RevisionBank: A Resource for Revision-based Multi-document Summarization and Evaluation

Jahna Otterbacher[*], **Dragomir Radev**[*][†]

[*]School of Information
[†] Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI USA
{jahna,radev}@umich.edu

Abstract

Multi-document summaries produced via sentence extraction often suffer from a number of cohesion problems, including dangling anaphora, sudden shifts in topic and incorrect or awkward chronological ordering. Therefore, the development of an automated revision process to correct such problems is a research area of current interest. We present the RevisionBank, a corpus of 240 extractive, multi-document summaries that have been manually revised to promote cohesion. The summaries were revised by six linguistic students using a constrained set of revision operations that we previously developed. In the current paper, we describe the process of developing a taxonomy of cohesion problems and corrective revision operators that address such problems, as well as an annotation schema for our corpus. Finally, we discuss how our taxonomy and corpus can be used for the study of revision-based multi-document summarization as well as for summary evaluation.

1. Introduction

Researchers in text summarization have proposed using a revision process in two different ways – to improve the informativeness and readability of automatically produced summaries and as an evaluation metric for summary quality. In this paper we present RevisionBank, a collection of 240 multi-document summaries produced by the MEAD summarizer (Radev et al., 2000) that have been manually revised by multiple human annotators. First, we introduce the corpus and describe how we built it. We also discuss some of the challenges involved in using revisions produced by human judges either for improving summarization or for the evaluation of summaries. Next, we address the issue of using revision as a method for summary evaluation and discuss our revision framework in relation to previous approaches. Finally, we conclude with our plans for future work on revision-based summarization using the corpus.

2. Revision-based Multi-document Summarization (MDS)

Mani and colleagues (Mani et al., 1998) first proposed that a revision process could be used to correct coherence problems and improve informativeness in single-document summaries. Similarly, (Jing and McKeown, 2000) noted that summaries produced by professionals could often be traced back to a series of cut-and-paste operations applied to the original document source. Inspired by this, we endeavored to see if a revision process could be used to improve the quality of extractive, multi-document summaries. Our motivations stem from two observations. Firstly, multi-document summaries produced via sentence extraction often suffer from an array of cohesion problems, including dangling anaphoric expressions and sudden shifts in topic. Therefore, summaries produced by sentence extraction are often not as fluent as those produced using more knowledge-intensive means, such as the extract and regenerate approach, which relies on finding pre-defined information types in filling in slots in the summary (Radev and McKeown, 1998). However, extractive summarizers that do not rely on any knowledge base, such as MEAD, can take input documents from any domain. Therefore, ultimately our aim is to see if a revision process can be developed that could work with existing extractive multi-document summarizers in order to promote cohesion and thus, more fluent summaries. To this end, we have built a resource for studying how humans revise automatically produced summaries.

3. Developing a Taxonomy of Cohesion Problems and Revision Operations

In (Otterbacher et al., 2002), we performed a corpus analysis of multi-document summaries produced by the default configuration of the MEAD summarizer. We manually revised a set of summaries of different lengths and produced from a variety of source documents. The goal was to identify common cohesion problems, as well as addition, modification and deletion operations that could address such problems. Following (Halliday and Hasan, 1976), we view cohesion as referring to "relations of meaning that exist within the text, and that define it as a text." Cohesion occurs "where the interpretation of some element in the discourse is dependent on that of another (p. 2)." Figure 1 shows an example of a unedited summary produced from a cluster of 11 documents related to the August 2000 crash of a Gulf Air jet in Bahrain. Several cohesion problems can be noted:

1. In sentences 3 and 4, the pronoun "he" has no antecedents.
2. The adverb "also," used in sentences 2 and 3, makes reference to previous events that have not been described in the summary.
3. There seems to be no transition between the sentences.

Figure 2 shows a possible revision of the same summary. It can be seen that four revision operations have been applied to the original summary:

1. In the second sentence, the phrase indicating the location of the event, "at Cairo airport," has been moved to the beginning of the sentence, in order to stress that the event and location described in the sentence are different than in sentence 1.

2. In sentences 2 and 3, the adverb "also" has been deleted.

3. In sentence 3, the pronoun "he" has been replaced with the antecedent to which "he" refers (namely, Bahrain's prime minister). This information was obtained from context in the source document.

4. In sentence 4, "he" has been replaced with the corresponding antecedent from the original document, namely "one witness."

