Columbia University at DUC 2004

Sasha Blair-Goldensohn, David Evans, Vasileios Hatzivassiloglou,
Kathleen McKeown, Ani Nenkova, Rebecca Passonneau, Barry Schiffman,
Andrew Schlaikjer, Advaith Siddharthan, Sergey Siegelman

Department of Computer Science
Columbia University

April 26, 2004

Abstract

We describe our participation in tasks 2, 4 and 5 of the DUC 2004 evaluation. For each task, we present the system(s) used, focusing on novel and newly developed aspects. We also analyze the results of the human and automatic evaluations.

1 Introduction

In this year's DUC evaluation, we participated in the traditional multi-document summarization (task 2) as well as the new multi-lingual (task 4) and question-focused (task 5) summarization. In the following sections, we detail the approach taken for each task and subtask, providing more detail in cases where the system applied involves significantly novel or newly developed techniques. Results of the various human and automatic evaluations are also analyzed, drawing conclusions about the relative performance of our systems where possible.

2 Task 2: Multi-Document Summarization from English Documents

We made two submissions for Task 2. The first was produced by essentially the same system that was described for DUC 2003 [Nenkova et al., 2003]. In brief, this system routes document clusters to one of two summarizers, DEMS [Schiffman et al., 2002] or MultiGen [Barzilay et al., 1999], thus we call it "DEMS-MG". Clusters with articles that are dated within a short time span are routed to MultiGen, and the rest to DEMS. In this year's DUC, 44 clusters went to DEMS and 6 to MultiGen. Details of the DEMS-MG system are not further discussed in this paper since they have been described in detail elswhere as noted above.

The second submission was produced by a newly developed system which uses sentence simplification and clustering, and we will call it SC. The sentence-clustering approach to multi-document summarization used in SC is similar to the one in MultiGen, with sentences in input documents being clustered according to their similarity. Larger clusters represent information that is repeated more often across input documents; hence the size of a cluster is indicative of the importance of that information.

The SC summarizer has four stages – using syntactic simplification software for preprocessing the original documents to remove relative clauses and appositives, clustering of the simplified sentences, selecting of one representative sentence from each cluster and deciding which of these selected sentences to incorporate in the summary.

The function of appositives and non-restrictive relative clauses is to provide background information on entities, and to relate entities to the discourse. Along with restrictive relative clauses, we feel that their inclusion in a summary should ideally be determined by a reference generating module, not a content selector. We use our syntactic simplification software [Siddharthan, 2002, Siddharthan, 2003] to remove relative clauses and appositives. It uses the LT TTT [Grover et al., 2000] for POS-tagging and simple noun-chunking and performs apposition and relative clause identification and attachment using shallow techniques based on local context and animacy information obtained from WordNet [Fellbaum, 1998].

Another issue in sentence-clustering based summarization is that the clustering is not always accurate. Clusters can contain spurious sentences, and a cluster's size might then exaggerate its importance. Improving the quality of the clustering can thus be expected to improve the content of the summary. We have experimentally confirmed that removing relative clauses and appositives results in a statistically significant improvement in SC's clustering. As an example of how clustering improves, our simplification routine simplifies:

> PAL, which has been unable to make payments on dlrs 2.1 billion in debt, was devastated by a pilots' strike in June and by the region's currency crisis, which reduced passenger numbers and inflated costs.

to:

> PAL was devastated by a pilots' strike in June and by the region's currency crisis.

Three other sentences also simplify to the extent that they represent PAL being hit by the June strike. The resulting cluster is:

1. PAL was devastated by a pilots' strike in June and by the region's currency crisis.
2. In June, PAL was embroiled in a crippling three-week pilots' strike.
3. Tan wants to retain the 200 pilots because they stood by him when the majority of PAL's pilots staged a devastating strike in June.
4. In June, PAL was embroiled in a crippling three-week pilots' strike.

Thus, the removal of background information makes more likely the obtaining of clusters based on the central events in the sentences.

