Deliverable #4: Final Summarization System

Paige Finkelstein **Jacob Hoffman** **Wesley Rose** **Joshua Tanner**
University of Washington
Department of Linguistics
{plfink, jrhoff, warose91, jotanner}@uw.edu

Abstract

In this deliverable report we outline changes and improvements made since deliverable #3 and discuss our final multi-document summarization system. For this deliverable, we have focused specifically on improving readability, rather than maximizing ROUGE score metrics. We were able to achieve small ROUGE score gains while eliminating sentence fragments and other grammatical readability issues and improving our ordering approach. In this report we also present our results for the held out evaluation dataset, a portion of the English Gigaword corpus, and provide detailed error analysis of the final system summarization results.

1 Introduction

We have finalized our multi-document summarization system that produces brief summaries (100 words or fewer) from topic-oriented document clusters of articles. For this deliverable, we have focused on improving the readability of our system's outputs, primarily through improvements to the content realization portion of our system and also through tweaks to our information ordering approach.

Our final system implementation achieves an average ROUGE-2 recall score of 0.08450 on the AQUAINT (devtest) data set, improving on our previous system's ROUGE-2 recall score of 0.08042. It achieves an average ROUGE-2 recall score of 0.10614 on the English Gigaword corpus (evaltest) data set. The main focus of our improvements for this final deliverable, however, have been centered on improving readability, even in cases where changes that improved readability did not impact or slightly decreased ROUGE scores.

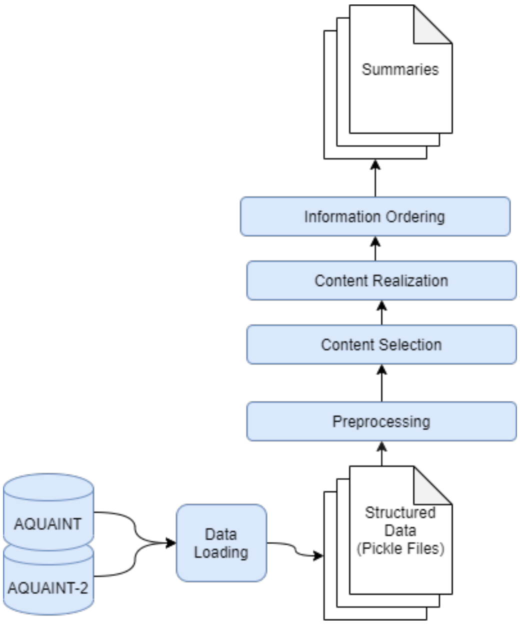

Figure 1: Final system architecture diagram

2 System Overview

Figure 1 shows our final summarization system architecture, which remains unchanged from deliverable #3. It is comprised of modules to fetch and load the articles for each topic cluster for the specified corpus, to perform data cleaning and preprocessing, and to perform each of the major components of summarization, namely, content selection, content realization, and information ordering.

3 Approach

3.1 Content Selection

The best method for content selection did not change from deliverable #3 and continues to be our ngram heuristic method, inspired by work such as sumBasic and HeirSum mentioned in Haghighi and Vanderwende (2009). Per-article selection with re-weighting and 2 sentences per article still yields the highest ROUGE-2 scores. The best weights remained as 0.2 for uni-grams and 0.9 for bi-grams. Query, headline overlap, and Cartesian pair scores negatively affected results. However, we did experiment with one more method, described in the following section.

3.1.1 Forward-Backward Selection

In this method, we do per-article selection twice. Once moving forward through the set of documents and once backwards. During each pass the data is re-weighted using the same scheme as in D3; when a uni-gram W is selected, $P(W)_{new} => P(W)^2_{old}$ as in Vanderwende et al. (2015). The probability distribution is restored to its original before the backward pass begins. A particular sentence only gains half of its average ngram-heuristic score per pass. The purpose of this method was to maybe address the possibility of missing out on good summary-candidate sentence that were not being chosen due to the re-weighting scheme. However, this method only negatively affected ROUGE-2 scores and so we did not use it in our final system.

3.2 Content Realization

The overall structure and approach of our content realization process remains unchanged since deliverable #3. However, a large portion of our work for this deliverable has been on improvements in content realization to produce more readable summary outputs, including experimenting with additional methods for condensing content described in Conroy et al. (2006).

The main changes that we have implemented since deliverable #3 to improve the readability of our summarization outputs include removing sentence fragments from the final output, even in cases when they increased the ROUGE score, as well as discarding the appositive removal functionality. We updated the word-initial removal to avoid removing words in an “exceptions” list for words that, despite their part of speech, are important for content and readability, such as “Now” and “Most.”

We also updated our removal of quotation sentences to *not* include sentences with scare quotes, that is, quotation marks around a word or phrase for emphasis or identification, rather than for marking an actual quote. We improved our redundancy filtering to use a lexical overlap comparison, so that short sentences that were entirely or almost entirely repeated in a longer sentence would be marked as redundant, despite a larger content difference overall due to the extra information in the longer sentence. Finally, we added handling to add back in sentences that had been marked as semi-redundant (but not too redundant) if we had a lot of extra space left in our summary at the end of the realization process.

