Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference

Ani Nenkova
Columbia University
1214 Amsterdam Ave
New York, NY 10027
ani@cs.columbia.edu

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 then 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. Results also indicate that most progress in the field has been achieved in generic multi-document summarization and that the most challenging task is that of producing a focused summary in answer to a question/topic.

Introduction

The Document Understanding Conference (DUC) has been run annually since 2001 and has been the major forum for comparing summarization systems on a shared test set. Four main groups of task have been addressed and manually evaluated for coverage (overlap between the system produced summary and a single human model): generic single-document summarization, generic multi-document summarization, headline generation and summarization focused by question or topic. Many summarization systems have been developed and evaluated in DUC, and an overview of the results each year have been provided by DUC organizers[1]. But no comprehensive overview of the results for specific tasks across the years exist to document the progress, or lack thereof, for certain tasks.

[1]The online proceedings of the conferences are available at http://duc.nist.gov

DUC coverage evaluation procedure

The most important evaluation measure used in DUC 2001 to 2004 is *coverage*, intended to capture how well summarizers perform *content selection* and not addressing issues such as readability and other text qualities of the summaries. The coverage scores are computed as the overlap between a human-authored summary (model) and a target summary to be evaluated (peer, produced either by a system or a human). The model summary is split in clauses and a human compares the model and peer summaries, and for each clause in the model summary answers the question “To what extent is the information conveyed in the model clause expressed in the peer summary?”. In 2001 the answer was given on a five point scale where 1 means “The model clause is not covered at all by the peer” and 5 means “The meaning of the model clause is fully expressed in the peer summary”. In later years, a six point scale was adopted, ranging from 0 to 1 with 0.20 increments. A choice X was interpreted to mean that X percent of the content of the model clause is covered by the peer. The coverage score for the entire summary is the average coverage of model clauses. In this paper, we compare the results across years on the coverage measure.

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

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (1)$$

Y_{ij} is the coverage score for system i for input j, and it is equal to μ, a grand mean coverage for any summary, adjusted for the effect of the summarizer (α_i) and the effect of the input (β_j) and some random noise ϵ_{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

sys	better than systems
1	15 16 17 18 23 25 30
A	1 15 16 17 18 21 23 25 31
B	1 15 16 17 18 19 21 23 25 27 28 29 31
C	1 15 16 17 18 19 21 23 25 27 28 29 31
D	1 15 16 17 18 19 21 23 25 27 28 29 31
E	1 15 16 17 18 19 21 23 25 27 28 29 31
F	15 16 17 18 23 25 29 31
G	1 15 16 17 18 19 21 23 25 27 28 29 31
H	1 15 16 17 18 19 21 23 25 27 28 29 31 A F J
I	1 15 16 17 18 19 21 23 25 27 28 29 31
J	15 16 17 18 23 25 29 31
15	17 25 30
16	17 30
18	17 30
19	15 16 17 18 23 25 30
21	15 16 17 18 25 30
23	17 25 30
25	17 30
27	15 16 17 18 23 25 30
28	15 16 17 18 23 25 30
29	16 17 18 25 30
31	16 17 18 25 30

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 summaries 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

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	30	25	25	30	29	30

Table 2: Summarizer code, coverage and number of summaries for the single document summarization task at DUC 2002

2	1 U W
A	1 2 L M O P R S U W Y Z
B	1 2 L M N O P R S U W Y Z
C	1 2 L M N O P R S U W Y Z
D	1 2 M O R S U W Z
E	1 2 H L M N O P R S T U W Y Z
F	1 2 H L M N O P R S T U W Y Z
G	1 2 L M N O P R S U W Y Z
H	1 U W Z
I	1 2 L M N O P R S T U W Y Z
K	1 2 L M N O P R S T U W Y Z
L	1 U W Z
M	1 U
N	1 M O R S U W Z
P	1 U W Z
R	1 U Y
S	1 U W
T	1 2 M O R S U W Z
Y	1 M O U W Z

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.

if systems are compared separately at different target summary lengths—would the results differ. For the 2001 data, we compared the groups of systems, humans, and baselines. At all lengths the humans performed significantly better than both the systems and the baselines. But the difference between the baselines and the automatic systems is significant only for summary lengths of 200 and 400 words and is not significant for target length 50 or 100 words. The same trend can be seen in the 2002 data as well (table 4). The inability to show significance in the difference between systems and baselines with shorter summary lengths might mean two things. The evaluation method might be too restrictive, measuring the overlap with a single human model. The work of (Nenkova & Passonneau 2004) discusses an alternative evaluation approach that looks for overlap between the summary to be evaluated and multiple human models, which is less restrictive and allows for the possibility that there might be more than one, different, but equally acceptable, summaries for the same input. It will be interesting to see if in upcoming DUC competitions the new pyramid scoring procedure will lead to greater ability to show significance. Another possible explanation is that as a shorter summary length is required, the task of choosing the most important information becomes more difficult and no approach works well consistently. Several interesting observations can be made from the data in table 3 (the results from 2001).

