Generating Update Summaries for DUC 2007

René Witte and **Ralf Krestel**
Faculty of Informatics
Institute for Program Structures
and Data Organization (IPD)
Universität Karlsruhe (TH), Germany
witte|krestel@ipd.uka.de

Sabine Bergler
The CLaC Laboratory
Department of Computer Science
and Software Engineering
Concordia University, Montréal, Canada
bergler@cs.concordia.ca

Abstract

Update summaries as defined for the new DUC 2007 task deliver focused information to a user who has already read a set of older documents covering the same topic. In this paper, we show how to generate this kind of summary from the same data structure—fuzzy coreference cluster graphs—as all other generic and focused multi-document summaries. Our system ERSS 2007 implementing this algorithm also participated in the DUC 2007 main task, without any changes from the 2006 version.

1 Introduction

The DUC 2007 competition included two tasks:[1] a main task, involving the generation of focused multi-document summaries, which was unchanged from the previous two years; and a novel update task, where summaries had to be generated for three consecutive document subsets, tracking the development of a single topic through time.

Our summarization system, ERSS (Bergler et al., 2003; Bergler et al., 2004; Witte et al., 2005; Witte et al., 2006), participated in most DUC tasks since 2003, with the only major system update in 2004 (to handle multi-document summarization). A particular feature of ERSS is that all different kinds of summaries required for the various DUC competitions, including short, long, focused, updates, cross-language, single- and multi-document summaries, are generated with the same system from the same data structure: fuzzy coreference cluster graphs, described in (Witte and Bergler, 2007). Thus, even though DUC 2007 included a novel task—update summaries—no changes to the system were necessary.

DUC 2007 Main Task. The main task, generating focused 250-word summaries for a set of 25 documents answering the question(s) in the context, was unchanged from the previous two years. Since we ran our last year's systems without any changes, we refer to (Witte et al., 2006) and (Witte and Bergler, 2007) for a detailed description of its algorithms, implementation, and evaluation. An example output on this year's data is shown in Figure 1.

DUC 2007 Update Task. For the update task, a system had to generate focused 100-word summaries for three sets of documents on the same topic. Each set contained new documents and the update summary should only contain information a reader of the previous documents had not seen before. The generation of update summaries in ERSS is described in the next section.

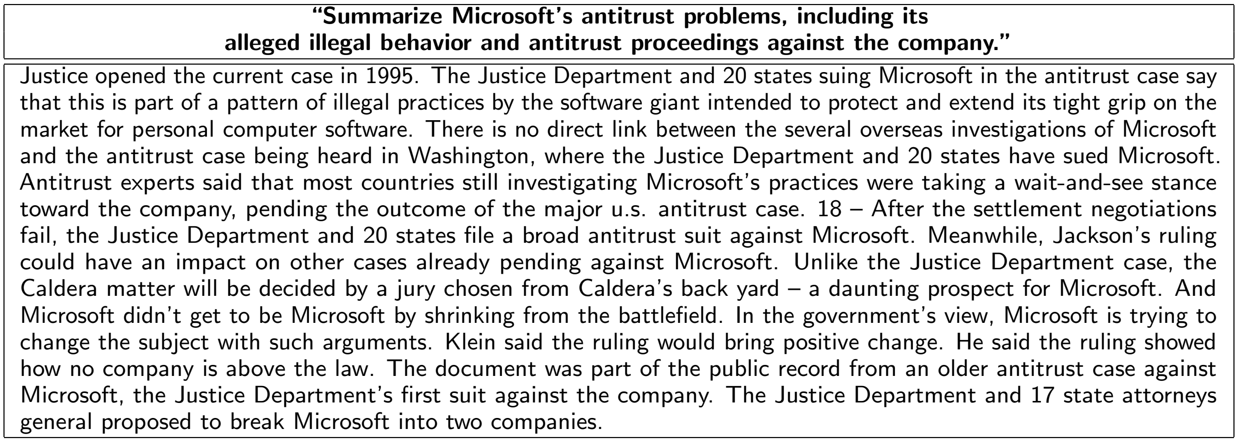
In this paper, we first describe how we generate update summaries based on cluster graphs and then present evaluation results in Section 3.