SC clusters the simplified sentences with *SimFinder* [Hatzivassiloglou et al., 1999]. To further tighten the clusters and ensure that their size is representative of their importance, we post-process them as follows. *SimFinder* implements an incremental approach to clustering. At each incremental step, the similarity of a new sentence to an existing cluster is computed. If this is higher than a threshold, the sentence is added to the cluster. There is no backtracking; once a sentence is added to a cluster, it cannot be removed, even if it is dissimilar to all the sentences added to the cluster in the future. Hence, there are often one or two sentences that have low similarity with the final cluster. We remove these with a post-process that can be considered equivalent to a back-tracking step. SC redefines the criteria for a sentence to be part of the final cluster such that it has to be similar (similarity value above the threshold) to *all* other sentences in the final cluster. We prune the cluster to remove sentences that do not satisfy this criterion. Consider the following cluster and a threshold of 0.65. Each line consists of two sentence ids (*P[sent_id]*) and their similarity value.

P37	P69	0.9999999999964279
P37	P160	0.8120098824183786
P37	P161	0.8910485867563762
P37	P176	0.8971370325713883
P69	P160	0.8120098824183786
P69	P161	0.8910485867563762
P69	P176	0.8971370325713883
P160	P161	**0.23330513256176**
P160	P176	**0.04479016583430**
P161	P176	0.7517636285580539

We mark all the lines with similarity values below the threshold (in bold font). We then remove as few sentences as possible such that these lines are excluded. In this example, it is sufficient to remove $P160$. The final cluster is then:

P37	P69	0.9999999999964279
P37	P161	0.8910485867563762
P37	P176	0.8971370325713883
P69	P161	0.8910485867563762
P69	P176	0.8971370325713883
P161	P176	0.7517636285580539

The result is a much tighter cluster with one sentence less than the original.

Having pruned the clusters, SC selects a representative sentence from each cluster based on *tf*idf*. We then incorporate these representative sentences into the summary in decreasing order of their cluster size. For clusters with the same size, we incorporate sentences in decreasing order of *tf*idf*. Unlike MultiGen [Barzilay et al., 1999], which is generative and constructs a sentence from each cluster using information fusion, SC implements extractive summarization and select one (simplified) sentence from each cluster.

2.1 Evaluation

There were 35 entries for the generic summary task (task 2), including ours. Under the automated ROUGE scoring, our simplification + clustering based summarizer (SC) outperformed the DEMS/MultiGen summarizer (DEMS-MG). DEMS/MuliGen on the other hand does very well on the human evaluation. In the SEE scoring by humans (Table 2, the DEMS-MG system was one of the top systems, and in a virtual tie for second place in average coverage, which reflects how well the human judges believe the system summary covers the points made in the human-written model. As only the first submission was evaluated manually, we do not have SEE scores for the SC summarizer.

At 95% confidence levels, our SC system was significantly superior to 23 systems and indistinguishable from the other 11 (using ROUGE-L). Using ROUGE-1, there was one system that was significantly superior to SC, 10 that were indistinguishable and 23 that were significantly inferior. We give a few *ROUGE* scores from DUC'04 in Table 1 below for comparison purposes. The 95% confidence intervals for SC are ± 0.0123 (*ROUGE-1*) and ± 0.0130 (*ROUGE-L*).

Summarizer	ROUGE-1	ROUGE-L
SC	0.3672	0.3804
DEMS-MG	0.3501	0.3580
Best Machine	0.3822	0.3895
Median Machine	0.3429	0.3538
Worst Machine	0.2419	0.2763
Av. of Human Summarizers	0.4030	0.4202

Table 1: ROUGE Scores for Task 2.

3 Task 4: Multi-Document Summarization of Cross-Lingual Documents

Task 4 consisted of three different subtasks, each involving different types of input about the same 25 event sets.

Summarizer	SEE coverage score
DEMS-MG	0.26152
Best System	0.30304
Average System	0.21497
Baseline	0.19964

Table 2: SEE Scores for Task 2

Task 4.1 asked for multidocument summaries of machine translated documents, task 4.2 is to summarize human translations of the documents used in task 4.1, and task 4.3 is to summarize the automatic translations possibly using supplied relevant English documents. We submitted two systems for each priority, with the second system in each case being a sort of baseline system.