Through the process of revising our corpus of summaries, we developed a taxonomy of cohesion problems and revision operations, which is briefly described in Table 1. The taxonomy consists of five major categories of problems for MDS. Examples of more specific problems within each category are shown in the table.[1]

```
[1] More than 130 bodies are reported to have
been recovered after a Gulf Air jet carrying 143
people crashed into the Gulf off Bahrain on
Wednesday. [2] Distraught relatives also gathered
at Cairo airport, demanding information.[3] He
also declared three days of national mourning. [4]
He said the jet fell "sharply, like an arrow."
```

Figure 1: Example of an unrevised summary.

```
[1] More than 130 bodies are reported to have
been recovered after a Gulf Air jet carrying 143
people crashed into the Gulf off Bahrain on
Wednesday. [2] At Cairo airport, distraught
relatives gathered, demanding information.[3]
Bahrain's Prime Minister Sheik Khalifa bin Salman
Al Khalifa declared three days of national mourning.
[4] One witness said the jet fell "sharply, like an
arrow."
```

Figure 2: A possible revision of example summary.

4. The RevisionBank Corpus

The corpus currently consists of a total of 240 summaries, 20 summaries produced from 4 different clusters of topically-related documents, revised by three annotators each. Six linguistics students were hired as paid annotators in revising the summaries. In addition, a fifth cluster, Milan, was used for development purposes and in particular, to verify that the annotation instructions and guidelines were clear. This cluster was annotated by the first author and one of her colleagues in the development stage of the study. Properties of the document clusters, including the number of source documents and news sources, and the time span of the publication of the source documents, are shown in Table 2.

The tag set we developed for the annotation task incorporated both the problems and the revision operators from our taxonomy. Thus, for a given revision made by a judge, we can tell what he or she deemed to be problematic and how it can be corrected. Two simple examples of problems and corrective operations are shown in Table 3. An example of a revised summary as marked up in our corpus is shown in Figure 3.

4.1. Interjudge Agreement

One challenge in finding the interjudge agreement in our corpus is that the particular revisions made by an annotator to a summary are not independent of one another. In other words, the revisions that one makes in an early sentence of a summary may influence those that need to be made later on in the summary, due to the discourse relations that hold between sentences. As a result, calculating an expected chance value of agreement cannot be done completely accurately. Another problem is that if one wishes to use the sentence as a unit for calculating judge agreement, there may be several revisions that apply to one sentence. If judges agree with respect to some but not all of these revisions, it is difficult to quantify the agreement between them.

Currently, we have used a simplified definition of agreement between judges for each of the summaries in the corpus. Specifically, we have calculated the degree to which judges agree that each sentence in a summary requires revision. Kappa (Carletta, 1996) for the corpus, given our definition, is 0.7234, indicating a rather satisfactory degree of agreement between our judges. However, it should be noted that, given the nature of our revision taxonomy and our annotation schema, interjudge agreement could be defined in various ways.

5. Revision for Evaluation

Recently, there has been some interest in using revision as an evaluation method for automatically produced summaries. In the Text Summarization Challenge (Okumura et al., 2003), human revisers were instructed to revise a set of automatically produced (single- and multi-document) summaries using three editing operations (insertion, deletion, and replacement). Unlike in our work, they were unrestrained as to the revisions they could make. In order to evaluate the summaries, the average number of operations performed as well as the average number of revised characters was computed. While such an evaluation method can be used to compare, on average, the quality of the summaries according to the human revisers, using a more constrained revision process, such as defined in our framework, could provide more insight as to how summaries differ. For example, our tags indicate not only what revision was performed by the judge but also the problem that he or she addressed. Therefore, one could compare the types of revisions that were performed on an automatically produced summary in comparison to a baseline (or gold standard)

[1] The full list of cohesion problems in our taxonomy as well as detailed examples can be found in (Otterbacher et al., 2002).

Problem category	Description	Example problems
Discourse	Concerns relationships between sentences and relationships of individual sentences to the overall summary	Sudden topic shift; redundant sentence; contrast/contradiction
Entities	Involves the resolution of referential expressions such that each entity can be easily identified	Repeated entity; bare anaphor
Temporal	Concerns the establishment of correct temporal relations between events	Incorrect temporal ordering; anachronism
Grammar	Concerns the corrections of grammatical errors, often due to previous revisions	Run-on sentence; mismatched verb tenses
Place/ setting	Involves establishing where each event takes place	Change of location; collocation

Table 1: Taxonomy of cohesion problems and revision operations.

Cluster	Topic/News Story	Documents	News Sources	Time span	Annotators
Milan (development cluster)	Crash of a small plane into a Milan skyscraper on April 18, 2002	9	5	3 hours	J,Z
GulfAir	Crash of GulfAir flight 072 in August 2000 in Bahrain	11	7	52 hours	B,E,F
RI	Rhode Island nightclub fire on February 20, 2003	10	5	21.5 hours	C,D,F
Shuttle	Columbia space shuttle disaster of February 2003	10	6	60 hours	A,B,C
Turkish	Turkish Airlines plane crash in January 2003	10	5	135 hours	A,D,E

Table 2: Clusters of articles in the RevisionBank corpus

summary, in addition to the average rates of revision, as in (Okumura et al., 2003).