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[2].
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-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

Set 34; H SUMMARY
Category 5 hurricanes include the 1935 storm striking the Florida Keys, 1969's Camille striking Mississippi and Louisiana, 1988's Gilbert striking from the Lesser Antilles to Mexico, and 1989's Hugo striking the Lesser Antilles and South Carolina. Other killer hurricanes struck Galveston in 1900, the Florida Keys in 1965 (Betsy), the Dominican Republic in 1987 (Emily), Central America in 1988 (Joan), the Gulf of Mexico in 1989 (Chantal), and Mexico in 1990 (Diana). Hurricanes also hit Miami in 1926, 1947, 1949, 1950 and 1964 (Cleo); the Florida Keys in 1960 (Donna) and 1987 (Floyd); Mississippi, Alabama and Florida in 1985 (Elena); and Mexico in 1985 (Debby).

Set 34; F SUMMARY
Hurricanes are spanned in the Atlantic off the coast of Africa. Average or heavy rains in the Sahel result in more, and more intense, hurricanes and dry seasons lead to fewer hurricanes according to Dr. William Gray a Colorado meteorologist. Since accurate records have been kept, three category five hurricanes have hit the Atlantic or Gulf coasts and caused great loss of life and billions of dollars of damages. In 1900, the Galveston hurricane killed 6,000. More and better satellites will improve hurricane tracking. However, the potential for loss of life and extensive damages are higher because of increased population and construction in coastal areas.

Figure 1: Summaries by humans F and H for the same input

A	*all automatic summarizers*
B	*all automatic summarizers*
C	*all automatic summarizers but 65*
D	*all automatic summarizers*
E	*all automatic summarizers*
F	*all automatic summarizers*
G	*all automatic summarizers*
H	*all automatic summarizers*
2	111 117
19	111 117
27	111
34	111 117
44	111 117 123 138 27
55	111 117 138 27
65	11 111 117 123 138 2 27 34
81	111 117 123 138 27
93	111 117 123 138 27
102	111 117 138 27
120	111 117 138 27
123	111
124	111 117 138 27
138	111

Table 7: Significant differences for 100 word multi-document summarizers in 2004

clear if the sets in 2002 were markedly different (i.e. harder) than those in 2004.

In order to gain some preliminary insight into what factors might make a set easier or harder to summarize, we looked at the two 2004 sets for which the peer coverage was highest (coverage = 0.49, set D30010) and lowest (coverage = 0.10, set D30049). Set D30010, for which the highest peer coverage was achieved, consists of articles reporting on a suicide bombing on a Jerusalem market. The set with lowest average coverage, D30049, consists of articles dealing with different aspects of the controversy surrounding the revival of North Korea's nuclear program. Thus, the set where high coverage was obtained was more scenario oriented, where the pieces of information that need to be supplied in order to give a summary of the event can be easily identified—where the attack took place, who was the perpetrator, what were the casualties. The set with low coverage, on the other hand, does not evoke a readily available scenario. Deeper analysis of the input factors is due in order to understand which are the subareas in which summarization is most successful. Previous research (Barzilay & Lee 2004) has shown that content selection and ordering can be approached in a very principled way for scenario-type inputs.

Generation of headlines

The headline generation task is to produce a 10-word summary, either as a sentence or as a set of keywords. In 2002, the headlines were generated for multiple documents on the same topic, while in 2003 they were for single documents. The average coverage per peer in the two years can be seen it tables 8 and 9 respectively. In 2002, one of the systems, 19, was not significantly worse than *either* of the humans and outperforms all other automatic systems.

For headline generation for single documents, the human written headline (available in the HEADLINE tag of the documents) was taken as a baseline. Thus the baseline was human written and very high, and in fact outperformed all automatic systems. Five out of ten human summarizers outperform the baseline.

Focused multi-document summaries

In 2003 and 2004, the multi-document task was modified so that the produced 100-word summary is not generic, but either responding to a topic description or a question (in 2004 the question was "Who is X?" where X is a person). The same baselines as for generic multi-document summarization are used and have the smallest number system code. For the question task in 2003, systems were also given a list of all the sentences that a human had marked as answering the question. Thus the task was actually to find the most important among the potential answers, while in 2004 the systems had only articles and had to identify and rank the answers.

Humans overwhelmingly outperform all of the systems (with a few exceptions for the modified question task in 2003). At the same time few systems perform better than the baselines: in 2003 (for the summaries focused by topic) systems 6, 13, 16, and 20 are better than baseline 2 (first 100 words of the first document), and only system 13 outperforms the baseline that takes first sentences from the input articles until the requested length is reached (code 3). In 2004, no system significantly outperforms the baseline as can be seen in table 10.