3.1 Task 4.1

Task 4.1 is to produce a short summary over machine-translated text. As the machine-translated text contains some errors such as strange word choice, or odd word order, parsing it is difficult and inaccurate, so we opted to use extraction to create the summary. Our approach is to apply our existing sentence-extraction based summarization system with reference re-writing, DEMS [Schiffman et al., 2002], to this data.

Baseline system The baseline system we submitted is a "summary" created by a simple *tf*idf* keyword extraction system run over the machine translated article set. The counts of all tokens from the document set are multiplied by their IDF values from a corpus of approximately 1 million Associated Press news articles from 1989–1997. We included the keyword runs to try to get an idea of how the various ROUGE metrics would score.

The *tf*idf* system that we submitted was the worst performing system by far, which is unsurprising. The DEMS run performs poorly as well, outperforming only 1 system, but using DEMS on translated text is not a focus of our research at Columbia. The summarizer was run as is, without any adaptation to the noisy machine-translated input.

3.2 Task 4.2

Baseline system We again submitted a baseline system of keyword runs using the same system as in task 4.1 over the manual translations. As expected, this system performed poorly.

Simplification and Clustering system In addition, we submitted a run for this task which used the same system used for task 2, described in detail Section 2.

Among all the entries for task 4.2, this system performed significantly worse than 2 systems, significantly better than 2 and was indistinguishable from 6 system on ROUGE-1. On ROUGE-L, there were 6 systems significantly superior to ours, 1 that was significantly worse, and 3 that were indistinguishable. It appears that a sample size of 24 document sets might be too small for automatic metrics to be reliable.

3.3 Task 4.3

Our submission for Task 4.3 uses similarity at the sentence level to identify sentences from the relevant English text that are similar to sentences in the machine translations, and include those in the summary. Since it is extremely difficult to find sentences in the related English documents containing exactly the same information as the translated sentences, we would prefer to perform similarity computation at a clause or phrase level. We opted to use full sentences from the translated text, but wanted to perform some more sophisticated processing on the English text. We ran the English text through sentence simplification software [Siddharthan, 2002] to reduce them in the hope that a single concept would be expressed by each resulting sentence, allowing us to mix and match simplified sentences that might have originally been from a single, more complicated sentence. The sentence simplification software breaks a long sentence into two separate sentences by removing embedded relative clauses from a sentence, and making a new sentence of the removed embedded relative clause. The overall system:

- Selects sentences for the summary from the machine translated documents using DEMS
- Performs sentence simplification on related English documents
- Computes similarity of selected sentences to simplified related English sentences
- Replaces selected summary sentences with English sentences that are very similar

Baseline system The baseline run for Task 4.3 is a run of DEMS modified to prefer sentences from the relevant English documents. The DEMS summarization strategy is not changed, but sentences from the translated documents are given a lower final weight multiplier arbitrarily set without any tuning. We would have tuned the weight multipliers, but no relevant training data was available.

The similarity-based system and modified DEMS runs scored statistically significantly worse than the other 4 systems, and were indistinguishable from each other. Further experiments performed when the evaluation data was released show that the parameters used for the submitted system result in the poorest scores compared to other parameter settings. For this task, only 10 of the 24 document sets contained relevant English articles to use in the summarization process. For those sets without any relevant English articles, our similarity-based system reverted to a DEMS summary.

3.4 A Note on the ROUGE Evaluation for Task 4

ROUGE is an automatic evaluation aimed at quantifying content selection. For the multilingual task, automatically evaluating content selection without taking the summary quality into account is fairly meaningless. The top submissions for task 4.1 appear to perform at 90% of human level when evaluated on ROUGE; this is obviously misleading. In an experiment, we ran our SC summarizer separately on the human translations, the ISI translations and the IBM translations. There was no significant difference in ROUGE scores.