2 Generating Update Summaries

Update summaries, as required for DUC 2007, are generated based on a context containing a single or set of questions and a number of document clusters (three for DUC 2007) containing topical documents sorted by their publication date. Each update summary (100 words length for DUC 2007) should inform the reader of new developments with respect to the focus questions, assuming that all documents in the previous set(s) have already been read (**note** that the DUC update task description explicitly mentions that the previous *documents* have been read, not just the previous *summaries* of those!).

To generate update summaries, we first generate the cluster graph data structure (Witte and Bergler, 2007) based on the context and the current set of documents (including all previous documents, i.e., not just the new ones). Figure 2 shows an example for a cluster graph generated for three documents with an additional set of context questions, modeled as another document d_0. For

[1] See http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html for the detailed task descriptions.



"Summarize Microsoft's antitrust problems, including its alleged illegal behavior and antitrust proceedings against the company."

Justice opened the current case in 1995. The Justice Department and 20 states suing Microsoft in the antitrust case say that this is part of a pattern of illegal practices by the software giant intended to protect and extend its tight grip on the market for personal computer software. There is no direct link between the several overseas investigations of Microsoft and the antitrust case being heard in Washington, where the Justice Department and 20 states have sued Microsoft. Antitrust experts said that most countries still investigating Microsoft's practices were taking a wait-and-see stance toward the company, pending the outcome of the major u.s. antitrust case. 18 – After the settlement negotiations fail, the Justice Department and 20 states file a broad antitrust suit against Microsoft. Meanwhile, Jackson's ruling could have an impact on other cases already pending against Microsoft. Unlike the Justice Department case, the Caldera matter will be decided by a jury chosen from Caldera's back yard – a daunting prospect for Microsoft. And Microsoft didn't get to be Microsoft by shrinking from the battlefield. In the government's view, Microsoft is trying to change the subject with such arguments. Klein said the ruling would bring positive change. He said the ruling showed how no company is above the law. The document was part of the public record from an older antitrust case against Microsoft, the Justice Department's first suit against the company. The Justice Department and 17 state attorneys general proposed to break Microsoft into two companies.

Figure 1: ERSS-generated focused summary for D0718D (context shown on top)

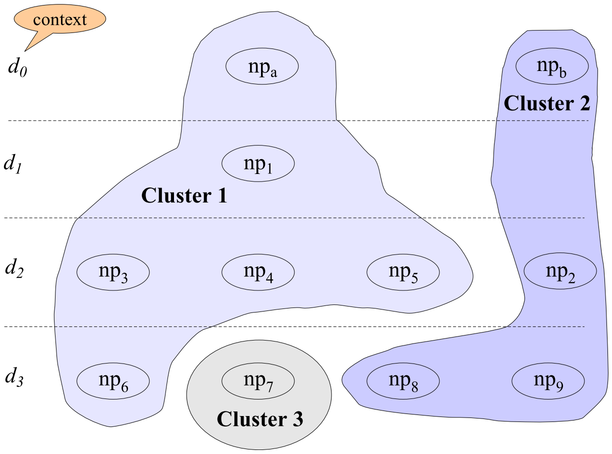


Figure 2: A cluster graph for three documents and a context

the first subset within an update cluster, summary generation is identical to a standard (main task) focused summary, as presented in (Witte et al., 2006). For each subsequent update subset, we re-generate the cluster graph, adding the new documents to the current set. When generating update summaries for these extended clusters, we select sentences based on the following ranking scheme:

1. The highest rank is given to sentences from clusters that overlap with the context (i.e., cover topics from the questions) but do not contain any elements from documents of a previous update (i.e., these are topical information *only* addressed in a new document).

2. A medium rank is given to sentences from clusters that overlap with the context and appear in the newly added (updated) set of documents (i.e., new information addressing a topic that has been covered before).

3. The lowest rank is given to all remaining sentences from clusters that overlap with the context (i.e., answer a question from the context).

An example is shown in Figure 2. Assume a user posed a number of questions, expressed in the context document d_0. Moreover, assume he already knows the contents of document d_1. After generating the cluster graph datastructure for two additional (update) documents d_2 and d_3, Cluster 2 would be ranked high because it overlaps with the context (d_0) and does not contain elements from a previous update (d_1). Thus, the sentences picked from d_2 will contain information regarding the focus questions that has not been addressed in a previous document (set), here, d_1.

