Baldwin Language Technology's DUC Summarization System

DUC Workshop, September 13, 2001

Breck Baldwin, Aaron Ross
breck@coreference.com, ross@coreference.com
(215) 546-5366

Abstract:

We approached the DUC multi-document summarization task from the perspective that sentences which contained 'interesting phrases' were good candidates for a sentence extraction summary. An interesting phrase was made up of interesting words which in turn were words that had a lower relative entropy in the corpus created by the 10 documents being summarized compared to the words entropy in a background corpus built up from similar data. Selected sentences were presented with '...' between them up to the various word limits for the summaries. Single document summaries were handled with the same basic algorithm.

Approach:

We have long felt that generic summaries are a difficult task. Generic implies a summary without specific purpose or direction. Evaluations of generic summaries reflect a set of unstated assumptions about what kinds of information matters. A summary for an intelligence analyst with rich background knowledge of a topic area is very different than one for a user that knows nothing of the area. We raise these points only to describe the reasoning behind the method our system used to generate 'generic' summaries.

Our hypothesis was that the differences between the 10 document sub-corpus formed by the topic clusters and a more general corpus would highlight what was unique and interesting about the topic cluster. This approach slants the summaries towards users who are familiar with the overall terminology of the background corpus, but likely to be interested in information that is more surprising given the general nature of the corpus. Roughly speaking, we see our users as well read and looking for distinguishing information from a topic based sub-corpus.

In contrast to our approach, one could generate a summary without reference to a background corpus. Done perfectly, such summaries would provide information without prejudice to a background set of information. One would expect such summaries to also contain more general information about the world which is effectively filtered somewhat by our reference to a background corpus.

Details:

We decided to experiment with generic multi-document summarization by selecting phrases that contained information more frequently found than in the background corpus. The distinguishing features of this approach are that we use a background corpus to select information and that the unit of selection is the phrase. In the end we use containing sentences for those phrases for presentation to the user by selecting sentences which contain the longest/most 'interesting phrases'. So the phrase selection basis of the approach is not evident to the end-user.

Interesting Words:

We first took the document collection to be summarized and recognized interesting words using the following definition of an 'interesting' word. A word is 'interesting' if it satisfies any of the following conditions:

1) The word is an honorific or corporate designator (Mr., Corp., Inc. etc).
2) The word is not mentioned in the background corpus.
3) The word was seen at least 5 times in the document corpus
4) The topic collection entropy[1] is at least 3 less than the back ground corpus entropy of the word.

Condition 1) was useful in growing interesting phrases that refer to people or companies. Condition 2) applies if the background corpus does not contain the document collection being summarized. In this case we used other TREC data as the background corpus and had instances of words from the collection to be summarized not being mentioned in the background corpus. It is fairly good bet that an unseen word is a good candidate for inclusion in the summaries even if the word has low entropy in the to be summarized data. . Condition 3) was a base line count requirement that words occur at some basic level of frequency.

Condition 4) reflects a key assumption that words which are low entropy in the collection being summarized relative to the background corpus are more likely to be the words which reflect the idiosyncratic content of the document collection. The value of 3 was chosen by examination of the words selected with the metric ranging from 0 to 6.

Interesting Phrases:

Once the individual words were classified as 'interesting', we grew interesting phrases by looking for adjacent interesting words, or uninteresting words with an interesting word on either side. Sentence boundaries were not spanned. The identification of interesting phrases gave us a point of further leverage into the documents to be summarized because we decided that multi-word collections of adjacent interesting words were likely to point us to even higher value information.

---

[1] Entropy defined as the base 2 log of the probability of the term occurring in the corpus. Low entropy means that a word is frequent in the corpus, high entropy means that a word is rare in a corpus.

The bracketing allowance was made upon examination of system output and often a mid-phrase word would be excluded because it was too common in the background corpus. Output appeared to be improved if some flexibility were added to the phrase recognition, but this was a judgment call.

We considered summarizing with just phrasal extraction but the summaries suffered from coherence problems for some classes of documents.

Producing Summaries:

In the end we decided to use the phrasal information to produce sentence selection summaries with the following algorithm:

```
Foreach sentence in the document collection
        Find the longest 'interesting phrase' that it contained and record the length

While the summary length is less than the length limit
        Pick the sentence with the longest match,
                if sentence length + summary length is less than the length limit
                        add the sentence to the summary

Print each line of the summary with '...' in between
```

We explicitly did not try and mandate as much phrase variation as possible in the summary generation because we did not see such a restriction making a difference in the quality of summaries. In particular, a document collection about a Mr. Smith would perhaps be hurt by heuristics that eschewed interesting word overlap.

Discussion:

We developed a very simple summarization system that cleanly reflected a set of base assumptions about what was an interesting generic summary for a small data collection. Those assumptions are perhaps the most interesting aspect of this work because they reflect what we believe is a coherent approach to generic summary generation. This, of course, means that we are not producing generic summaries at all, but rather summaries that focus on what is idiosyncratic about the sub-corpus defined by the 10 or so documents in the topic as contrasted with the background corpus.

Speaking at a higher level of generalization, our use of entropy models informativeness at some level, and we built a system that focuses on words and phrases that are less informative in the summary collection than in the background corpus. This somewhat counter intuitive observation makes sense if you consider that we are selecting terms which are usually rare, but more frequent in the summary collection. These terms are probably important to the summary collection since they are being used more than expected and as a result sentences which mention them should be used in the generated summaries.

This same basic idea could be couched in a much fuller featured system that did not rely on such simplistic summarization generation techniques. There is no reason the recovered phrases could not be part of a generation system that sought to assemble disjoint phrases into coherent sentences. The basis of phrasal selection could also be constituent based from parse tree output or predicate argument structure. What remains at the core of such elaborated system is that a background corpus, which stands in for a background set of world knowledge or already processed information, can provide an important direction to what counts as being important in a summary.

One last point worth mentioning is that generic summaries might better be thought of as subject to one of the successful approaches in question answering: different questions have different strategies for producing an answer. In the context of summarization, perhaps we should say that different data collections/topics are amenable to different types of summarization approach—I know that NIST made efforts to mix and match the topics to span several kinds of multi-document collections. Perhaps a better sense of what a generic summary is will evolve, when a finer grained analysis is made of the topics/collections that we are trying to summarize.