Machine translations and human translation might use the same vocabulary, but at the same time machine translated text is far less readable than human translations. On the other hand, summaries generated by substituting natural English text for machine translated text are likely to be more readable. But this improvement will not be reflected in improved ROUGE scores.

In general, automatic scoring methods make assumptions about the input that are violated in the multilingual case. As an analogy, standard readability metrics like Flesch would indicate that summaries of machine translations and human translations are equally readable since they have similar distributions of sentence and word lengths.

4 Task 5: Question-Focused Summarization

Task 5 introduces a summarization problem constrained by simple questions of the type "Who is X?", where X is the name of a person or group of people. As with DUC 2003's question constrained summarization task, DefScriber, a self-contained component of our AQUAINT (Advanced Question Answering for Intelligence) project system, was modified for use here [Blair-Goldensohn et al., 2003, Blair-Goldensohn et al., 2004].

Unlike last year's question constrained summarization task, this year's task 5 places more emphasis on information retrieval and filtering in that no small "relevant and novel" sentence set is made available for input questions. Consequently, DefScriber's techniques addressing problems of information filtering, omitted from our DUC 2003 system, were included and updated for this year's task 5 system.

4.1 System Overview

DefScriber's is a system which provides multi-sentence answers to questions of the form "What is X?" through a combination of goal-driven and data-driven techniques. The data-driven techniques shape answer content in a bottom-up manner, according to themes found in the data, using statistical techniques including centroid-based similarity [Radev et al., 2000] and clustering [Hovy and Lin, 1997]. The goal-driven techniques apply a top-down method, using a set of *definitional predicates* to indentify types of information ideally suited for inclusion in a definition, such as hierarchical information (i.e., "X is a kind of Y distinguished by Z."). Two methods are used to automatically identify instances of these predicates in text: feature-based classification from machine-learned decision rules, and pattern recognition using patterns manually extracted from a hand-marked corpus. A detailed description of DefScriber can be found in [Blair-Goldensohn et al., 2004].

Because of the similarity between definitional ("What is X?") questions and the "Who is X?" questions of this task, DefScriber is naturally applicable here. However, certain changes were made to improve the quality of summaries for questions relating to individuals and groups of people, as opposed to the more general class of terms which DefScriber is meant to handle.

Our summarization process follows these steps:

1. **Identify and extract relevant sentences** containing "definitional" information for the target individual or group X, as identified by definitional predicate classifiers.

2. **Incrementally cluster extracted sentences** with cosine distance metric, employing global and local word-stem IDF features.

3. **Select sentences for output summary**, using a fitness function which maximizes inclusion of core definitional predicates and sentences from clusters which are statistically closest to the centroid of all definitional sentences. The summary character length restriction is used as a stopping criteria for this process.

4. **Apply rewriting techniques** to references to people in extracted sentences to improve readability of summary.

The key modifications made to DefScriber for task 5 were in the first and last steps of this pipeline.

In the initial step, identification of definitional material performs an information filtering function. Since we rely on string matching on the target of our question (i.e., the "X") to anchor the detection of definitional information, we needed to adapt the system to an X which was a person's name. In particular, we loosened the criteria for matching instances of the target term X, as it was important to allow for the fact that names are often shortened or abbreviated when referenced repeatedly in text. By relaxing sentence filtering to accept sentences containing partial matches of the target name, we observed that recall of relevant sentences in the training sets was drastically improved.

A more significant modification was the addition of the final step of the pipeline. There, we used a system for the rewriting references to people [Nenkova and McKeown, 2003] to make DefScriber's initial output more coherent. This experimental addition reflected our belief that reference resolution and cohesion, always an issue with extracted text, can be particularly treacherous when the core concept being defined is a person. While the rewriting system we used was previously deployed in the context of general news summarization [McKeown et al., 2003], this was our first effort at integrating it with our question-answering architecture.

4.2 Discussion

Since we evaluated only a single submission in this task, it is difficult to assess the individual contribution of the modifications discussed above. However, we did perform manual examination on a sample of the submission output to get a sense of the effect of reference rewriting.