Summaries are generated by selecting sentences from each rank, until it has been exhausted, proceeding with the next-lowest ranked ones until the length limit (100 words for DUC 2007) has been reached or all candidate sentences have been exhausted.

An example for an update summary generated with this algorithm is shown in Figure 3.

3 Evaluation

We present results based both on the evaluation performed by NIST, using the automatic ROUGE (Lin, 2004) and BE (Hovy et al., 2005) metrics and the manual Responsiveness measure,[2] as well as the results of additional post-submission experiments.

3.1 Main Task Evaluation

Table 1 summarizes the performance of our system ERSS for the main task. When compared with previous years (see Section 3.3), it is interesting to note its apparent decrease in performance. However, as the system was essen-

[2] See http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html for more details on the evaluation performed by NIST for each task.

"Summarize Microsoft's antitrust problems, including its alleged illegal behavior and antitrust proceedings against the company."

Docsubset D0711C-A

Ever since ibm selected Microsoft to supply the operating system for the first ibm pc in 1981, ibm has been out-hustled and outsmarted by Microsoft in the pc business – not the sort of behavior the nation's antitrust laws were meant to prohibit. Antitrust experts note, it is certainly legal for Microsoft to offer a distributor – as the ibm pc business was – incentives to aggressively distribute its product. If accepting those incentives means featuring Microsoft's Windows over os/2 – made by another unit of ibm – that is mainly a business decision for ibm to make, they say.

Docsubset D0711C-B

The case is assigned to Jackson, who seeks to expedite the proceedings by setting an early trial date. Microsoft denies all charges of wrongdoing. To Microsoft in 1997, is a harsh critic of the company. Meanwhile, Jackson's ruling could have an impact on other cases already pending against Microsoft. And Microsoft didn't get to be Microsoft by shrinking from the battlefield. Tuesday's courtroom statements – uncompromising attacks from each side – only seemed to underline the gap between the government and Microsoft. In the government's view, Microsoft is trying to change the subject with such arguments.

Docsubset D0711C-C

The company said it would appeal. Klein said the ruling would bring positive change. He said the ruling showed how no company is above the law. The document was part of the public record from an older antitrust case against Microsoft, the Justice Department's first suit against the company. The Justice Department and 17 state attorneys general proposed to break Microsoft into two companies. Local press reports said that attorneys for the Justice Department and the 19 states that successfully sued Microsoft for antitrust violations are considering ways to break up the company as a method to curb anticompetitive practices.

Figure 3: ERSS-generated update summary for D0718D (context shown on top)

Measure	ERSS	mean	best / worst	rank
ROUGE-1	0.3789	0.3973	0.4526 / 0.2428	25/32
ROUGE-2	0.0791	0.0949	0.1245 / 0.0381	28/32
ROUGE-SU4	0.1354	0.1475	0.1772 / 0.0739	26/32
Basic Elements	0.0394	0.0477	0.0664 / 0.0010	27/32
Linguistic quality	2.8489	3.2364	4.2978 / 2.0978	28/32
Responsiveness	2.3560	2.6167	3.400 / 1.5560	25/32

Table 1: Evaluation results overview for ERSS 2007 (System ID #10) main task

tially unchanged from last year, we believe this is due to a combination of two factors:

- ERSS is basically a heuristics-based system. As such, it requires no training data, but at the same time does not improve in performance when more training data becomes available, as is the case with the DUC main task (focused multi-document summarization), which has been running essentially unchanged for two years (with an additional, very similar Task 5 in 2004).

- ERSS still does not contain any significant post-processing strategies to remove redundant information from summaries. When the source articles contain a number of slightly different, but very similar sentences that are recognized as relevant, the generated summaries exhibit a large degree of redundant information, which greatly impacts both manual (Responsiveness) and automatic (ROUGE/BE) scores. This happened to a larger extend with the 2007 test data set.

