Summarization: an Application for NL Generation

Beryl Hoffman
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, U.K.
hoffman@cogsci.ed.ac.uk

1 Introduction

In this paper, I will be exploring techniques for automatically summarising texts, concentrating on selecting the content of the summary from a parsed (semantic) representation of the original text. Summarization is a particularly nice application for natural language generation because the original text can serve as the knowledge base for generating the summary. In addition, we only need to develop a lexicon limited to the words and senses in the original text (as long as we use the same words in the same context as the original text). This simplifies the generation task somewhat.

However, summarization is not a trivial task. We must first analyze the original text using a robust grammar that can produce a reliable semantic interpretation of the text. To simplify this investigation, I will not tackle the many problems of NL analysis, but will use already parsed texts from the TAG Tree Bank (UPenn, 1995). I use a perl script to convert the syntactic structures in this parsed corpus into a list of logical forms that roughly indicate the predicate-argument structure of each clause in the text.[1] We can generate a summary by choosing a subset of this list of LFs. However, choosing the *right* subset is not easy.

The problem is how to judge which clauses are *important*. Sophisticated discourse analysis is needed in order to interpret the intentional and rhetorical structure of the original text and then prune it in the appropriate ways. However, discourse analysis is a hard task that requires an immense amount of world knowledge (Sparck-Jones, 1993). I investigate ways to generate a summary without full interpretation of the original text. I use Centering Theory to roughly segment the text, as described in the next section. Then, as described in section 3, a set of pruning rules based on centers and discourse relations are used to select the content of the summary. First, those segments that are about the most frequent centers of attention are selected, and then these segments are pruned by recognizing non-critical *elaborations* among the propositions. Another heuristic used is to select *restatements* among the propositions for the summary, since restatement is a good indicator of important information. The proposed summarization heuristics are tested out on a sample text in section 4; an implementation to test out these heuristics is in progress.

[1] A parser which directly produces the pred-arg structure is probably preferable to this method. Note that the parser probably would not have to resolve all syntactic ambiguities in the the summarization task, because we can preserve the same ambiguities in the summary, or delete some of the problem phrases such as PPs in the summary anyway.

2 Discourse Segmentation

Centering Theory (Grosz, Joshi, and Weinstein, 1995) is a computational model of *local* discourse coherence which relates each utterance to the previous and the following utterances by keeping track of the center of attention. The most salient entity, the center of attention, at a particular utterance is called the backward looking center (Cb). The Cb is defined as the highest thematically ranked element in the previous utterance that also occurs in the current utterance. If there is a pronoun in the sentence, it is preferred to be Cb.

Centering Theory can be used to segment a discourse by noting whether the same center of attention, Cb, is preserved from one ut-

terance to another. Basically, we can either CONTINUE to talk about the same entity or SHIFT to a new center. A SHIFT indicates the start of a new discourse segment.[2]

In the method that I am proposing, the original text is first divided into segments according to Centering Theory. Then, as described in the following sections, the segments which are about the most frequent Cb(s) in the text are selected for the summary, and then the discourse relations of elaboration and restatement are used to further prune and select information for the summary.

3 Content Selection

3.1 Frequent Centers

After the text has been segmented, we need to decide which of the discourse segments are important for the summary. The most prevalent discourse topic will play a big role in the summary. Thus, the most frequent Cb can be used to select the important segments in the text. I propose the following heuristic:

> **Heuristic 1:** Select those segments that are about the most frequent Cb in the text[3] for the summary.

Picking the most frequent Cb gives better results than simply picking the most frequent words or references as the most important topics in the text. For example, in the sample text (see Section 4) about a new electronic surveillance method being tried on prisoners that will allow them to be under house-arrest, "wristband" occurs just as frequently as "surveillance/supervision", however "surveillance/supervision" is a more frequent Cb than "wristband", and this reflects the fact that it is a more central topic in the text.