Also, in using our framework for evaluative revisions, interjudge agreement can be examined. For example, if two judges have performed the same number of editing operations on a given summary, this does not necessarily imply that they corrected the same problem, or even that they edited the same sentences. It could be the case that one judge identified two problems in the summary while the other found only one but edited the same number of characters. Similarly, two different summaries with the same average number of edit operations may not be equally troubled. Finally, in a particular summarization task, some problems may be more serious than others, and assessors may wish to assign different penalties for them. In short, using a more restrained revision framework for evaluation might offer more flexibility in summary evaluation in addition to providing more insight into what kinds of problems summarization researchers should address in order to improve the quality of summaries.

6. Future Work and Conclusion

Our future work related to the RevisionBank project will involve the machine learning of revision rules for revision-based MDS. To this end, we plan to use transformation-based learning (Brill, 1995) in automatically deriving revision rules. Here, we briefly outline our plans for this process:

1 Find an appropriate way to parse the sentences, such that the revision rules can adequately be formalized.

2 Using Brill's transformational approach to learning, one revision rule would be applied to raw text at a given iteration. Then, a distance metric would be used to quantify the error between the automatically revised text, and the target text - that revised by the human revisers. If the application of the rule decreases the error, the rule is kept.

3 The set of revision rules established by the learning technique will be implemented as a revision module. The module will behave as a post-processor of the output of the MEAD extractive summarizer.

4 Finally, we need to know how much our revision process improves the summaries. Eventually, a user study should be conducted in which readers score the summaries before and after the revision process. This will allow us to assess the contribution of the work.

In summary, we have established a corpus of 240 manually revised multi-document summaries, annotated with respect to our previously developed taxonomy of revision operations. More information about the corpus can be obtained from $http : //www - personal.umich.edu/{\sim}jahna/revision.htm$.

7. References

Brill, Eric, 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics*.

Carletta, Jean, 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *CL*, 22(2):249–254.

Halliday, M. and R. Hasan, 1976. *Cohesion in English*. London: Longman.

Jing, Hongyan and Kathleen R. McKeown, 2000. Cut and Paste-Based Text Summarization. In *Proceedings of the*

Problem	Revision operation(s)	Example
Redundant entity	Delete NP; add pronoun	Mrs. Lo announced that the the number of young people abusing drugs fell in 1999. <DEL-np> Mrs. Lo </DEL-np> <ADD-pro> She </ADD-pro> said “The number of drug abusers under age 21...”
Ambiguous anaphor	Delete pronoun; add NP	If <DEL-pro> it </DEL-pro> <ADD-np> Pinatubo </ADD-np> does erupt, its primary means of causing death...

Table 3: Examples of revisions

```
<?xml version ='1.0' encoding='UTF-8'?>
<!DOCTYPE DOCSENT SYSTEM
"/clair4/projects/revision_corpus/dtd/revision1.dtd">
<REVDOC DID='11.gulfair' LANG='ENG'>
<HEADER JUDGE="J" CLUSTER="gulfair" />
<BODY>
<TEXT>
<S SNO="1">  <DEL CAT="dis" TYPE="rem"> ABCNews.com: 10 Dead 100 Hurt in
Huge R.I. Club Fire </DEL> </S>
<S SNO="2">  <DEL CAT="dis" TYPE="rem"> FOXNews.com  Fire Engulfs R.I.
Nightclub; At Least 10 Dead, 164 Injured </DEL> </S>
<S SNO="3">  <DEL CAT="loc" TYPE="stamp"> WEST WARWICK, R.I. </DEL>  An
immense fire tore through a Rhode Island nightclub, killing at least 10 and
injuring 164, authorities said. </S>
<S SNO="4"> The blaze broke out close to 11 p.m. Thursday during a
pyrotechnics display during a Great White concert at The Station in West
Warwick, about 15 miles southwest of Providence. </S>
</TEXT>
</BODY>
</REVDOC>
```

Figure 3: Example of a formatted summary.

6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics. Seattle, WA.

Mani, Inderjeet, Barbara Gates, and Eric Bloedorn, 1998. Using Cohesion and Coherence Models for Text Summarization. In Eduard Hovy and Dragomir R. Radev (eds.), *Proceedings of the AAAI Symposium on Intelligent Text Summarization*. Stanford, CA: AAAI Press.

Okumura, Manabu, Takahiro Fukusima, and Hidetsugu Nanba, 2003. Text Summarization Challenge 2: Text Summarization Evaluation and NTCIR Workshop 3. In *Proceedings of HLT-NAACL 2003*. Edmonton.

Otterbacher, Jahna, Dragomir Radev, and Airong Luo, 2002. Revisions that Improve Cohesion in Multi-document Summaries: A Preliminary Study. In *Proceedings of the Workshop on Text Summarization at the 40th Meeting of the Association for Computational Linguistics*. Philadelphia, PA.

Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska, 2000. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. In *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA.

Radev, Dragomir R. and Kathleen R. McKeown, 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, 4:469–500.