ERSS at TAC 2008

Sylvain Bellemare, Sabine Bergler and René Witte
CLaC Lab
Concordia University
Montreal, Canada

Abstract

ERSS 2008 attempted to rectify certain issues of ERSS 2007. The improvements to readability, however, do not reflect in significant score increases, and in fact the system fell in overall ranking. While we have not concluded our analysis, we present some preliminary observations here.

1 Introduction

Last year's DUC 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; Witte et al., 2007), participated in DUC tasks since 2003, with the only major design change in 2004. 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. We reused the same system with small enhancements for the TAC 2008 Summarization task.

Performance of ERSS 2007 was respectable, the conclusions pointed to linguistics features that should be improved. Accordingly, ERSS 2008 evaluated the importance of improving the observed shortcomings in our summaries, such as repetitive material and sentence splitting errors. The basic summarization system was unchanged.

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

ERSS selects sentences for summaries based on contained entities that are coreferred to frequently. The basic data structure is the fuzzy cluster graph (Witte and Bergler, 2007), which links all NPs of the texts in two dimensions: coreference within the same document, and coreference with other documents. For focused summaries, the focus text is added as a new document.

(Witte et al., 2007) describes how update summaries based on cluster graphs are created. In short, 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). For the first subset within an update cluster, summary generation is identical to a standard (non-update) 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 addressed 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).

Summaries are generated by selecting sentences from each rank, until it has been exhausted, proceeding with the next-lowest ranked ones until the 100 word length limit

has been reached or all candidate sentences have been exhausted.

2 Changes for 2008

Certain of our summaries last year showed repetitive information. For ERSS 2008, we reduced duplication through an explicit filter and a cosine similarity measure used during sentence selection.

We improved a summary compression strategy that was implemented last year, removing date phrases. The revised strategy removes a wider set of temporal phrases, because many of these (eg. last Tuesday) are wrong when taken out of context. Other temporal phrases (eg. November 15, 1999) were eliminated because they contribute little information.

2.1 Sentence Similarity Detection

In order to avoid generating summaries that would contain redundant information, we compare the initial and final word sequences of two sentences. The length of these sequences is determined through a parameter. If the sequences are equal, a cosine similarity measure is applied to all the stemmed non-stop words.

Two different types of vector representations were considered:

- one using binary weights (0 or 1), and
- one using tfidf weights.

Pre-competition evaluation of both measures showed the binary weights to yield higher scores.

2.2 Requiring a minimum number of words per sentence

ERSS 2007 generated several summaries using very short sentences without much content. ERSS 2008 requires a minimum number of words for each sentence to be incorporated into a summary. During pre-competition testing, we noticed that when imposing a somewhat high threshold (such as 12 words per sentence), the ROUGE and BE scores were better. Thus, for our submission we required a minimum of 10 words per sentence in the summary.

3 Results

Table 1 summarizes the performance of ERSS 2008 for the update task.

The submitted results were not compiled with all modules running correctly. We thus report results for four different systems: ERSS 2007 (last year's system), ERSS 2008 (the submitted system), ERSS 2008+ (the intended system for 2008), and a system with optimized parameter settings, ERSS 2008++.

Measure	baseline	ERSS	best/worst	rank
ROUGE-2	.058	.060	.103/.033	58/71
ROUGE-SU4	.093	.102	.137/.065	55/71
BE	.032	.032	.064/.013	59/71
Ling. quality	3.333	2.427	3.073/1.312	23/57
Responsiveness	2.073	2.104	2.667/1.198	50/57

Table 1: Evaluation results overview for TAC 2008 (System ID #5) update task

Table 2 shows the performance of ERSS 2007, ERSS 2008 used for TAC 2008 submission, ERSS 2008 corrected (with stemmer), and of a tentative optimized ERSS 2008, on the TAC 2008 data.

	ROUGE-2	Rouge-SU4	BE
ERSS 2007	0.05721	0.09835	0.03045
ERSS 2008	0.06077	0.10199	0.03274
ERSS 2008+	0.06112	0.10262	0.03275
ERSS 2008++	0.06090	0.10318	0.03320

Table 2: Evaluation results overview for different configurations of ERSS on TAC 2008 data for the update task

The improvements we brought to the system, while measurable, do not influence its rankings.

3.1 Main versus Update Summaries

We observe that our system performs differently on the main (traditional) summaries and on the update summaries.

Type	5	4	3	2	1
Main	0	4	12	16	16
Update	2	5	7	17	17

Figure 1: Responsiveness scores for main and update summaries (5: very good, 1: very poor)

We have not been able to determine, whether this observation can lead to further insights. Four of our Main summaries (-A) have no SCUs and 0 pyramid score, while eight of our update summaries (-B) show no overlap with model summaries in the pyramid score. This suggests that our exclusion strategy for sentences with topics already present in the Main summary is too strict.

4 Example summaries

We demonstrate the range of our summaries from very good to very poor with the main and update summaries for the same cluster.

4.1 Very Good Responsiveness

We obtained an update summary judged "very good" in responsiveness score.

Stewart, who built a billion-dollar empire in homemaking, reported that her job at the federal prison camp in Alderson, West Virginia, has been cleaning, including sweeping, vacuuming and raking leaves. As part of her sentence, Stewart was ordered to serve five months at the federal women's camp in Alderson. Since entering federal prison Martha Stewart has tried her hand at ceramics, learned to crochet and become an expert on vending-machine snacks. Stewart is reportedly in fighting shape, having lost 20 pounds in prison while impressing the locals with her down-to-earth attitude.

4.2 Very Poor Responsiveness

The main summary of the same topic, surprisingly, is "very poor" and illustrates well the points for improvement: the summary has repetitive content and no flow.

After she is released from prison, Stewart will also have to undergo five months of home confinement. When Stewart leaves prison and begins serving five months on house arrest, she will be allowed to work a limited number of hours per week. The lawyer said he did not know what she is writing, but Stewart has suggested since her conviction that she might write a book about her recent experience with the legal system. Dellinger said he did not know what she is writing, but Stewart has hinted she may write a book about her experience with federal law enforcement.

5 Discussion

The type of readability improvement we focused on, while successful, does not translate into a competitive advantage. This is not surprising, because the content of our summaries is not affected by much.

With the update task maturing, however, we see the need to adapt several of our basic assumptions. Our treatment of the focus text, for instance, consists of adding the focus to the document cluster as an additional document and to demand that any sentence chosen for the summary have overlap with the focus "document". This was a very good strategy, when there were many focus questions. Now, however, the focus text is very short and often contains no named entities, on which our coreference-based approach perform best. This leads to selection of sentences that often contain the same NP, leading to repetitiveness and highlighting of different attributes of that topic, rather than a more comprehensive summary. At the same time, with opposite effect, our attempt to cover as many topics as possible within these constraints leads to inclusion of spurious material.

Quite by design, the task has shifted more towards the question answering format. ERSS 2008 was not adapted to this format and we will test whether we can add some question answering techniques.

The overall shift away from named entities and towards more general topics also suggests to rethink the dominating feature of ERSS 2008, the coreference cluster size.

Acknowledgments

Yonatan Cohen helped to develop ERSS 2008. This research was funded by NSERC.

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.

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.

R. Witte, R. Krestel, and S. Bergler. 2007. Generating update summaries for duc 2007. In *On-line Proceedings of the Document Understanding Conference (DUC), Workshop at NAACL-HLT 2007*, Rochester, NY, April.