3.3 Information Ordering

Information ordering for next sentence selection remains largely unchanged from deliverable #3, and is still based on the “preference learning” method outlined in Bollegala et al. (2012) that incorporates weights from a number of “experts.” We did, however, make some changes to how the initial sentence of the summary is selected, which affects the subsequent sentences selected by the expert weight formula as well.

Specifically, we added sentence length as a positive feature for selecting the initial summary sentence, after analyzing the summary outputs and noting the many of the most effective initial sentences tended to be longer in word count. Secondly, we added negative feature weights for very unlikely and semi-unlikely start words, after noticing that effective initial summary sentences almost never began with pronouns such as “He”, “She”, and “They,” or words such as “That” and “Since,” as sentences beginning with such terms were likely to require prior context for readability.

4 Results

4.1 ROUGE Recall Scores

As Tables 1 and 2 show, we have increased our average ROUGE-2 recall score by 0.00408 for the AQUAINT corpus summaries since deliverable #3, and our average ROUGE-2 recall score for the English Gigaword corpus dataset (evaltest) is 0.10614.

Table 1: Average ROUGE Recall Scores for D3 and D4 on the AQUAINT (devtest) data set

Deliverable	ROUGE-1 R	ROUGE-2 R
D3	0.25856	0.08042
D4	0.26493	0.08450

Table 2: D4 Average ROUGE Recall Scores for the English Gigaword (evaltest) data set

ROUGE-1 R	ROUGE-2 R
0.31445	0.10614

4.2 Human Readability Judgments

Our group obtained the highest overall average human readability judgment scores for our summaries. These scores are displayed in Table 3. All of our average scores fell between the range of 3.8 and 4.6, denoting, at a minimum, an evaluation of "mostly understandable; a few minor readability or flow issues." We provide some analysis of our lowest scoring summary in the Error Analysis section below.

Table 3: Human Readability Judgments

Summary No.	**Average Eval Score**
1105	3.8
1106	4.6
1107	4.2
1108	4
1109	4.2
1110	4.6
Overall Average	4.233333

5 Discussion

5.1 Content Realization Updates

Here we present a few examples of ways in which our improvements to content realization have improved our summary outputs for deliverable #4.

For one, our changes to limit the words being selected for "word-initial" removal at the beginning of sentences helped correct for readability issues previously introduced from too-aggressive removal. For example, the part-of-speech for "Most" in the following sentence had been incorrectly tagged as an adverb: *Most had eaten cooked hot dogs the month before they became ill.* This led to the word being removed and the resulting ungrammatical formulation of: *Had eaten cooked hot dogs the month before they became ill.* Now, however, we specify such tricky cases and do not remove them, despite their part of speech tags.

Secondly, our improvement of handling the removal of sentences that contain quotations to no longer include scare quote cases has enabled us to retain some information-dense sentences that would have been discarded other wise, such as the following:

The "bird flu" has claimed four victims here, killing two, including a 54-year-old man who died Friday.

Ronald "Duffy" Hambleton, a former methamphetamine user, said Blake solicited him to kill his wife.

One final example of summary output improvements from our content realization updates is a decrease in repetitive content in short sentences. Previously, because we were only computing similarity across two full sentences, in the case where one sentence was mainly or entirely subsumed by another but there was also extra content in the longer sentence, the sentences weren't being registered as redundant. This led to ending up with summaries that contained two highly repetitive sentences such as:

Norway, which is the world's third-largest oil exporter after Saudi Arabia and Russia and which normally has an average daily production of three million barrels, has seen its oil production reduced by some 205,000 barrels per day (bpd) since the platforms were closed.

Norway is the world's third-largest oil exporter after Saudi Arabia and Russia.

With the additional lexical overlap redundancy checker, however, we no longer end up with subsumed redundant sentences like the second one shown here.

5.2 Information Ordering Updates

While the changes and improvements that we made for information ordering during this deliverable had less than a 0.0005 point impact on our ROUGE-2 recall scores, we believe that they make a noticeable difference on the readability of the final summary outputs. For example, consider the ordering differences for one summary that had the same content selected for deliverables #3 and #4. Here is the output for deliverable #3:

The destination is Medan, capital of North Sumatra, Indonesia. The first batch of 100-ton cargoes was airlifted to Colombo, capital of Sri Lanka, Wednesday morning. The official of the ministry said that at least 476,619 people were refugees, and the figure could increase as there were still many others uncounted by the officials. An extremely powerful earthquake and the tsunami that followed devastated many parts of North Sumatra and Aceh on Dec. 26 last year, killing more than 100,000 people there. Bangladesh has already sent medicines and other relief goods for the tsunami victims in Sri Lanka and the Maldives.