Overall, we observed that felicitous rewritings outnumbered the errors which were introduced. Still, we did encounter occasional significant mistakes, for instance in the well-known difficult case where discrete named entities with extremely similar names occur close together. In the summary for document set 155 ("Who is JFK, Jr.?"), our system attempts to rewrite the name cannonically, with disastrous results:

Q: "Who is JFK Jr.?"

A: **President John F. Kennedy** was traveling with his wife, Carolyn Bessette Kennedy, and sister-in-law, Lauren Bessette, to a family wedding when their plane disappeared over the ocean on its way to Martha's Vineyard, Mass. ...

However, such errors were outnumbered by successful rewritings, even when two similarly named individuals are involved. Table 3 shows our summary for document set 192 ("Who is Sonia Gandhi?"), where the system navigates rewriting in the context of two Gandhis (Rajiv and Sonia) quite well. The left side of the table shows the original output of DefScriber after Step 3 in the pipeline; the right side shows the corresponding sentences in the automatically rewritten final output. As can be seen, official titles (when identified) are rendered with the first mention of a specific named entity, and shortened forms are used for subsequent references to the same named entity.

Manual Metric	Our Rank	Sig. Worse	Better
Mean SEE Coverage	6	2	0
Mean Responsiveness	8	2	1
Qual Question 1	9	1	2
Qual Question 2	7	0	3
Qual Question 3	7	0	1
Qual Question 4	10	4	3
Qual Question 5	6	3	3
Qual Question 6	10	1	5
Qual Question 7	2	7	0
Mean Qual Quests	7.3	1.9	2.4

Table 4: Our results, rankings, and the number of systems doing significantly worse and better than ours for the manual metrics evaluated on task 5. (15 automatic systems were evaluated manually.)

4.3 Evaluation

Our modified DefScriber system performed well in the automated ROUGE evaluation, achieving the top score among the 23 automated systems on all ROUGE variants except ROUGE-4. Using the provided confidence interval, we significantly outperformed 19 of the 22 other systems on ROUGE-1, and 20 of 22 on ROUGE-L and ROUGE-W-1.2 ($p < .05$). No ROUGE metric produced scores where DefScriber scored significantly worse than any other system.

On the manual evaluation metrics, our results were close to average (Table 4). However, using a paired t-test, we found there were fairly few significant differences ($p < .05$), so it appears that these evaluations may not have been very discriminative. On the traditional SEE coverage question, we rank sixth but no system is significantly better; on the responsiveness question we score significantly better than two systems and worse than one.

In the quality questions, we tended to fall in a large middle group of about ten systems, with one or two systems (not always the same systems) standing out as signif-

Summary Before Rewriting	Automatically Rewritten Final Output
Gandhi, who married into what was once India's most powerful political family, is the first non-Indian since independence 50 years ago to lead the Congress.	**Congress President Sonia Gandhi**, who married into what was once India's most powerful political family, is the first non-Indian since independence 50 years ago to lead the Congress.
After **Rajiv Gandhi** was assassinated in 1991, **Sonia Gandhi** was persuaded by the Congress to succeed her husband to continue leading the party as the chief, but she refused.	After **Prime Minister Rajiv Gandhi** was assassinated in 1991, **Gandhi** was persuaded by the Congress to succeed her husband to continue leading the party as the chief, but she refused.
The BJP had shrugged off the influence of the 51-year-old Sonia Gandhi when she stepped into politics early this year, dismissing her as a "foreigner."	The BJP had shrugged off the influence of the 51-year-old Sonia Gandhi when she stepped into politics early this year, dismissing her as a "foreigner."
Sonia Gandhi is now an Indian citizen.	Sonia Gandhi is now an Indian citizen.
Mrs. Gandhi, who is 51, met her husband when she was an 18-year-old student at Cambridge in London, the first time she was away from her native Italy.	**Gandhi**, who is 51, met her husband when she was an 18-year-old student at Cambridge in London, the first time she was away from her native Italy.