3.2 Pruning Elaborations

While doing the centering analysis of my sample texts, I noticed that it is the segment boundaries, the SHIFTs, that are important for summarization in the discourse analysis of the original text. In fact, the CONTINUE transitions in Centering often correspond to Elaboration relations in RST (Mann and Thompson, 1987). A restricted type of the elaboration relation between sentences can be restated in Centering terms:

> **Elaboration on the same topic:** the subject of the clause is a pronoun that refers to the subject of the previous clause – a CONTINUE in centering.

Thus, I propose the following heuristic for pruning the segments in the summary:

> **Heuristic 2:** Delete elaborations on the same topic (as defined above) in the summary.

For example, the second sentence below can be left out of the summary because it is an elaboration on the same topic.

(1) a. Most county jail **inmates** did not commit violent crimes.
(Cb = inmates, SHIFT)

b. **They**'re in jail for such things as bad checks or stealing.
(Cb = they = inmates, CONTINUE)

3.3 Restatement

Another RST relation that is very important for summarization is Restatement, because restatements are a good indicator of important information. Good authors often restate the thesis, often at the beginning and at the end of the text, to ensure that the point of the text gets across. The heuristic used is:

> **Heuristic 3:**
> Select repeated or semantically synonymous LFs (i.e. predicate-argument relations) in the original text for the summary.

One way to find restatements in the text is to simply search for repeated phrases. However, most good authors restate phrases rather simply repeating them. That is why I propose we search for repeated LFs rather than repeated words or phrases. Since LFs capture the primary relations in a whole clause, their frequency captures dependencies that traditional statistical approaches such as bigrams

[2]There are other cues to discourse segmentation (not yet included in this study) such as tense and aspect continuity and the use of cue words such as "and".

[3]More than one frequent Cb can be picked if there are no clear winners.

and trigrams would miss. However, some inference would be necessary in order to infer whether LFs are semantically synonymous.

For example, the following two sentences from the sample text are very similar. Their semantic representations contain the propositions *call(computer,prisoner)* and *plug-in(prisoner)*, after anaphora resolution and inferences such as that *call(computer,prisoner)* is equivalent to *make(a computerized call,to a former prisoner's home)*. Notice that a simple trigram would not recognize "that person answers by plugging in" in (2)b as a restatement of the "prisoner plugs in". We need to consider the predicate-argument relations instead of simple word collocations.

(2) a. Whenever a computer randomly calls them from jail, the former prisoner plugs in to let corrections officials know they're in the right place at the right time.

b. When a computerized call is made to a former prisoner's home, that person answers by plugging in the device.

Searching for similar LFs captures important information that is restated many times in the text.[4] This method is similar to *aggregation* methods used in NL generation. Summarization can be seen as a massive application of aggregation algorithms. We need to look for shared elements, agents, propositions, etc. in the semantic representation of the original text in order to aggregate similar elements as well as to recognize important elements that the author restates many times.

4 An Example Text

The following is a sample text from the Penn Treebank. The Δ and alternating normal and italicized script mark segment breaks in the text as determined by Centering Theory. Embedded subsegments are shown with brackets. The Cbs are shown in bold.

TEXT:

