td>sums</td>
<td>624</td>
<td>622</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td>624</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>summarizer</th>
<th>8</th>
<th>9</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>.51</td>
<td>.27</td>
<td>.49</td>
<td>.57</td>
<td>.56</td>
<td>.53</td>
<td>.61</td>
<td>.60</td>
<td>.57</td>
<td>.52</td>
<td>.52</td>
<td>.49</td>
</tr>
<tr>
<td>sums</td>
<td>623</td>
<td>624</td>
<td>184</td>
<td>192</td>
<td>189</td>
<td>187</td>
<td>183</td>
<td>182</td>
<td>185</td>
<td>192</td>
<td>188</td>
<td>190</td>
</tr>
</tbody>
</table>

### Table 9: Summarizer codes, average coverage and number of produced summaries for single document headlines in DUC 2003
further from human performance in single document summarization than in the multi-document task. This is to a certain extent counter-intuitive to the general feeling that multi-document summarization is the more difficult of the two tasks. The result can be partly explained by the fact that repetition across input document can be used as an indication of importance in multi-document summarization, while such cues are not available in single document summarization. At the same time, the coverage scores of humans for the single document summarizations are significantly higher than human coverage score for multi-document summarization (cf. tables 2 and 6), indicating that there is better agreement between humans on content selections decisions in single document summarization than in multi-document and that thus the single document task is easier and better defined for humans.

Future DUC tasks will continue the emphasis on focused summarization. Past experience indicates that this task is difficult and it might make sense to break it down into simpler, more manageable subtasks.

In the analysis of variance model, the input set for multi-document summarization and the document for single document summarization were significant factors. Thus, some inputs are more amenable to summarization (or at least lead to better coverage scores) than others. No systematic study has been done to identify types of “easy” or “hard” inputs. Such a study can bring insight for the development of more specialized summarizers.

As a final note, we want to mention again that all the results reported on here concern the summarization of newswire and that multi-document summaries were produced for inputs of about 10 articles. The results might not carry over to different domains or to sets containing dramatically more input articles.

Discussion and conclusions

Several interesting conclusions follow from the presented results, that should be kept in mind when developing future systems and when defining tasks. In generic multi-document summarization we observed that differences between systems become more significant when longer (100 or 200) word summaries are produced. But for summaries of such length, text qualities such as coherence and cohesion become important. While some quality aspects of summaries have been evaluated in DUC, little attention has been paid to them, with the main focus always being on coverage. Thus, in order to be able to develop practically useful summarization systems, quality issues need to be addressed and assessed as well. The need for concentrating more on text quality is also supported by the fact that one of the systems for multi-document headlines was not significantly worse in terms of coverage than any human. But the ability of a system to be as useful as a human summarizer would also depend on the overall readability of the automatic headlines.

Another somewhat surprising result is that systems are further from human performance in single document summarization than in the multi-document task. This is to a certain extent counter-intuitive to the general feeling that multi-document summarization is the more difficult of the two tasks. The result can be partly explained by the fact that repetition across input document can be used as an indication of importance in multi-document summarization, while such cues are not available in single document summarization. At the same time, the coverage scores of humans for the single document summarizations are significantly higher than human coverage score for multi-document summarization (cf. tables 2 and 6), indicating that there is better agreement between humans on content selections decisions in single document summarization than in multi-document and that thus the single document task is easier and better defined for humans.

Future DUC tasks will continue the emphasis on focused summarization. Past experience indicates that this task is difficult and it might make sense to break it down into simpler, more manageable subtasks.

In the analysis of variance model, the input set for multi-document summarization and the document for single document summarization were significant factors. Thus, some inputs are more amenable to summarization (or at least lead to better coverage scores) than others. No systematic study has been done to identify types of “easy” or “hard” inputs. Such a study can bring insight for the development of more specialized summarizers.

As a final note, we want to mention again that all the results reported on here concern the summarization of newswire and that multi-document summaries were produced for inputs of about 10 articles. The results might not carry over to different domains or to sets containing dramatically more input articles.

References


<table>
<thead>
<tr>
<th>summarizer</th>
<th>109</th>
<th>116</th>
<th>122</th>
<th>125</th>
<th>147</th>
<th>16</th>
<th>24</th>
<th>30</th>
<th>43</th>
<th>49</th>
<th>5</th>
<th>62</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>.24</td>
<td>.17</td>
<td>.18</td>
<td>.19</td>
<td>.22</td>
<td>.16</td>
<td>.21</td>
<td>.20</td>
<td>.20</td>
<td>.21</td>
<td>.19</td>
<td>.20</td>
</tr>
<tr>
<td>sums</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>summarizer</td>
<td>71</td>
<td>86</td>
<td>96</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td>.21</td>
<td>.15</td>
<td>.22</td>
<td>.48</td>
<td>.55</td>
<td>.35</td>
<td>.47</td>
<td>.47</td>
<td>.52</td>
<td>.54</td>
<td>.44</td>
<td></td>
</tr>
<tr>
<td>sums</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>18</td>
<td>18</td>
<td>25</td>
<td>18</td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 10: Coverage for the 100 words “Who is X?” multi-document summaries at DUC 2004

Table 11: Significant differences for 100 word multi-document summarizers (focused by question) in 2004

A  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96
B  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96 C
C  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96 C
D  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96 C
E  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96 C
F  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96 C
G  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96 C
H  | 109 116 122 125 147 16 24 30 43 49 5 62 71 86 96
109 | 16 86