Table 3: An example of rewriting in our question-focused summary for document set 192 ("Who is Sonia Gandhi?")

icantly worse or better on each question. Interestingly, we did not fare especially well on the questions which specifically ask about the quality of noun phrase references. On questions four (which asks whether noun phrases should have been in a longer form) and five (which asks the opposite), we were only average (significantly better than three systems, worse than and equal to three and four respectively). While we would like to imagine our rewriting step is helping our scores on these questions, the precise impact is difficult to assess without having scores for our non-rewritten summary.

We were pleased to do well in the ROUGE evaluation, but puzzled that our strong results there did not carry over to the manual evaluation. We found this particularly vexing since a key aspect of our system, named-entity rewriting, did not distinguish itself on the Quality Questions which asked about it. We hope that further examination of the evaluation data, as well as insights from other DUC participants, can shed more light on this apparent divergence between manual and automated evaluations.

5 Conclusion

Our participation in DUC 2004 allowed us to evaluate the performance of various summarization systems being developed at Columbia University. These included systems for the traditional multi-document summarization, as well as the new tasks of multi-lingual and question-focused summarization.

The results of the evaluations highlight that our work in summarization in both traditional and newly added tasks is quite competitive, but that there is also room for improvement.

6 Acknowledgments

The authors acknowledge the support of this work by the DARPA TIDES program (contract N66001-00-1-8919) and ARDA AQUAINT program (contract MDA908-02-C-0008). However, any opinions or errors belong solely to the authors. In addition, we are thankful for the support provided by our colleagues in the Natural Language Processing group at Columbia University.

References

[Barzilay et al., 1999] Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, College Park (MD), USA.

[Blair-Goldensohn et al., 2003] Blair-Goldensohn, S., McKeown, K., and Schlaikjer, A. (2003). A hybrid approach for QA track definitional questions. In *Proceedings of 12th Text Retrieval Conference TREC 2003*.

[Blair-Goldensohn et al., 2004] Blair-Goldensohn, S., McKeown, K., and Schlaikjer, A. (2004). Answering definitional questions: A hybrid approach. In Maybury, M., editor, *New Directions In Question Answering*, chapter 4. AAAI Press.

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

[Grover et al., 2000] Grover, C., Matheson, C., Mikheev, A., and Moens, M. (2000). LT TTT – A flexible tokenisation tool. In *Proceedings of Second International Conference on Language Resources and Evaluation*, pages 1147–1154, Athens, Greece.

[Hatzivassiloglou et al., 1999] Hatzivassiloglou, V., Klavans, J., and Eskin, E. (1999). Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing And Very Large Corpora EMNLP99*.

[Hovy and Lin, 1997] Hovy, E. and Lin, C. (1997). Automated text summarization in SUMMARIST. pages 18–24. In ACL '97 workshop on Intelligent Scalable Text Summarization.

[McKeown et al., 2003] McKeown, K., Barzilay, R., Chen, J., Elson, D., Evans, D., Klavans, J., Nenkova, A., Schiffman, B., and Sigelman, S. (2003). Columbia's newsblaster: New features and future directions (demo). In *Proceedings of NAACL-HLT'03*.

[Nenkova and McKeown, 2003] Nenkova, A. and McKeown, K. (2003). References to named entities: A corpus study. In *Proceedings of NAACL-HLT 2003*. short paper.

[Nenkova et al., 2003] Nenkova, A., Schiffman, B., Schlaiker, A., Blair-Goldensohn, S., Barzilay, R., Sigelman, S., Hatzivassiloglou, V., and McKeown, K. (2003). Columbia at the document understanding conference 2003. In *Proceedings of the Document Understanding Workshop DUC 2003*.

[Radev et al., 2000] Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents. In ANLP-NAACL workshop on summarization.

[Schiffman et al., 2002] Schiffman, B., Nenkova, A., and McKeown, K. (2002). Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference HLT02*.

[Siddharthan, 2002] Siddharthan, A. (2002). Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Workshop, 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 60–65, Philadelphia, USA.

[Siddharthan, 2003] Siddharthan, A. (2003). *Syntactic simplification and Text Cohesion*. PhD thesis, University of Cambridge, UK.