Δ**Computerized phone calls** [which do everything from selling magazine subscriptions to reminding people about meetings] have become the telephone equivalent of junk mail, but a new application of the **technology** is about to be tried out in Massachusetts [to ease crowded jail conditions]. ΔΔ *Next week some* **inmates** *[T released early from the Hampton County jail] in Springfield will be wearing a wristband [that T hooks up with a special jack on their home phones]. [Whenever a computer randomly calls* **them** *from jail],* **the former prisoner** *plugs in [[to let corrections officials know] [***they***'re in the right place at the right time]].* Δ **The device** is attached to a plastic wristband. **It** looks like a watch. **It** functions like an electronic probation officer. Δ *[When a computerized call is made to a former prisoner's home phone], that person answers by plugging in the device.* Δ The wristband can be removed only by breaking its clasp and [if that's done] the inmate immediately is returned to jail. Δ *The description conjures up images of big brother watching,* Δ but Jay Ash, [deputy superintendent of the Hampton County jail in Springfield], says [the surveillance system is not that sinister]. Such **supervision**, [according to Ash], is a sensible cost effective alternative to incarceration [that T should not alarm civil libertarians]. Δ **Dr. Norman Rosenblatt**, *[dean of the college of criminal justice at Northeastern University], agrees.* **Rosenblatt** *expects electronic surveillance in parole situations to become more wide spread, and* **he** *thinks [eventually people will get used to the idea].* Δ Springfield jail deputy superintendent Ash says [[although **it** will allow some prisoners to be released a few months before their sentences are up], concerns that may raise about public safety are not well founded]. ΔΔ **Most county jail inmates** *did not commit violent crimes.* **They***'re in jail for such things as bad checks or stealing.* **Those on early release** *must check in with corrections officials fifty times a week according to Ash [who says about half the contacts for a select group will now be made by the computerized phone calls].* Δ Initially the program will involve only a handful of inmates. Ash says the ultimate goal is to use **it** [to get about forty out of jail early]. Δ *The Springfield jail [T built for 270 people] now houses more than 500.* Δ

The content of the summary is selected by picking the two segments with the most fre-

[4] Many restatements in the texts involve the most frequent Cb which may serve as an additional heuristic.

quent Cb, the inmate(s)/prisoner. These are marked with two ΔΔs at the beginning of the segments above. Then, elaborations (i.e. CONTINUEs) in these segments are deleted. Essentially, this leaves the first sentence of each segment with the Cb the inmates. In addition, we search for restatements in the text. As a result, the following sentences from the text are selected for the summary. The first and third sentences are the first sentences in the segments about the most frequent Cb, the inmates; the second sentence as well as part of the first sentence is given by recognizing restatements in the text.

Summary:

ΔNext week some inmates released early from the Hampton County jail in Springfield will be wearing a wristband that hooks up with a special jack on their home phones. Δ When a computerized call is made to a former prisoner's home phone, that person answers by plugging in the device. Δ Most county jail inmates did not commit violent crimes. Δ

The summary above just shows the relevant portions of the original text (in the original order) selected for the summary. The heuristics for content selection actually operate on LFs; the selected LFs will then be sent to a generator which can plan a more coherent summary than what is produced above.[5]

5 Conclusions

In this paper, I have outlined the following method for content selection in the summarization task. The content of the summary is selected from a parsed (semantic) representation of the original text. Centering Theory is used to segment the text. Segments that are about the most frequent centers and LFs that are restated in the text are selected as important information for the summary. These segments are then pruned by recognizing elaborations.

1. Parse the original text into a list of logical forms.
2. Divide the original text into segments according to Centering Theory and do anaphora resolution.
3. Select the segments that are about the most frequent Cb(s) in the text.
4. Delete elaborations (i.e. CONTINUEs in Centering terms) in these selected segments, and substitute antecedents for all anaphora in the LFs for these segments.
5. Simplify the LFs in these selected segments by pruning unnecessary adjuncts and embedded clauses.
6. Find restated propositions in the semantic representation of the original text by searching for repeated or semantically synonymous LFs.
7. Generate the summary from the LFs produced by the last two steps.

I believe that the method proposed above shows promise in selecting important information from the original text for the summary. However, a rigorous evaluation of the summaries produced by the method is now needed. I have assumed that in the summarization task the computer does not have to fully understand the original text if it can reuse the same words, phrases, and predicate-argument relations. However, the summary will improve as we undertake deeper (rhetorical and intentional) analysis of the original text and as we move from simply selecting information from the text to inferencing and generalizing from the information in the text.

[5]The selected LFs for each sentence should also be simplified by pruning unnecessary adjuncts and embedded clauses.

References

Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics.*

Karen Sparck Jones. 1993. What might be in a summary? In Knorz, Krause, and Womser-Hackr, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26.

William Mann and Sandra Thompson. 1987. Rhetorical structure theory: A framework for the analysis of texts. Technical Report RS-87-185, ISI.

Penn TreeBank. 1995. University of Pennsylvania. copyrighted.