3.2 Update Task Evaluation

Results very similar to the main task were obtained for the update task (Table 2). We suspect this is due to the similarity of the main and update tasks, since an update summary can be trivially generated using a system developed for the main task by simply running it on each of the document subsets, generating a normal focused summary for each. However, compared to the main task ERSS performed slightly better with respect to the other participating systems.

We also evaluated how much better the update cluster strategy described in Section 2 performs on the update tasks when compared with the standard focused summarization (main) task. Here, we re-generated the update summaries using the standard focused summarization algorithm as used for the main task (except with the shorter summary size of 100 words). This can be seen as a "baseline" result for evaluating the performance of an update algorithm (assuming a corresponding focused, but non-update, algorithm has already been developed). The results are shown in Figures 4 (ROUGE metric) and 5 (BE metric) for each document subset; Table 3 contains the average results over all clusters. As can be seen from

Measure	ERSS	mean	best / worst	rank
ROUGE-1	0.2961	0.3262	0.3768 / 0.2621	19/24
ROUGE-2	0.0531	0.0745	0.1117 / 0.0365	20/24
ROUGE-SU4	0.0957	0.1128	0.1430 / 0.0745	20/24
Basic Elements	0.0241	0.0410	0.0721 / 0.0177	21/24
Responsiveness	1.9670	2.3278	2.9670 / 1.6670	20/24

Table 2: Evaluation results overview for ERSS 2007 (System ID #39) update task

	Measure					
	ROUGE			Basic Elements		
	ROUGE-1	ROUGE-2	ROUGE-SU4	Recall	Precision	F-Measure
main algorithm	0.279	0.045	0.088	0.018	0.020	0.019
update algorithm	0.296	0.054	0.096	0.024	0.027	0.025

Table 3: Comparison between the update algorithm and standard main algorithm on the update task data set

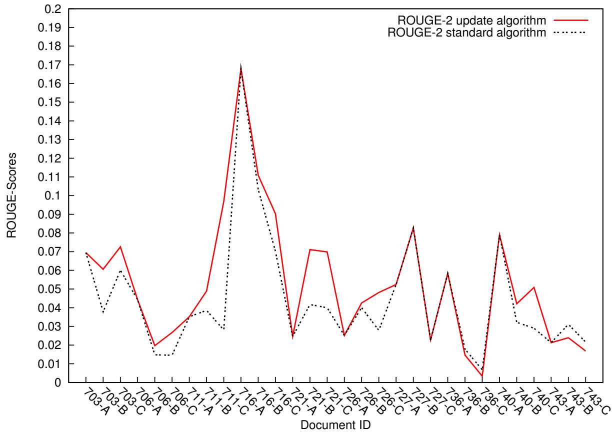


Figure 4: Comparison between the main and update algorithms on the update task test data (ROUGE metric)

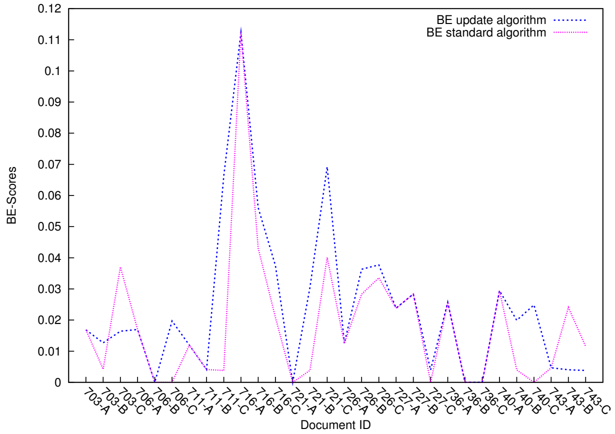


Figure 5: Comparison between the main and update algorithms on the update task test data (BE metric)

the two plots, the scores are identical for the "A"-type clusters (no previous knowledge), as is to be expected, since in that case our update strategy correspond to the standard focused strategy. When previous knowledge becomes available (clusters of type "B" and "C"), our update algorithm significantly outperforms the standard strategy.

An alternative baseline is to run the main task algorithm on only the *new* data for each update cluster. This experiment is still on-going.