And here is the ordering for deliverable #4:

An extremely powerful earthquake and the tsunami that followed devastated many parts of North Sumatra and Aceh on Dec. 26 last year, killing more than 100,000 people there. Bangladesh has already sent medicines and other relief goods for the tsunami victims in Sri Lanka and the Maldives. The official of the ministry said that at least 476,619 people were refugees, and the figure could increase as there were still many others uncounted by the officials. The destination is Medan, capital of North Sumatra, Indonesia. The first batch of 100-ton cargoes was airlifted to Colombo, capital of Sri Lanka, Wednesday morning.

We think that the new ordering is a pretty obvious improvement, especially with regard to the initial sentence which now provides the necessary context to make sense of the following sentences. While the ordering is not ideal for all of the summaries, we observe many cases of marked improvements overall, particularly for initial sentence selection.

6 Error Analysis

Despite how far our summarization system has come since its initial implementation in terms of producing content-dense and readable summaries, there are still plenty of instances of errors and room for improvement.

6.1 Redundant Sentence Issues

There are still cases of repetitive sentences that are not caught by our lexical overlap metric or the spaCy-based or BERT-based similarity comparison methods, such as the following two sentences:

An Iraqi reporter threw his shoes at visiting U.S. President George W. Bush and called him a "dog" in Arabic during a news conference with Iraqi Prime Minister Nuri al-Maliki in Baghdad on Sunday.

President George W. Bush ducked a pair of shoes hurled at his head–one shoe after the other–in the middle of a news conference with Iraqi Prime Minister Nouri al-Maliki.

In order to lower any of the redundancy thresholds to a limit that will mark these sentences as repetitive, however, we also end up excluding many sentences as redundant that our human evaluative judgments tell us are *not* actually redundant, and end up with much less cohesive sumarries overall. The question of how to improve the redundancy checker to be sophisticated enough to identify that sentences are redundant despite structural and lexical differences but not overly-sensitive is an outstanding issue for our system.

6.2 Named Entity Redundancy

The summary with the lowest average evaluation score from the human readability evaluations that we received was summary number 1105, which reads:

An Adam Air Boeing 737-400 plane with 102 people on board crashed in a mountainous area near the town of Polewali late Monday on its way from Surabaya to Manado. The Adam Air Boeing 737-400 crashed Monday afternoon, but search and rescue teams only discovered the wreckage early Tuesday. The Boeing 737-400 belonging to low-cost carrier Adam Air made a distress call Monday evening and disappeared from the radar screen minutes later. A Boeing 737-400 plane with 102 people onboard crashed into a mountain in the West Sulawesi province Monday, killing at least 90 people.

In addition to some repetition in the sentences' content, we think that the biggest issue here is the fact that the main entity in all four of the sentences in this summary, i.e. the Adam Air Boeing 737-400 plane, is referred to by its full or nearly full entity name each time. That is, we see "An Adam Air Boeing 737-400," "The Adam Air Boeing 737-400," "The Boeing 737-400," and "A Boeing 737-400 plane" appear in each sentence, respectively. Aside from taking up precious space with repetitive information, this leads to a jarring reader experience because one would normally expect the full entity name to be used likely only for its first mention and referred to afterwards by an abbreviated form such as "the plane" or "the Boeing." If we were to continue improving our summarization system, this compression of subsequent named entities is likely the first place we would focus on fixing.

6.3 Referential Issues

One of the most obvious outstanding issues with our system is the handling of co-referents and other referential issues. While our improved selection of summary initial sentences helps with readability by removing some of the worst referential issues from the start of the text, we still see sentences such as the following examples that contain referential words (highlighted in bold) with no antecedents to be found in the summary:

*More than 100 Giant Pandas live **there**.*

***Each time**, computer security managers and users have cleaned up the damage and patched holes in systems.*

***The destination** is Medan, capital of North Sumatra, Indonesia.*

While the content of the surrounding sentences makes it possible to guess about what might be the earlier events/situations/places being referenced in these examples, their absence in the immediately preceding sentences makes for awkward reading.

7 Conclusion

In this deliverable we have finalized our multi-document summarization system, focusing on producing outputs that are cohesive and optimize for readability. In analyzing the final output summaries produced from both the AQUAINT corpus and English Gigaword corpus datasets, we note that there are still some summaries for which we believe the ordering could be more effective and that there are instances of more repetition in the realized content than we would ideally like. However, overall, we believe that our system produces reasonable extractive summaries that are acceptably readable and on-topic in terms of content.

References

Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2012. A preference learning approach to sentence ordering for multi-document summarization. *Inf. Sci.*, 217:78–95.

John M Conroy, Judith D Schlesinger, Dianne P O'Leary, and Jade Goldstein. 2006. Back to basics: Classy 2006. In *Proceedings of DUC*, volume 6, page 150.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Lucy Vanderwende, Arul Menezes, and Chris Quirk. 2015. An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, Denver, Colorado. Association for Computational Linguistics.