The coverage for human summarizers for the question-focused task in 2004 seems higher than the human cov-

sys	50 words	100 words	200 words
A	16 2 20 24 25 29 3	16 2 20 24 25 26 29 3	16 2 20 24 25 29 3
B	16 19 2 20 24 25 26 28 29 3	6 19 2 20 24 25 26 28 29 3	6 19 2 20 24 25 26 28 29 3
C		16 2 20 25 26 29	16 19 2 20 24 25 26 28 29 3
D	16 2 20 24 25 29 3)	16 2 25 29	16 19 2 20 24 25 26) 28 29 3
E	16 2 20 24 25 26 28 29 3		2
F	16 2 20 24 25 3	16 2 25 29	16 2 25 29
G	16 2 20 24 25 29 3		16 2 25 29
H	16 2 20 24 25 26 28 29 3	16 2 20 24 25 26 29 3	16 19 2 20 24 25 26 28 29 3
I		16 2 24 25 26 28 29 3	16 19 2 20 24 25 26 28 29 3
J		16 2 25 29	16 2 25 29
3		16 2 25	16 2 25 29
19	16 20 24 25 3	16 25 29	16 20 24 25 29
20			16 2 25
24		26	16 2 25 29
26	16 20 24 25	2 20	16 2 25 29
28	16 20 25	16 2 20 25 29	16 2 25 29

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

summarizer	16	19	2	20	24	25	26	28	29	3
50 wrds	.12	.23	.14	.12	.13	.10	.21	.21	.16	.14
100 wrds	.13	.24	.13	.16	.19	.12	.19	.23	.14	.20
200 wrds	.15	.25	.13	.21	.22	.16	.24	.23	.17	.22
sums	59	59	59	59	59	59	59	59	59	59

summarizer	A	B	C	D	E	F	G	H	I	J
50 wrds	.37	.44	.25	.40	.46	.33	.44	.32	.26	.22
100 wrds	.37	.43	.38	.28	.25	.28	.36	.31	.39	.26
200 wrds	.37	.42	.45	.39	.27	.30	.29	.37	0.41	.26
sums	6	6	5	6	6	5	5	6	6	6

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

summarizer	102	11	111	117	120	123	124	138	19	2	27	34	44
coverage	.24	.22	.05	.16	.24	.17	.26	.17	.23	.20	.17	.22	.26
sums	50	50	50	50	50	50	50	50	49	50	50	50	50

summarizer	55	65	81	93	A	B	C	D	E	F	G	H
coverage	.24	.30	.25	.26	.48	.44	.38	.50	.42	.46	.45	.47
sums	50	50	50	50	18	18	25	18	18	17	18	18

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

summarizer	16	19	20	25	26	29	A	B	C	D	E	F	G	H	I	J
coverage	.12	.39	.13	.14	.26	.09	.40	.65	.55	.45	.57	.48	.32	.62	.30	.50
sums	59	59	59	59	59	59	6	6	5	6	6	6	5	6	6	6

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.

summarizer	1	10	13	15	17	18	21	22	24	25	26	7
coverage	.48	.15	.25	.17	.40	.36	.32	.31	.24	.29	.38	.3
sums	624	622	624	624	624	624	624	624	564	624	624	624

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

Table 9: Summarizer codes, average coverage and number of produced summaries for single document headlines in DUC 2003

summarizer	109	116	122	125	147	16	24	30	43	49	5	62
coverage	.24	.17	.18	.19	.22	.16	.21	.20	.20	.21	.19	.20
sums	50	50	50	50	50	50	50	50	50	50	50	50

summarizer	71	86	96	A	B	C	D	E	F	G	H
coverage	.21	.15	.22	.48	.55	.35	.47	.47	.52	.54	.44
sums	50	50	50	18	18	25	18	18	17	18	18

Table 10: Coverage for the 100 words "Who is X?" multi-document summaries at DUC 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
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
H	109 116 122 125 147 16 24 30 43 49 5 62 71 86 96
109	16 86

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

erage for generic summarization the same year— three of the human summarizers in the focused task have coverage score above 0.5 (the highest score achieved in the generic task). But the overall human average is 0.45 for the generic and 0.47 for the focused task, which is not significant (the Wilcoxon rank-sum test p-value is 0.23), suggesting that the intuition that non-generic summarization can lead to better inter-human agreement (Sparck-Jones 1998) is not supported by the 2004 data. At the same time, the average machine scores for the focused summaries are lower than those for generic summaries (p=0.075), that is—the focused task is harder for machines than the generic summary one.

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 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 task 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 summaries 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

Barzilay, R., and Lee, L. 2004. Catching the drift: probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL-HLT'04*.

Harman, D., and Over, P. 2004. The effects of human variation in duc summarization evaluation. In *Text summarization branches out workshop at ACL'2004*.

Nenkova, A., and Passonneau, R. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*.

Sparck-Jones, K. 1998. Automatic summarizing: factors and directions. In *Advances in automatic text summarization, eds. Mani and Maybury*.