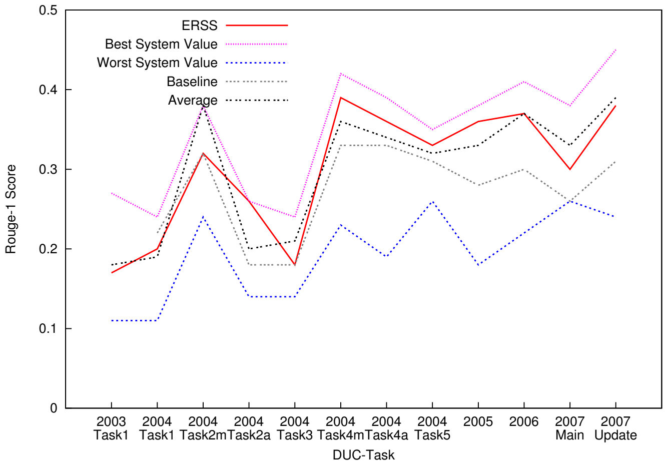


Figure 6: ERSS performance overview from 2003–2007

3.3 Performance of ERSS from 2003–2007

An overview of ERSS' performance from 2003–2007 is shown in Figure 6. Here, we use the ROUGE-1 score to allow a comparison for all five years. This evaluation also confirms the observation noted above: using a single algorithm for all tasks throughout the years results in an overall excellent performance. Still, when enough training data becomes available for statistical systems, they can outperform our cluster-based algorithm. Essentially, statistical systems become more robust within a single task with respect to changing input data, while our cluster approach is more robust with respect to changing tasks, independent of the data.

4 Conclusions

In this paper we demonstrated how a novel task like update summaries can be solved from an essentially unchanged system by relying on an expressive and flexible data structure for summarization, fuzzy coreference cluster graphs.

As a mainly rule-based system, the strength of ERSS is its capability of quickly adapting to novel and changing tasks, without requiring any training data. However, a major weakness of ERSS is its almost complete lack of post-processing for the generated summaries. Especially on the 2007 dataset, a large number of summaries exhibited redundant information stemming from similar sentences within the various articles. Improving this part of the system should result in a significant increase for the various evaluation metrics.

References

Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. 2003. Using Knowledge-poor Coreference Resolution for Text Summarization. In *Proceedings of the HLT/NAACL Workshop on Text Summarization (DUC 2003)*. Document Understanding Conference. http://www-nlpir.nist.gov/projects/duc/pubs/2003final.papers/concordia.final.pdf.

Sabine Bergler, René Witte, Zhuoyan Li, Michelle Khalife, Yunyu Chen, Monia Doandes, and Alina Andreevskaia. 2004. Multi-ERSS and ERSS 2004. In *Proceedings of the HLT/NAACL Workshop on Text Summarization (DUC 2004)*. Document Understanding Conference. http://www-nlpir.nist.gov/projects/duc/pubs/2004papers/concordia.witte.pdf.

E. Hovy, C. Lin, and L. Zhou. 2005. Evaluating DUC 2005 using Basic Elements. In NIST, editor, *Proceedings of the HLT/EMNLP Workshop on Text Summarization DUC 2005*, Vancouver, BC, Canada, October 9–10. http://duc.nist.gov/pubs.html#2005.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25–26. http://www.isi.edu/~cyl/ROUGE/.

René Witte and Sabine Bergler. 2007. Fuzzy clustering for topic analysis and summarization of document collections. In Z. Kobti and D. Wu, editors, *Proc. of the 20th Canadian Conference on Artificial Intelligence (Canadian A.I. 2007)*, LNAI 4509, pages 476–488, Montréal, Québec, Canada, May 28–30. Springer.

René Witte, Ralf Krestel, and Sabine Bergler. 2005. ERSS 2005: Coreference-Based Summarization Reloaded. In *Proceedings of Document Understanding Workshop (DUC)*, Vancouver, B.C., Canada, October 9–10. http://duc.nist.gov/pubs/2005papers/ukarlsruhe.witte.pdf.

René Witte, Ralf Krestel, and Sabine Bergler. 2006. Context-based Multi-Document Summarization using Fuzzy Coreference Cluster Graphs. In *Proceedings of Document Understanding Workshop (DUC)*, New York City, NY, USA, June 8–9. http://duc.nist.gov/pubs/2005papers/ukarlsruhe.witte.pdf.